

Equivariant Gaussian Processes as Limiting Convolutional Networks with Infinite Number of Channels

Andrey Demichev^{a,*}

^a*Skobeltsyn Institute of Nuclear Physics, M.V. Lomonosov Moscow State University,
1(2) Leninskie gory, Moscow 119991, Russian Federation*

E-mail: demichev@theory.sinp.msu.ru

In this work we establish a relationship between the many-channel limit for $SO(2)$ -equivariant convolutional neural networks (CNNs) and the corresponding equivariant Gaussian process (GP) in the case of Fourier space quadratic nonlinearity. The approach used provides explicit equivariance at each step of the relationship derivation.

*The 5th International Workshop on Deep Learning in Computational Physics
28-29 June, 2021
Moscow, Russia*

*Speaker

1. Introduction

Studies in the works [1–4] have established the correspondence between asymptotic (infinitely wide) fully connected neural networks (NN) and Gaussian processes (GP). In particular, the results of [1, 2] showed that the function defined by a single-layer fully-connected NN with infinitely many hidden units, and random independent zero-mean weights and biases is equivalent to a GP. Later these results were extended to arbitrarily deep fully-connected NN with infinitely many hidden units in each layer [3, 4]. This discovery of the asymptotic equivalence between NN and GPs provides obtaining an analytic form for the prior over functions encoded by NN architectures and initializations. Rather than considering NNs to be determined by a learning parameter space distribution, this relationship allows one to consider NNs themselves as samples from a function space distribution. Such uncovering of an explicit prior over functions can be analytically investigated and therefore provides means for a theoretical understanding of deep learning. For example, in the work [5] predictions for learning curves of deep NNs trained on polynomial regression problems were derived, and in [6] predictions of NN outputs with high accuracy were obtained by using the NN-GP relationship.

The basic idea behind these studies is that a set of NNs $f_{\theta,n} : \mathcal{X} \rightarrow \mathcal{Y}$ with learning parameters θ and width n admits a limit $n \rightarrow \infty$ in which the networks are equivalent to GP, that is f have a Gaussian distribution on the function space: $f \sim \exp \left\{ - \int d^{|X|}x d^{|X|}x' f(x)K(x, x')f(x') \right\}$, where $K(x, x')$ is the GP kernel. Clearly, in practice one is interested in networks with finite width n . It is supposed that they can be drawn from a distribution that receives $1/n$ corrections relative to the Gaussian distribution, i.e., from a non-Gaussian process (NGP), see *e.g.*, [7]. It is worth noting that from the technical point of view studying neural networks with close-to-Gaussian distribution on function space are to some extent analogous to perturbative quantum field theory ([8] and refs therein).

However, fully-connected networks are rarely used in practice, as they are unable to exploit such important properties of data as their possible invariance with respect to continuous or discrete groups of transformations. This disadvantage is eliminated in the architecture of convolutional neural networks (CNN; see, *e.g.*, [9]). In particular, classical CNNs are very successful in the image recognition domain because they ensure equivariance to translations. Therefore, recently the equivalence between NNs and GPs was extended to deep convolutional neural networks [10, 11]. In this case, if each hidden layer has an infinite number of convolutional filters (that is infinite number of channels), the network prior is equivalent to a GP. There have been many recent attempts [12–18] to generalize translation-invariant usual CNNs to other symmetry transformations. In particular, the important result of the work [15] is the proof that convolutional structure is not just a sufficient, but also a necessary condition for equivariance to the action of a compact group. Equivariance guarantees that exactly the same filters are applied to each part the input image regardless of position and that the network can detect any given object equally well regardless of its location respecting data symmetry properties. The formalism used in this studies is based on concepts from representation theory and non-commutative Fourier harmonic analysis. From a mathematical point of view, the relevance of Fourier theoretic ideas in this domain is a consequence of the fact that neuron activations form representation space of the underlying symmetry group. It is worth noting that the Fourier formulation also exists for the usual CNNs with translational equivariance [19, 20],

and the corresponding neural network implementations demonstrate good performance.

It is important that the above mentioned works [10, 11] on establishing the relations between CNNs and GP deal only with translational equivariance and only in the real spaces (RS) of images (not in the Fourier spaces (FS)). On the other hand there exists investigations of equivariant neural processes [21] but without established relations with CNNs in the appropriate limit. The present work is the first step towards filling the gap between equivariance of CNNs and the corresponding GPs.

2. Many-channel CNNs in Fourier Space

2.1 General Setup

For simplicity of presentation we consider 1D convolutional networks with layers and filters being circles S^1 . In this particular case one can consider usual discrete real space with circularly-padded activations [10, 22] and discrete Fourier transform [19]. However, having in mind generalization of the present approach to more complicated symmetries (e.g., spherical $SO(3)$ symmetry), which does not consistent with discretization, we shall consider activations and filters as functions on *continuous* index space as in [15, 16]. In this case the discretization needed for practical realization of the CNNs is achieved by appropriate cutoff of the Fourier series [14, 16].

Our analysis is straightforward to extend to higher dimensional torus with $SO(2)$ symmetry along each dimension. Generalization to the cases of more complicated symmetries (like, spherical symmetry $SO(3)$ or Euclidean group $E(2)$) will be considered in subsequent works.

We consider a series of $L + 1$ convolutional layers, $l = 0, \dots, L$. Since we consider a toy model intended for developing the general approach, for further simplicity we omit bias terms in the CNNs:

- on the one hand, non-zero biases can be easily restored in the expressions below;
- on the other hand, CNNs do not necessarily require non-zero biases, and there exist well known architectures like ResNet with realizations where the convolution layers have no biases¹.

The parameters of the network. These are the convolutional filters, $\omega_{ij}^l(\nu)$ with channel indices i, j and filter spatial location $\nu = \varphi/2\pi \in \mathbb{R}/\mathbb{Z}$, where φ is the angular coordinate on S^1 . Further details about our notation are as follows:

- We use Roman letters to index channels and Greek letters for spatial location.
- To make expressions as easy readable as possible, we denote functions on spatial space and their Fourier transforms by the same letters but with spatial variable in parenthesis while the respective Fourier components are denoted by letters with Greek indices. For example, $\omega(\nu)$ denotes spacial function and ω_α denotes α -component of its Fourier transform.

¹<https://github.com/KaimingHe/deep-residual-networks/blob/master/prototxt/ResNet-152-deploy.prototxt>

- According to general approach to the NN-GP correspondence [1–4, 10, 11], it is assumed the Gaussian prior on the filter weights in the Fourier space,

$$\omega_{ij,\beta}^l \sim \mathcal{N}\left(0, \frac{\sigma_\omega^2}{n^l}\right), \quad (2.1)$$

where

- σ_ω^2 is the weight variance,
- n^l is the number of channels (filters) in layer l .

Thus all the $\omega_{ij,\beta}^l$ are *i.i.d.* normally distributed random variables.

Functions on a circle and the network inputs. It is well known (see, e.g., [23, 24]) that the Fourier components f_α of a function on S^1

$$f(v) = \sum_{\alpha=-\infty}^{\infty} f_\alpha e^{i2\pi\alpha v}, \quad v \in \mathbb{R}/\mathbb{Z}, \quad (2.2)$$

transforms under the action of $SO(2)$ as a one-dimensional representations:

$$\begin{aligned} T_{g(\mu)} f(v) &\equiv f(v + \mu) \\ &= \sum_{\alpha=-\infty}^{\infty} \left(T_\alpha(g(\mu)) f_\alpha\right) e^{i2\pi\alpha v} = \sum_{\alpha=-\infty}^{\infty} f_\alpha e^{i2\pi\alpha(v+\mu)} \\ &\Rightarrow T_\alpha(g(\mu)) f_\alpha = e^{i2\pi\alpha\mu} f_\alpha. \end{aligned} \quad (2.3)$$

Let \mathcal{X} denote a set of input data (images) (training set, validation set, or both). Input data $x_i(v) \in \mathcal{X}$ on S^1 can be expanded into Fourier series

$$x_i(v) = \sum_{\alpha=-\infty}^{\infty} x_{i,\alpha} e^{i2\pi\alpha v}, \quad (2.4)$$

where i is the channel index. Since $x_i(v)$ are real,

$$\bar{x}_{i,\alpha} = x_{i,-\alpha}, \quad (2.5)$$

the bar over a character denotes the complex conjugate. Finite dimensionality of the input data is achieved by cutting off the Fourier series at some value Q of the component index

$$x_i^{(Q)}(v) = \sum_{\alpha=-Q}^Q x_{i,\alpha} e^{i2\pi\alpha v}, \quad x_i^{(Q)}(v) \in \mathcal{X}^{(Q)} \quad (2.6)$$

The set of such functions with finite series will be denoted by $\mathcal{X}^{(Q)}$ and its dimension is $|\mathcal{X}^{(Q)}| = n^0(Q + 1)$ (for real-valued quantities). For intermediary expressions it is convenient to use the complete Fourier series (2.4), and to perform transition to finite-dimensional quantities at the final stage.

Pre-activations, and activations. The network has activations $y^l(\nu; x)$ and pre-activations $z^l(\nu; x)$ for each input image $x \in \mathcal{X}$. They are defined by the following mappings from layer to layer:

$$y_i^l(\nu; x) = \begin{cases} x_i(\nu) & l = 0 \\ \phi^{(RS)}(z_i^{l-1}(\nu; x)) & l > 0 \end{cases}, \quad z_i^l(\nu; x) = \sum_{j=1}^{n^l} \int_{-1/2}^{1/2} d\nu' \bar{\omega}_{ij}^l(\nu') y_j^l(\nu + \nu'; x). \quad (2.7)$$

As it is seen, the pre-activations $z_i^l(\nu; x)$ are defined by the cross-correlation² for the functions on S^1 . Though in the real space all functions are real-valued, we keep the genuine definition of the cross-correlation with complex conjugate first multiplier because we use the complex-valued form of the Fourier transform. Also we explicitly indicate the dependence of $y_i^l(\nu; x)$ and $z_i^l(\nu; x)$ on the input x . The function $\phi^{(RS)}$ defines a nonlinearity in the real space; following to [16] we do not require that it has pointwise nature.

After the Fourier transform similar to that of the input data (2.4), the relations (2.7) take the form

$$y_{i,\alpha}^l(x) = \begin{cases} x_{i,\alpha} & l = 0 \\ \phi_\alpha(z_{i,\bullet}^{l-1}(x), z_{i,\bullet}^{l-1}(x)) & l > 0 \end{cases}, \quad z_{i,\alpha}^l(x) = \sum_{j=1}^{n^l} \bar{\omega}_{ij,\alpha}^l y_{j,\alpha}^l(x), \quad (2.8)$$

where $z_{i,\bullet}^{l-1}$ means that ϕ_α may depend on any of the Fourier components. According to the general property, the cross-correlation in (2.7) becomes the pointwise multiplication (2.8) in the Fourier space.

Following to [16], we choose the *quadratic* nonlinearity in the Fourier space:

$$y_{i,\alpha}^l(x) = \sum_{\beta=0}^{\infty} \bar{z}_{i,\beta}^{l-1}(x) z_{i,\alpha+\beta}^{l-1}(x). \quad (2.9)$$

A few comments to the chosen form of the nonlinearity are in order:

- we have chosen the simplest form of the quadratic FS nonlinearity which does not lead to increasing in the number of channels in a subsequent layer (*cf.* [16]);
- the left hand side of (2.9) transforms according to the definite (defined by the Fourier component index α) irreducible representation (irrep) of the SO(2) group;
- since the nonlinearity (2.9) has the form of cross-correlation in the Fourier space, in the real space it corresponds to local (pointwise) quadratic activation function; as it was noticed in [19], together with appropriate linear term such a nonlinearity can approximate the famous ReLU function [25]; for simplicity we omit such a linear term in (2.9).

Taking into account the transformation of Fourier modes under the rotation group (2.3) and the fact that the nonlinearity (2.9) has equivariant form, it is readily seen that the defining equations (2.8) (and hence the CNN) are explicitly equivariant. In particular, by induction starting from $y_i^0 = x_i(\nu)$ we have

$$\begin{aligned} z_{i,\alpha}^l(T_{g(\mu)}x) &= e^{i2\pi\alpha\mu} z_{i,\alpha}^l(x), \\ y_{i,\alpha}^l(T_{g(\mu)}x) &= e^{i2\pi\alpha\mu} y_{i,\alpha}^l(x), \end{aligned} \quad (2.10)$$

²Remind that though the networks are called ‘‘convolutional’’, the mapping for pre-activations are usually defined by the cross-correlation (closely related to the convolution operation).

Activation covariance. One of the most important quantity for the derivation of the CNN-GP correspondence is the uncentered covariance matrix K^l of the activations y^l , defined as [10]:

$$K_{\alpha,\alpha'}^l(x,x') \stackrel{\text{def}}{=} \frac{1}{n^l} \sum_{i=1}^{n^l} \bar{y}_{i,\alpha}^l(x) y_{i,\alpha'}^l(x'). \quad (2.11)$$

Notice that

- K^l for $l > 0$ are random variables indexed by two inputs x, x' and two Fourier component indices α, α' ;
- K^0 is the deterministic covariance of inputs;
- according to the definition, K^l is transformed under rotation of the inputs as follows

$$K_{\alpha,\alpha'}^l(T_{g(\mu)}x, T_{g(\mu')}x') = e^{i2\pi(\alpha'\mu' - \alpha\mu)} K_{\alpha,\alpha'}^l(x, x'). \quad (2.12)$$

2.2 Convergence of top-layer pre-activations to normal random vector with explicit equivariance

In this section, following to the general approach presented in [10], we show that the top-layer pre-activations z_α^L converge in distribution to a normal random vector with a particular covariance matrix as number of the channels in *hidden* layers are taken to infinity: $\min\{n^1, \dots, n^L\} \rightarrow \infty$ (the number of channels in the top-layer n^{L+1} remains fixed). Our main distinction from the work [10] is that in the passage to the limit we keep the explicit equivariance with respect to the symmetry of input data ($SO(2)$ group in the considered case).

First of all we notice that since $\omega_{ij,\alpha}$ defined in (2.1) are *i.i.d.* variables, they satisfy

$$\mathbb{E}[\omega_{ij,\alpha}^l \omega_{i'j',\alpha'}^l] = \mathbb{E}[\bar{\omega}_{ij,\alpha}^l \bar{\omega}_{i'j',\alpha'}^l] = 0, \quad (2.13)$$

$$\mathbb{E}[\bar{\omega}_{ij,\alpha}^l \omega_{i'j',\alpha'}^l] = \mathbb{E}[\omega_{ij,\alpha}^l \bar{\omega}_{i'j',\alpha'}^l] = \frac{\sigma_w^2}{2n^l} \delta_{ii'} \delta_{jj'} \delta_{\alpha\alpha'}, \quad (2.14)$$

The pre-activations z^l are linear combinations of the Gaussian variables ω^l , specified by the previous layer's activations y^l . Thus for the conditioned on y^l pre-activations one has

$$\mathbb{E}[\bar{z}_{i,\alpha}^l(x) z_{i,\alpha'}^l(x')] = \frac{\sigma_w^2}{2} \delta_{\alpha\alpha'} K_{\alpha\alpha'}^l(x, x'). \quad (2.15)$$

In the derivation of (2.15) the relations (2.8), (2.13), (2.14) and (2.11) were used. Since a linear combination of Gaussian variables is itself a Gaussian we can conclude that

$$(z^l | y^l) \sim \mathcal{N}_C(0, \tilde{\Gamma}), \quad (2.16)$$

where $\mathcal{N}_C(0, \tilde{\Gamma})$ is the circularly-symmetric central complex normal distribution (see, e.g., [26]), and the covariance $\tilde{\Gamma}$ reads

$$\tilde{\Gamma}_{\alpha\alpha'}^l(x, x') = \frac{\sigma_w^2}{2} \delta_{\alpha\alpha'} K_{\alpha\alpha'}^l(x, x') \otimes I_{n^{l+1}}. \quad (2.17)$$

According to (2.16) and (2.17) the normal distribution of $(z^l|y^l)$ depends only on K^l . Thus similarly to [10] we can conclude that the random variable $(z^l|K^l)$ has the same distribution: $(z^l|K^l) \sim \mathcal{N}_C(0, \tilde{\Gamma})$.

Now by means of the relation (2.8) we can express the covariance (2.11) via the pre-activations, take $n^l \rightarrow \infty$ limit and use the weak law of large numbers (cf. [10]):

$$\begin{aligned}
K_{\alpha\alpha'}^l(x, x') &= \frac{1}{n^l} \sum_{i=1}^{n^l} \bar{\phi}_\alpha \left(\bar{z}_{i,\cdot}^{l-1}(x), z_{i,\cdot}^{l-1}(x) \right) \phi_{\alpha'} \left(\bar{z}_{i,\cdot}^{l-1}(x'), z_{i,\cdot}^{l-1}(x') \right) \\
&\xrightarrow{n^l \rightarrow \infty} \mathbb{E}_{z \sim \mathcal{N}_C(0, \Gamma^{l-1})} \left[\bar{\phi}_\alpha \left(\bar{z}_{\cdot,\cdot}^{l-1}(x), z_{\cdot,\cdot}^{l-1}(x) \right) \phi_{\alpha'} \left(\bar{z}_{\cdot,\cdot}^{l-1}(x'), z_{\cdot,\cdot}^{l-1}(x') \right) \right] \\
&= \frac{1}{\mathcal{Z}} \int \left[\prod_{\gamma \geq 0, x} d\bar{z}_{\gamma, x} z_{\gamma, x} \right] \sum_{\beta} z_{\beta, x} \bar{z}_{\alpha+\beta, x} \sum_{\beta'} \bar{z}_{\beta', x'} z_{\alpha'+\beta', x'} \\
&\quad \times \exp \left\{ - \sum_{\delta, \bar{x}; \delta', \bar{x}'} \bar{z}_{\delta, \bar{x}} \left(\Gamma^{l-1} \right)_{\delta, \bar{x}; \delta', \bar{x}'}^{-1} z_{\delta', \bar{x}'} \right\}, \quad (2.18)
\end{aligned}$$

where \mathcal{Z} is the normalization factor, $\Gamma_{\alpha, x; \alpha', x'}^{l-1} \equiv \Gamma_{\alpha\alpha'}^{l-1}(x, x')$ is defined by the relation

$$\Gamma_{\alpha\alpha'}^{l-1}(x, x') = \frac{\sigma_w^2}{2} \delta_{\alpha\alpha'} K_{\alpha\alpha'}^{l-1}(x, x'), \quad (2.19)$$

and we have used the nonlinearity (2.9). The expression (2.18) is written in the matrix form for easier perception.

As usual, for the calculation of the Gaussian-like integral we add to the exponential of (2.18) the terms with external sources, namely

$$\sum_{\delta, x} \left(\bar{J}_{\delta, x} z_{\delta, x} + \bar{z}_{\delta, x} J_{\delta, x} \right), \quad (2.20)$$

so that the integral (2.19) can be expressed as follows

$$\begin{aligned}
K_{\alpha\alpha'}^l(x, x') &= \sum_{\beta} \frac{d}{dJ_{\beta, x}} \frac{d}{d\bar{J}_{\alpha+\beta, x}} \sum_{\eta} \frac{d}{dJ_{\eta, x}} \frac{d}{d\bar{J}_{\alpha'+\eta, x'}} \exp \left\{ \sum_{\gamma, x; \gamma', x'} \bar{J}_{\gamma, x} \Gamma_{\gamma x, \gamma' x'}^{l-1} J_{\gamma', x'} \right\} \Big|_{\bar{J}=J=0} \\
&= \frac{\sigma_w^4}{4} \left[\delta_{\alpha 0} \delta_{\alpha' 0} \sum_{\beta} K_{\beta, \beta}^{l-1}(x, x) \sum_{\eta} K_{\eta, \eta}^{l-1}(x', x') \right. \\
&\quad \left. + \delta_{\alpha\alpha'} \sum_{\beta} \bar{K}_{\beta, \beta}^{l-1}(x, x') K_{\alpha+\beta, \alpha'+\beta}^{l-1}(x, x') \right], \quad (2.21)
\end{aligned}$$

Thus for $\alpha, \alpha' \neq 0$ we have

$$\begin{aligned} K_{\alpha\alpha'}^l(x, x') &= \frac{\sigma_w^4}{4} \delta_{\alpha\alpha'} \left[\sum_{\beta} \bar{K}_{\beta,\beta}^{l-1}(x, x') K_{\alpha+\beta, \alpha'+\beta}^{l-1}(x, x') \right] \\ &\equiv \frac{\sigma_w^4}{4} \delta_{\alpha\alpha'} [K^{l-1} \star K^{l-1}]_{\alpha, \alpha'}(x, x'), \end{aligned} \quad (2.22)$$

so that we can explicitly express $K_{\alpha\alpha'}^L(x, x')$ via $K_{\alpha\alpha'}^0(x, x')$ as follows

$$K_{\alpha\alpha'}^L(x, x') = \left(\frac{\sigma_w^2}{2} \right)^{2L} \delta_{\alpha\alpha'} \left[\underbrace{K^0 \star K^0 \star \dots \star K^0}_{2^L \text{ times}} \right]_{\alpha, \alpha'}(x, x'), \quad \alpha, \alpha' \neq 0. \quad (2.23)$$

For $K_{00}^l(x, x')$ we have the recursive relation:

$$K_{00}^l(x, x') = \frac{\sigma_w^4}{4} \left[\sum_{\beta} K_{\beta,\beta}^{l-1}(x, x) \sum_{\eta} K_{\eta,\eta}^{l-1}(x', x') + \sum_{\beta} \bar{K}_{\beta,\beta}^{l-1}(x, x') K_{\beta,\beta}^{l-1}(x, x') \right]. \quad (2.24)$$

Unfortunately, we do not have an explicit solution to this recursion.

According to [15, 16], the last convolutional layer has to produce invariant with respect to transformations of the input data. In the Fourier space formulation, that means using only zero mode of the activation:

$$z_i^{L+1}(x) = \sum_{j=1}^{L+1} \omega_{ij}^L y_{j,0}^L(x), \quad (2.25)$$

The zero modes have real values (see (2.5)); therefore if ω_{ij}^L is real, all the quantities in (2.25) are also real. The respective covariance reads:

$$\begin{aligned} K_{00}^{L+1}(x, x') &= \frac{1}{n^{L+1}} \sum_{i=1}^{L+1} \bar{y}_{i,0}^{L+1}(x) y_{i,0}^{L+1}(x') \\ &= \frac{\sigma_w^4}{4} \left[\sum_{\beta} K_{\beta,\beta}^L(x, x) \sum_{\eta} K_{\eta,\eta}^L(x', x') + \sum_{\beta} \bar{K}_{\beta,\beta}^L(x, x') K_{\beta,\beta}^L(x, x') \right], \end{aligned} \quad (2.26)$$

and also obviously real. Thus the covariance for the equivalent Gaussian process has the form

$$\tilde{\Gamma}^{L+1}(x, x') = \frac{\sigma_w^2}{2} K_{00}^{L+1}(x, x') \otimes \mathbb{1}_{n^{L+2}}, \quad (2.27)$$

and the output of the $SO(2)$ -equivariant convolutional neural network can be expressed as a Gaussian process in terms of its input,

$$z_i^{L+1}(x) \sim \mathcal{GP} \left(0, \tilde{\Gamma}^{L+1} \right). \quad (2.28)$$

Proving this was the main goal of this work.

3. Conclusion

In this work we have derived the many-channel limit for CNNs with $SO(2)$ symmetry in the Fourier space with explicit equivariance at each step of the derivation and calculated the corresponding equivariant GP kernel in the case of Fourier space quadratic nonlinearity. All the subtleties and mathematically rigorous proofs for the expressions obtained are quite similar to the case of classical CNNs and can be found in [10].

In this paper we considered only scalar activations as in [15]. Many applications require using more general vector-valued neuron activations with the appropriate properties of equivariance. Such NNs are described in the framework of steerable convolutional neural networks [13, 27] which use induced representations. A derivation of the relationship of such CNNs with the appropriately defined vector-valued equivariant GPs [21] will be presented in a subsequent work.

Acknowledgments

This work was carried out in the framework of R&D State Assignment No.115041410196.

References

- [1] R.M. Neal, *Priors for infinite networks*, in *Bayesian Learning for Neural Networks*, pp. 29–53, Springer (1996).
- [2] R.M. Neal, *Bayesian learning for neural networks*, vol. 118, Springer Science & Business Media (2012).
- [3] Jaehoon Lee, Jascha Sohl-Dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz and Yasaman Bahri, *Deep neural networks as gaussian processes*, in *International Conference on Learning Representations*, 2018, <https://openreview.net/forum?id=B1EA-M-0Z>.
- [4] A.G.G. Matthews, M. Rowland, J. Hron, R.E. Turner and Z. Ghahramani, *Gaussian process behaviour in wide deep neural networks*, *arXiv preprint arXiv:1804.11271* (2018) .
- [5] O. Cohen, O. Malka and Z. Ringel, *Learning curves for deep neural networks: a gaussian field theory perspective*, *arXiv preprint arXiv:1906.05301* (2019) .
- [6] G. Naveh, O. Ben-David, H. Sompolinsky and Z. Ringel, *Predicting the outputs of finite networks trained with noisy gradients*, *arXiv preprint arXiv:2004.01190* (2020) .
- [7] S. Yaida, *Non-gaussian processes and neural networks at finite widths*, in *Mathematical and Scientific Machine Learning*, pp. 165–192, PMLR, 2020.
- [8] J. Halverson, A. Maiti and K. Stoner, *Neural networks and quantum field theory*, *arXiv preprint arXiv:2008.08601* (2020) .
- [9] I. Goodfellow, Y. Bengio, A. Courville and Y. Bengio, *Deep learning*, vol. 1, MIT press Cambridge (2016).

- [10] R. Novak, L. Xiao, J. Lee, Y. Bahri, G. Yang, J. Hron et al., *Bayesian deep convolutional networks with many channels are gaussian processes*, *arXiv preprint arXiv:1810.05148* (2018) .
- [11] A. Garriga-Alonso, C.E. Rasmussen and L. Aitchison, *Deep convolutional networks as shallow gaussian processes*, *arXiv preprint arXiv:1808.05587* (2018) .
- [12] T. Cohen and M. Welling, *Group equivariant convolutional networks*, in *International conference on machine learning*, pp. 2990–2999, 2016.
- [13] T.S. Cohen and M. Welling, *Steerable cnns*, *arXiv preprint arXiv:1612.08498* (2016) .
- [14] T.S. Cohen, M. Geiger, J. Köhler and M. Welling, *Spherical cnns*, *arXiv preprint arXiv:1801.10130* (2018) .
- [15] R. Kondor and S. Trivedi, *On the generalization of equivariance and convolution in neural networks to the action of compact groups*, in *International Conference on Machine Learning*, pp. 2747–2755, PMLR, 2018.
- [16] R. Kondor, Z. Lin and S. Trivedi, *Clebsch–gordan nets: a fully fourier space spherical convolutional neural network*, in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds., vol. 31, Curran Associates, Inc., 2018, <https://proceedings.neurips.cc/paper/2018/file/a3fc981af450752046be179185ebc8b5-Paper.pdf>.
- [17] S. Ravanbakhsh, J. Schneider and B. Poczos, *Equivariance through parameter-sharing*, in *International Conference on Machine Learning*, pp. 2892–2901, PMLR, 2017.
- [18] C. Esteves, *Theoretical aspects of group equivariant neural networks*, *arXiv preprint arXiv:2004.05154* (2020) .
- [19] S.O. Ayat, M. Khalil-Hani, A.A.-H. Ab Rahman and H. Abdellatef, *Spectral-based convolutional neural network without multiple spatial-frequency domain switchings*, *Neurocomputing* **364** (2019) 152.
- [20] S. Li, K. Xue, B. Zhu, C. Ding, X. Gao, D. Wei et al., *Falcon: A fourier transform based approach for fast and secure convolutional neural network predictions*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8705–8714, 2020.
- [21] P. Holderrieth, M. Hutchinson and Y.W. Teh, *Equivariant conditional neural processes*, *arXiv preprint arXiv:2011.12916* (2020) .
- [22] L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. Schoenholz and J. 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*, vol. 80 of *Proceedings of Machine Learning Research*, (Stockholmsmässan, Stockholm Sweden), pp. 5393–5402, PMLR, 10–15 Jul, 2018, <http://proceedings.mlr.press/v80/xiao18a.html>.

- [23] N. Vilenkin, *Special functions and the theory of group representations*, vol. 22, American Mathematical Soc. (1978).
- [24] R. Goodman and N.R. Wallach, *Symmetry, representations, and invariants*, vol. 255, Springer (2009).
- [25] V. Nair and G.E. Hinton, *Rectified linear units improve restricted boltzmann machines*, in *Icml*, 2010.
- [26] A. Lapidoth, *A foundation in digital communication*, Cambridge University Press (2017).
- [27] T.S. Cohen, M. Geiger and M. Weiler, *Intertwiners between induced representations (with applications to the theory of equivariant neural networks)*, *arXiv preprint arXiv:1803.10743* (2018) .