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

Andrey Demichev

SINP MSU

DLCP-2021, Moscow

The Problem Context

- The general topic within which this work was carried out: establishing **relationships between various methods** of machine learning (ML)
  - ultimate goal $=$ a better **theoretical understanding** of these methods and their improvements
- In particular, a correspondence has recently been established between the **appropriate asymptotics** of deep neural networks (**DNNs**), including convolutional ones (**CNNs**), and the ML method based on **Gaussian processes** (**GPs**)
- Gaussian processes are mathematically equivalent to free (Euclidean) quantum field theory (**QFT**) $\Rightarrow$ potential for using a broad range of QFT methods for analyzing DNNs

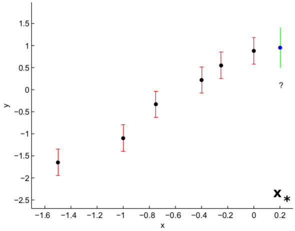
Posing the Problem and Main Result

- An important feature of CNNs is their equivariance (*consistency*) with respect to the symmetry transformations of the input data
- In this work, we have established a **relationship** between the **many-channel limit** of equivariant **CNNs** and the corresponding equivariant **Gaussian processes (GPs)**, and hence the **QFT** with the appropriate **symmetry**
- The approach used provides **explicit equivariance** at each stage of the derivation of the relationship

Gaussian Processes for Machine Learning: Example of Regression (1/2)

- GP ML $\subset$ **kernel method** approach in ML
- In the GP regression rather than claiming $f(x, \theta)$ relates to some specific models
  - e.g., linear, quadratic or even non-polynomial

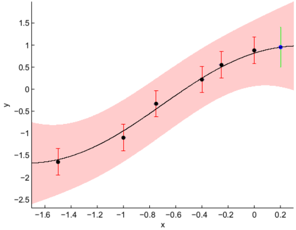
  one can consider **every** possible function that matches data
- but in order that $f(x, \theta)$ be **not too wiggly** (overfitting, *etc.*) $\Rightarrow$ **covariance matrix**
  - to ensure that values that are **close** together in **input** space will produce **output** values that are close together



- An example of simple regression task (*Ebden, arXiv:1505.02965*):
  - given noisy data points $\Rightarrow$ estimating the value at additional point $x_* = 0.2$

Gaussian Processes for Machine Learning: Example of Regression (2/2)

- GP assumes that data set $p(y_1, \ldots, y_N)$ is jointly Gaussian, with some **mean** and **covariance** $k(x_i, x_j; \theta) \equiv$ positive definite kernel function
- using a number of **nice GP properties**, including
  - conditional Gaussian = Gaussian
  - marginal Gaussian = Gaussian
  - integrability
- + some rather lengthy matrix algebra
- one **can find** $\sim p(y_* | x_*, x, y)$
- ML: **optimization** of θ in $k(x_i, x_j; \theta)$ using Bayes' theorem



- Result of the GP-regression (*Ebden, arXiv:1505.02965*):
  - solid line: mean of y_* for 1000 values of x_*
  - shaded: 95% confidence interval

Fully-Connected Neural Networks $\Leftrightarrow$ GPs

- R.M.Neal (1996,2012): the function defined by a **single-layer** fully-connected NN with
  - **infinitely many hidden units** (= *shallow and* ∞*-wide*)
  - *i.i.d.* zero-mean weights and biases as network prior

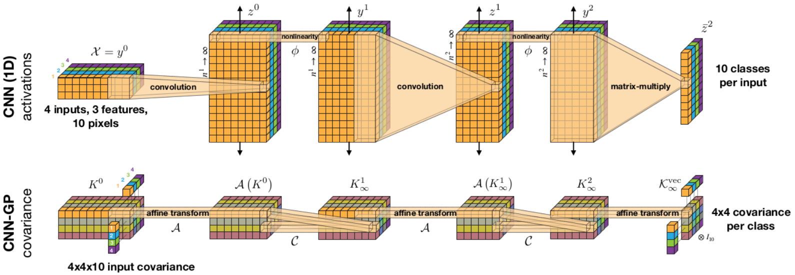
  is **equivalent** to a GP
- J.Lee et al (2018), A.G.Matthews et al (2018): extended these results to **arbitrarily deep fully-connected NN** with **infinitely many hidden units** in each layer
  - provide an explicit form for the prior over functions encoded by NN architectures and initializations
  - $\Rightarrow$ analytical investigation and means for a theoretical understanding of DL, *e.g.*:
    - O.Cohen (2019) et al: predictions for learning curves of DNNs trained on regression problems
    - G. Naveh (2020) et al: predictions of the outputs of some finite networks with high accuracy

Finite NNs $\Leftrightarrow$ GPs

- in practice one is interested in networks with **finite width** N:
  - It is supposed (not rigorously proven so far) 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.*, S.Yaida (2020)
  - It is worth noting: 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 (QFT)**,
    - J.Halverson et al (2020): experimental evidences for the (NGPs/perturbative QFT) $\Leftrightarrow$ (finite-width FCNNs) relationship

Convolutional Neural Networks (CNNs) ⇔ GPs

- fully-connected networks (FCNNs) are rarely used in practice
- CNN ⇒ **localized** filter, essentially **not** very wide!
- R.Novak et al (2018), A.Garriga-Alonso et al (2018): if each hidden layer has an **infinite number** of convolutional **filters** (that is infinite number of **channels**), the **CNN** prior is **equivalent** to a **GP**



The figure is borrowed from R.Novak et al (2018)

Step aside: equivariance in CNNs (1/2)

- Well-known fact: usual CNNs are **translational equivariant**

- Recent years: **huge** activity to extend this to **other symmetries**
  - *e.g.*, rotations in 2D & 3D, Euclidean motions, Lorentz group, *etc*
  - works by Kondor, Trivedi, Cohen, Welling, Esteves, Ravanbakhsh,... and *many others*
- the main ingredient of these extensions is appropriate **generalization of the convolution operation** from plane grids to other homogeneous spaces and even to arbitrary manifolds

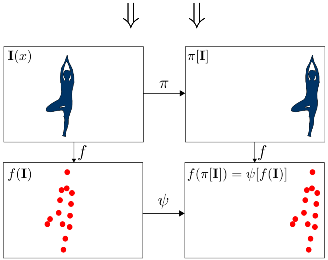
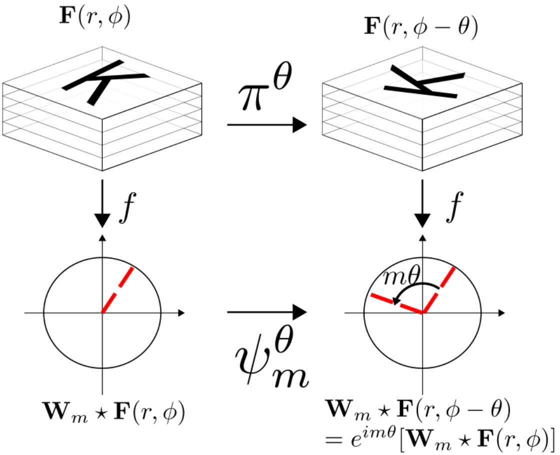


Illustration of translational equivariance of classical CNNs

The figure is borrowed from D.E.Worrall et al (2017)

Step aside: equivariance in CNNs (2/2)



A demonstration of the meaning of equivariance (2D rotational symmetry)

The figure is borrowed from D.E.Worrall et al (2017)

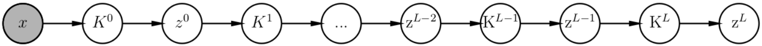
A note on the terminology

Please do not confuse the two notions that sound somewhat similar:

- **equivariance** $\sim$ consistency with symmetry transformations
- **covariance** $\sim$ 2d moment of a distribution

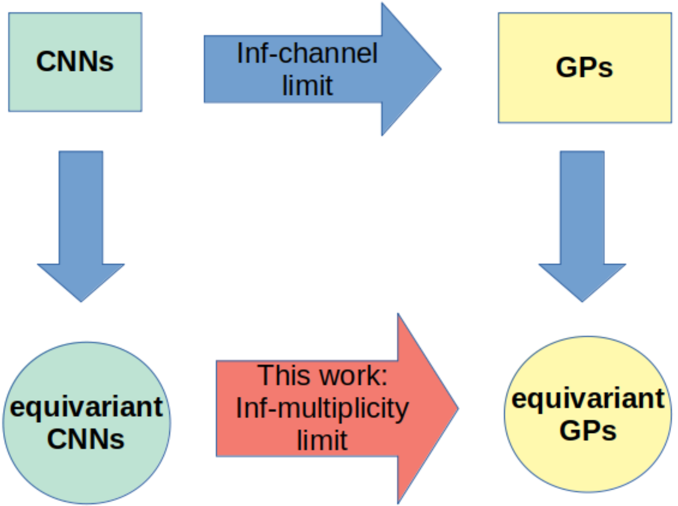
Equivariant CNN with Infinite Number of Channels = Equivariant GPs (1/4)

- All the preceding seminal works on the CNN-GP relationship **did not take into account equivariance**
  - neither generalized nor even **explicit** translational equivariance
- *On the other hand*, there exists investigations of **equivariant GPs** (*e.g.*, P.Holderrieth et al (2020)) but **without** established relations with CNNs in the appropriate limit
  - **The present work is intended to fill the gap between equivariance of CNNs and that of the corresponding GPs**
- the method constituents are
  - layer-by-layer derivation of GP covariances in the many-channel limit by using the law of large numbers that results in the **recursive relation** for the top-layer covariance
  - keeping **explicit equivariance** at each step of the derivation



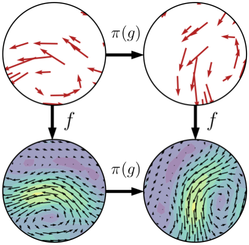
The figure is borrowed from J.Lee et al (2018)

Equivariant CNN with Infinite Number of Channels = Equivariant GPs (2/4)



Equivariant CNN with Infinite Number of Channels = Equivariant GPs (3/4)

- the **main question** in our work is how to deal with **vector-valued** functions
- the point is that such vectors (of **finite** dimensionality) are also treated as channels, so the question is how one can go to the **infinite**-channel limit
- **our solution** is based on using the so called steerable CNNs (T.Cohen & M.Welling (2016)) which in turn heavily use induced representations of symmetry groups
- all-in-all this allows us to **separate channels** indices in **two categories**:
  1. the indices that numerate the vector components within an *irrep* and used to describe their transformations under matrix representations of a symmetry group;
  2. the indices that numerate different irreducible representations (of the same or different types);



from P.Holderrieth et al (2020)

Equivariant CNN with Infinite Number of Channels = Equivariant GPs (4/4)

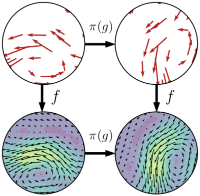
- the 2d type of the indices are **not restricted** and can be used for the **limiting transition to the corresponding GP**
- as result we obtain the **equivariant GP** as **the limit of (steerable) CNNs** with the covariance

$$K(\vec{x}, \vec{x'}) = K(\vec{x} - \vec{x'}, \vec{0}) \equiv \widehat{K}(\vec{x} - \vec{x'})$$

$$\widehat{K}(R\vec{x}) = \rho(R)\widehat{K}(\vec{x})\rho(R)^T$$

R = a transformation; $\rho(R)$ = matrix irrep

- these relations **provide the required equivariance** $\Rightarrow \Rightarrow \Rightarrow$
- and thereby **fill the gap** between **many-channel** CNNs and **equivariant GP** introduced in P.Holderrieth et al (2020)



The figure is borrowed from P.Holderrieth et al (2020)

Example of (recursion) relations for the GP kernel

- for the rotation equivariant CNN and a specific choice of nonlinearity (quadratic nonlinearity in the Fourier space)
- Fourier components of the NN-GP kernel (Gaussian covariance) are expressed via data covariance K^0 as follows

$$K^L_{\alpha\alpha'}(x,x') = \left(\frac{\sigma_w^2}{2}\right)^{2^L} \delta_{\alpha\alpha'} \Big[\underbrace{K^0 \star K^0 \star \cdots \star K^0}_{2^L \text{ times}}\Big]_{\alpha,\alpha'}(x,x') \qquad \alpha,\alpha' \neq 0$$

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

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

- All the terms transforms according to SO(2) irreps ⇒ **explicit equivariance**

Conclusion

- Currently there exists rather promising **new trend** in ML based on the relationship between FCNN/CNNs and GPs
  - many related subtopics, *e.g.*, signal propagation in NNs, learning curve, QFT methods in ML
- In this work we have derived the many-channel limit for CNNs with **symmetry** on Euclidean plane (translations+**rotations**)
  - with explicit equivariance at each step of the derivation
  - calculated the corresponding equivariant GP kernel in the case of specific nonlinearities
- thereby **filled the gap** between many-channel **equivariant** CNNs and independently introduced **equivariant** GP
- many subtleties and mathematically rigorous proofs were dropped in the report but essentially they go in parallel with the case of classical (translationally equivariant) CNNs