Probabilistic equivariance and convolutional neural networks

Sean La

December 16, 2019

1 Background

Deep learning has revolutionized machine learning due to its great accuracy and generalizability on large, high dimensional datasets. The convolutional neural network (CNN) architecture has become the mainstay for constructing computer vision platforms. The chief design principle of CNNs is the incorporation of convolutional layers that allow the network to detect certain features in the input image. In particular, these convolutional layers are equivariant under translation of the image, in the sense that translation of the input image causes the activations of the convolutional layers to translate in the same way. This design feature is the key reason why convolutional neural networks have achieved excellent generalizability on a variety of datasets.

In the last decade, there has been much work in expanding mathematical understanding of deep learning algorithms. Machine learning algorithms in general can be formalized under two separate points of views: they can be viewed as deterministic functions $f : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{X}$ is the set of input features and $\mathcal{Y}$ is the set of labels, or they can be viewed from a probabilistic point of view so that $Y = f(\eta, X)$, where X and Y are the input and output random variables, and η is some outsourced noise random variable. Viewing deep learning algorithms deterministically, Kondor and Trivedi proved that f containing a convolution operation of the form $(h * g)(u) = \int h(uv^{-1})g(v)\, dv$ is both necessary and sufficient for the layers of a neural network $y = f(x)$ to be equivariant under the action of a compact group G[**kondor2018generalization**].

On the other hand, when viewed from a probabilistic point of view, Bloem-Reddy and Teh provide necessary and sufficient conditions for a neural network $Y = f(\eta, X)$ to be invariant or equivariant under the action of a compact group [**bloemreddy2019probabilistic**]. The conditions posed by Bloem-Reddy and Teh on the structure of $Y = f(\eta, X)$ are quite general and do not, by themselves, restrict f to be a function involving convolution, as is the case in of the results of Kondor and Trivedi. Then, a natural question to ask is: *what further conditions on $Y = f(\eta, X)$ are needed so that f is restricted to functions involving convolutions?*

In this project, we answer this question by providing a condition that is both simple and intuitive: linearity of f in the second argument (i.e. $f(\eta, X_1 + X_2) = f(\eta, X_1) + f(\eta, X_2)$).

The project report proceeds as follows. First, we give a brief overview of group theory that will be essential for understanding the main result of this project. Then, we summarize the main results of [**bloemreddy2019probabilistic**]. After that, we provide key definitions and theorems from [**kondor2018generalization**] that will be useful for proving our main result. As the main scholarly contribution of this project, we prove that linearity of $Y = f(\eta, X)$ in the second argument, along with equivariance of X and Y, are sufficient for $Y = X * \chi(\eta)$, i.e. Y is a convolution involving X. Finally, we provide open questions and future research directions.

1.1 A gentle introduction to group theory

The results of both [**bloemreddy2019probabilistic**] and [**kondor2018generalization**] are very mathematics-heavy and involve results from a wide variety of fields of mathematics, namely group theory, fourier analysis, representation theory, and of course, probability theory. In this report, we assume the reader has a good working knowledge of probability, but it may be too much to ask for a background in these other subjects as well. So then, before we get into the thick of things, we give a brief overview of group theory, and how it relates to symmetry, which was derived from [**RomanSteven2012FoGT**]. Note that we will not cover fourier analysis nor representation theory, as they are not necessary to understand the main definitions and theorems of [**bloemreddy2019probabilistic**] or [**kondor2018generalization**].

Group theory is the study of - *you guessed it!* - mathematical objects called *groups*, which we define below.

Definition 1. *A group is a set of elements G along with a binary operator $\cdot : G \times G \to G$ that satisfies the following properties:*

(a) (Closure under the binary operator) For every $u, v \in G$, we have $u \cdot v \in G$.

(b) (Existence of an identity element) There exists an element $e \in G$ (which we call the identity element) so that for every $u \in G$, $e \cdot u = u \cdot e = u$, and

(c) (Existence of inverses) For every $u \in G$, there exists its inverse $u^{-1} \in G$ so that $u \cdot u^{-1} = u^{-1} \cdot u = e$.

For succintness, we denote a group by the pair $(G, \cdot)$.

Example 1. An example of a group is the integers equipped with addition, i.e. $(\mathbb{Z}, +)$ is a group. In this case, 0 is the identity since $z + 0 = z$ for every $z \in \mathbb{Z}$, and every element in $\mathbb{Z}$ has an inverse $-z := z^{-1}$ so that $z + (-z) = 0$. Note that the integers equipped with multiplication is not a group, since the only integer that has an inverse that is also an integer is 1.

Groups are a natural tool to formalize and study the mathematics of symmetry and symmetry-preserving transformations.

Example 2. Consider the set of rotations of a triangle in two dimensional space that are integer multiples of 120°, and call it G. Clearly, if we rotate a triangle by any multiple of 120°, the image of the triangle will be preserved; in other words, these transformations preserve the symmetry of the triangle. Notice that we can view the set of rotations by multiples of 120° as a group in and of itself.

1. G is closed under composition of rotations. If we rotate a triangle by $n120°$ and then rotate it again by $m120°$, we have rotated it in total by $(n + m)120°$, another integer multiple of 120°. Therefore, composition of rotations is itself a rotation.
2. G has an identity element, the $0 \cdot 120°$ rotation.
3. Every rotation $n120°$ has an inverse rotation $-n120°$.

Example 2 gives an example of how a group can be formed by a set of symmetry-preserving *transformations* of some object - in other words, the group *acts* on this object, in a way. We formalize this notion of *group action* below.

Definition 2. *Consider a group $(G, \cdot)$ and a set $\mathcal{X}$. A group action $\mathbb{T} : G \times \mathcal{X} \to \mathcal{X}$ is a function that satisfies the following properties*

1. *$\mathbb{T}(e, x) = x$ for all $x \in \mathcal{X}$, where e is the identity element, and*
2. *$\mathbb{T}(u \cdot v, x) = \mathbb{T}(u, \mathbb{T}(v, x))$ for all $u, v \in G$, and $x \in \mathcal{X}$.*

When the context is clear, we will write $gx := \mathbb{T}_g(x) := \mathbb{T}(g, x)$.

From here, we define two key concepts that are important in understanding the results of [**bloemreddy2019probabilistic**] and [**kondor2018generalization**].

Definition 3. *Consider a group G and set $\mathcal{X}$ upon which it acts through its group action $\mathbb{T}$. Take an element $x \in \mathcal{X}$.*

1. *The orbit of x is the subset of $\mathcal{X}$ defined by*

$$G \cdot x := \{\mathbb{T}(g, x) \in \mathcal{X} : g \in G\}.$$

2. *The stabilizer of x is the subset of G defined by*

$$G_x := \{g \in G : \mathbb{T}(g, x) = x\}.$$

The orbit of an element x with respect to a group G can be thought of as all elements in $\mathcal{X}$ that can be reached from x using G. One interesting property of stabilizers is that they form a *subgroup* of G - they satisfy all the properties described in definition 1. The proof of this is left as an exercise to the reader.

We now describe the penultimate concept that will be important in understanding the main result of this project - the idea of a quotient group.

Definition 4. *Consider a group $(G, \cdot)$ and a subgroup H of G. A left coset of H is the set $gH = \{g \cdot h : h \in H\}$ for some $g \in G$. Then we define the left quotient group*

$$G/H := \{gH : g \in G\}$$

which itself forms a group under the binary operation $\cdot$ defined by

$$g_1 H \cdot g_2 H := (g_1 \cdot g_2)H.$$

Notice that if $g_2 \in g_1 H$, we can equivalently write $g_2 H = g_1 H$, so there does not necessarily exist a unique representation of the elements of G/H; the choice of representation is up to us. No matter; it is easy to see that the binary operation on G/H borrowed from G is still well-defined.

Finally, we can view groups as *topological objects* in similar vein as $\mathbb{R}$ or $\mathbb{C}$.

Definition 5. *A topological space is defined as a set $\mathcal{X}$ and a collection of open sets $\mathcal{C}$. $(\mathcal{X}, \mathcal{C})$ is then said to be compact if for every collection of open sets $\mathcal{C}' \subseteq \mathcal{C}$ that covers $\mathcal{X}$, i.e.*

$$X = \cup_{C \in \mathcal{C}'} C$$

there exists a finite subcover $C'' \subseteq C'$ so that

$$X = \cup_{C \in \mathcal{C}''} C.$$

A compact group is then a group $(G, \cdot)$ equipped with a collection of open sets $\mathcal{C}$ so that G is compact with respect to $\mathcal{C}$.

This concludes are brief foray into group theory. With these tools, we are now able to grasp the main results of [**bloemreddy2019probabilistic**] and [**kondor2018generalization**].

1.2 Probabilistic symmetry

Symmetry is a natural concept in probability. For example, iid random variables exhibit an obvious symmetry in that the joint distribution of a collection of iid random variables $(X_1, \ldots, X_n)$ does not depend on the order in which we write the random variables, since

$$\mathbb{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i) = \mathbb{P}(X_{\sigma(1)}, \ldots, X_{\sigma(n)})$$

for any permutation σ on $\{1, \ldots, n\}$. In other words, this probability distribution is invariant under permutations of the random variables, when they are iid.

An even stronger concept is *exchangeability*. We say that a sequence of random variables $(X_1, \ldots, X_n)$ is *exchangeable* if

$$\mathbb{P}(X_1, \ldots, X_n) = \mathbb{P}(X_{\sigma(1)}, \ldots, X_{\sigma(n)})$$

for any permutation σ. Note that in the definition of exchangeability, we do not assume that the the random variables X_i are iid.

We see that the joint distribution of an exchangeable sequence is *invariant* when the exchangeable sequence is acted upon by the symmetry group S_n, the group of all permutations of $\{1, \ldots, n\}$. We can generalize this idea by considering the group action of a group G on a set $\mathcal{X}$ of random variables, as per definition 2. Moreover, we can further expand the idea of group action into the language of probability by considering group actions that are G-measurable functions. We take the following definition from [**bloemreddy2019probabilistic**]:

Definition 6. *A group G along with a σ-algebra $\sigma(G)$ is measurable if the group operations*

- *inversion: $g \mapsto g^{-1}$, and*
- *composition: $(g, h) \mapsto g \cdot h$*

are measurable functions with respect to $\sigma(G)$-measurable. Moreover, we say that G acts measurably on $\mathcal{X}$ if $\mathbb{T}$ is a measurable function from $\sigma(G) \times \mathcal{B}_{\mathcal{X}} \to \mathcal{B}_{\mathcal{X}}$.

Bloem-Reddy and Teh [**bloemreddy2019probabilistic**] provide two definitions of probabilistic symmetry that are especially relevant to machine learning applications:

Definition 7. *Consider random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ and a group G which acts measurably on both $\mathcal{X}$ and $\mathcal{Y}$. The conditional distribution $\mathbb{P}_{Y|X}$ is then said to be*

1. *G-invariant if $Y|X \overset{\text{d}}{=} Y|gX$,*
2. *G-equivariant if $Y|X \overset{\text{d}}{=} gY|gX$.*

It is useful to assume that the marginal distribution of X is invariant under G, so that $\mathbb{P}(X) = \mathbb{P}(gX)$ for all $g \in G$. This assumption, along with definition 7, implies that the joint distributions are invariant

$$\mathbb{P}_{X,Y}(X, Y) = \mathbb{P}_{X,Y}(gX, Y),$$

or equivariant

$$\mathbb{P}_{X,Y}(X, Y) = \mathbb{P}_{X,Y}(gX, gY),$$

respectively.

Given two random variables X and Y whose conditional probability distribution $\mathbb{P}_{Y|X}$ is G-invariant or G-equivariant, a natural question to ask is: how is Y related to X?

Since we are working in a machine learning setting where X is the input data and Y is the output of a layer in some neural network, it would be useful to be able to write Y as some sort of function involving X. *Noise outsourcing* allows us to do just that - given two random variables X and Y in Borel spaces, there exists a measurable function f so that

$$(X, Y) \overset{\text{a.s.}}{=} (X, f(\eta, X))$$

for some $\eta \sim \text{Unif}[0, 1]$. η can be viewed as the variation in Y that cannot be explained by X.

The version of this result that is used extensively by [**bloemreddy2019probabilistic**] involves the situation where there exists a $\mathcal{B}(\mathcal{X})/\mathcal{B}(\mathcal{S})$-measurable statistic $S : \mathcal{X} \to \mathcal{S}$ so that X and Y are conditionally independent given this statistic.[1] A slightly-paraphrased of this lemma from [**bloemreddy2019probabilistic**] follows:

Lemma 1. *Let X and Y be random variables with joint distribution $\mathbb{P}_{X,Y}$. Let $\mathcal{S}$ be a standard Borel space and $S : \mathcal{X} \to \mathcal{S}$ a measurable map. Then X and Y are conditionally independent given S if and only if there is a measurable function $f : [0, 1] \times \mathcal{S} \to \mathcal{Y}$ so that*

$$(X, Y) \overset{\text{a.s.}}{=} (X, f(\eta, S(X)))$$

where $\eta \sim \text{Unif}[0, 1]$ and $\eta \perp\!\!\!\perp X$.

Another tool that [**bloemreddy2019probabilistic**] uses is the notion of *maximal invariants*.

Definition 8. *Consider a space $\mathcal{X}$ and a group G which acts on $\mathcal{X}$. A maximal invariant M is a function that satisfies $M(x_1) = M(x_2)$ for $x_1, x_2 \in \mathcal{X}$ implies $x_2 = \mathbb{T}_g(x_1)$ for some $g \in G$.*

If one looks closely at the above definition, it easy to see that maximal invariants take on the different values on *orbits* (see definition 3 for a quick refresher) of elements in $\mathcal{X}$ with respect to G. It is then straightforward to create a maximal invariant for any group G and space $\mathcal{X}$ by just defining a function S that takes on unique values on each orbit.

Using noise outsourcing and maximal invariants, [**bloemreddy2019probabilistic**] determines the following relationship between random variables X and Y when their joint probability $\mathbb{P}_{Y|X}$ is G-invariant:

[1] Edit 14 December 2019: changed $\sigma(\mathcal{X})$ and $\sigma(\mathcal{S})$ to $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\mathcal{S})$

Theorem 1. *Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be random variables from Borel spaces $\mathcal{X}$ and $\mathcal{Y}$. Assume that $\mathbb{P}_X$ is G-invariant, and let $M : \mathcal{X} \to \mathcal{S}$ be a maximal invariant where $\mathcal{S}$ is another Borel space. Then $\mathbb{P}_{Y|X}$ is G-invariant if and only if there exists a measurable function $f : [0,1] \times \mathcal{S} \to \mathcal{Y}$ such that*

$$(X, Y) \overset{\text{a.s.}}{=} (X, f(\eta, M(X)))$$

with $\eta \sim \text{Unif}[0,1]$ and $\eta \perp\!\!\!\perp X$.

To prove a similar result for the case when $\mathbb{P}_{Y|X}$ is G-equivariant, [**bloemreddy2019probabilistic**] defines a function similar to maximal invariants for equivariant symmetry:

Definition 9. *Consider a set $\mathcal{X}$ and a group G which acts upon $\mathcal{X}$. A function $\tau : \mathcal{X} \to G$ is said to be a maximal equivariant if it satisfies*

$$\tau(gx) = g \cdot \tau(x).$$

for $g \in G$ and $x \in \mathcal{X}$.

[**bloemreddy2019probabilistic**] then proves the following

Theorem 2. *Let G be a compact group acting measurably on Borel spaces $\mathcal{X}$ and $\mathcal{Y}$. Let $\tau : \mathcal{X} \to G$ be a maximal equivariant. Consider random elements $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ such that the stabilizers of X and Y (see definition 3 for a reminder) satisfy $G_X \subseteq G_Y$ almost surely. Suppose $\mathbb{P}_X$ is G-invariant. Then $\mathbb{P}_{Y|X}$ is G-equivariant if and only if there exists a measurable function $f : [0,1] \times \mathcal{X} \to \mathcal{Y}$ such that*

$$gY \overset{\text{a.s.}}{=} f(\eta, gX)$$

for all $g \in G$, and $\eta \sim \text{Unif}[0,1]$ and $\eta \perp\!\!\!\perp X$.

The utility of theorems 1 and 2 is further enhanced by the following result, which states that there is a straight-forward way to combine equivariant and invariant functions to yield more equivariant or invariant functions.

Proposition 1. *Let X, Y, Z be random variables such that $X \perp\!\!\!\perp_Y Z$, and G be a group which acts upon X, Y, and Z. Suppose further that $\mathbb{P}_X(X) = \mathbb{P}_X(gX)$ for any $g \in G$.*

1. *If Y is conditionally G-invariant given X, and Z is conditionally G-equivariant given Y, then Z is conditionally G-equivariant given X.*
2. *If Y is conditionally G-equivariant given X, and Z is conditionally G-invariant given Y, then Z is conditionally G-invariant given X.*

With the multitude of results that are theorems 1, 2 and proposition 1, we now have a methodology to form deep learning algorithms that are invariant or equivariant under the action of a group G:

1. Find a class of datasets $\mathcal{X}$ whose distribution $\mathbb{P}_{\mathcal{X}}$ is G-invariant for some group G.
2. For each layer ℓ in the deep learning algorithm, determine whether it should be invariant or equivariant under the action of G, and then construct appropriate maximal invariants or equivariants.
3. Model each layer Y_ℓ as a noise-outsourced function of the previous layer so that $f : [0,1] \times \mathcal{Y}_{\ell-1} \to \mathcal{Y}_\ell$ with
$$Y_\ell = f(\eta, Y_{\ell-1})$$
where $\eta \perp\!\!\!\perp Y_{\ell-1}$.

Notice in the statement of theorem 2, the only condition on f is that it is some measurable function from $[0,1] \times \mathcal{X}$ to $\mathcal{Y}$; no other conditions on what classes of functions make for acceptable f are given. An open question is: what further conditions on f are needed so that f is constrained to be certain classes of functions? In particular, what are conditions are necessary for f to be a convolution? As stated previously, the incorporation of convolutions is a popular way to create equivariant layers in a neural network, and so an answer to this question could prove useful to machine learning practitioners.

1.3 Deterministic symmetry in machine learning

In the deterministic view, a machine learning algorithm is considered to be a function $f : \mathcal{X} \times \Theta \to \mathcal{Y}$ which maps an input object $x \in \mathcal{X}$ and parameters of the algorithm $\theta \in \Theta$ to a label $y \in \mathcal{Y}$, so that $y = f(x; \theta)$. There is no randomness in this formulation, and so we provide no distributional assumptions on $\mathcal{X}$ nor $\mathcal{Y}$.

Each neuron of a neural network is constructed by a linear transformation of the previous layer followed by a linear/non-linear function that is evaluated pointwise; that is, the xth neuron in layer ℓ is of the form

$$f_x^\ell = \sigma\left(b_x^\ell + \sum_y w_{x,y}^\ell f_y^{\ell-1}\right)$$

where

- b_x^ℓ is the bias term for the xth neuron in layer ℓ,
- y denotes the yth neuron in the $(\ell-1)$th layer,
- $w_{x,y}$ are weights that scale the outputs of the neurons $f_y^{\ell-1}$ of the previous layer, and
- σ is a function that is evaluated pointwise, which is often times non-linear such as a sigmoid function.

Kondor and Trivedi [**kondor2018generalization**] prove that any neural network whose layers are *equivariant* under the action of *any compact group* (see definition 5) must necessarily have layers which are convolutions of the previous layer; that is, the linear transformations are of the form

$$\phi_\ell(f^{\ell-1})_{(i,j)} = \sum_{u_1=1}^{w}\sum_{u_2=1}^{w} f^{\ell-1}_{(i-u_1,j-u_2)}\chi_\ell(u_1,u_2)$$

where $f^{\ell-1}$ is previous layer in the neural network, and χ_ℓ is some *filter* function that scales the values of the neurons in the previous layer.

The proof of this result is very involved, drawing upon areas such as fourier analysis, group theory, and representation theory, so we will limit our discussion of Kondor and Trivedi's results. We will, however, discuss key definitions and theorems that will be needed for our main result in this paper.

Kondor and Trivedi take a more abstract point of view in that they consider the layers of a neural network as functions of their index sets of their neurons, and not as functions of the values of the previous layer. For example, let f_ℓ be the ℓth layer of a neural network. $f_\ell(i)$ gives the value of the ith neuron in layer ℓ. Kondor and Trivedi first define a neural network in this slightly more abstract formalism.

Definition 10. *Given an index set $\mathcal{I}$ and a vector space V, $L_V(\mathcal{I})$ denotes the space of functions $\{f : \mathcal{I} \to V\}$. Let $\mathcal{I}_0, \ldots, \mathcal{I}_L$ be a sequence of index sets, $V_0, \ldots, V_L$ vector spaces, $\phi_1, \ldots, \phi_L$ linear maps so that*

$$\phi_\ell : L_{V_{\ell-1}}(\mathcal{I}_{\ell-1}) \to L_{V_\ell}(\mathcal{I}_\ell)$$

and $\sigma_\ell : V_\ell \to V_\ell$ pointwise linear or nonlinear functions. A multi-layer feed-forward neural network (MFFNN) is a sequence of maps $f_0 \mapsto f_1 \mapsto f_2 \mapsto \ldots, \mapsto f_L$, where

$$f_\ell(i) = \sigma_\ell(\phi_\ell(f_{\ell-1})(i))$$

for index $i \in \mathcal{I}_\ell$.

Next, Kondor and Trivedi define group actions of a group G on linear maps in a neural network.

Definition 11. *Consider a function $f \in L_V(\mathcal{I})$ and a group that acts on $\mathcal{I}$ by a group action $\mathbb{T}$. The induced group action of G on f is given by*

$$gf := \mathbb{T}_g(f) := f(\mathbb{T}_g(i)) = f(gi)$$

for $g \in G$ and $i \in \mathcal{I}$.

Definition 12. *Let G be a group and $\mathcal{I}_1, \mathcal{I}_2$ be two sets with corresponding G-actions*

$$T_g : \mathcal{I}_1 \to \mathcal{I}_1,\ T'_g : \mathcal{I}_2 \to \mathcal{I}_2.$$

Let V_1 and V_2 be vector spaces, and $\mathbb{T}$ and $\mathbb{T}'$ be the induced actions of G on $L_{V_1}(\mathcal{I}_1)$ and $L_{V_2}(\mathcal{I}_2)$. We say that a map $\phi : L_{V_1}(\mathcal{I}_1) \to L_{V_2}(\mathcal{I}_2)$ is equivariant with the action of G (or G-equivariant for short) if

$$\phi(\mathbb{T}_g(f)) = \mathbb{T}'_g(\phi(f))\ \forall f \in L_{V_1}(\mathcal{I}_1)$$

for any group element $g \in G$.

Then, Kondor and Trivedi defines what it means for a neural network to be equivariant under the action of a group G.

Definition 13. *Let $\mathcal{N}$ be a feed-forward neural network as defined in Definition 10, and G be a group that acts on each index space $\mathcal{I}_0, \dots, \mathcal{I}_L$. Let $\mathbb{T}^0, \dots, \mathbb{T}^L$ be the corresponding group actions on $L_{V_0}(\mathcal{I}_0), \dots, L_{V_L}(\mathcal{I}_L)$. $\mathcal{N}$ is a G-equivariant feed-forward network if, when the inputs are transformed $f_0 \mapsto \mathbb{T}^0_g(f_0)$ for any $g \in G$, the neurons of the other layers correspondingly transform as $f_\ell \mapsto \mathbb{T}^\ell_g(f_\ell)$.*

Kondor and Trivedi assume that the action of G on any index set $\mathcal{I}_\ell$ is *transitive*, so that it is possible to travel from element in $\mathcal{I}$ to another by a sequence of group actions. That is, they assume that for any $i, j \in \mathcal{I}_\ell$, we have the existence of $g \in G$ so that $j = \mathbb{T}_g(i)$. With this in mind, we can arbitrarily set some $x_\ell \in \mathcal{I}_\ell$ to be the "origin", and *identify* (represent) $\mathcal{I}_\ell$ by the *quotient group* G/G_{x_ℓ} (see definition 4).

Convoluting two functions is a popular tool in engineering fields such as signals engineering, where the convolution involves integration over $\mathbb{C}$. We can extend the notion of convolutions over complex space to convolutions over groups in the following way.

Definition 14. *Let G be a finite or countable group, $\mathcal{I}_\ell$ and $\mathcal{I}_k$ be (left or right) quotient spaces of G, $f : \mathcal{I}_\ell \to \mathbb{C}$, and $g : \mathcal{I}_k \to \mathbb{C}$. Define $f \uparrow^G : G/G_{x_\ell} \to \mathbb{C}$ by*

$$f \uparrow^G (u) := f(\mathbb{T}_u(x_\ell)) = f(x)$$

where $u \in G/G_{x_\ell}$, $x = \mathbb{T}_u(x_\ell)$, and x_ℓ is the "orign" of $\mathcal{I}_\ell$. We define $g \uparrow^G$ similarly. The convolution of f with g is then defined as

$$(f * g)(u) = \sum_{v \in G} f \uparrow^G (u \cdot v^{-1}) g \uparrow^G (v)$$

for some $u \in G$

With this definition, Kondor and Trivedi defines convolutional neural network in this abstract formalism.

Definition 15. *Let G be a compact group and $\mathcal{N}$ an $L+1$ layer feed-forward network in which the ℓth index set identified with G/G_{x_ℓ} where G_{x_ℓ} is the stabilizer of the "origin" index $x_\ell \in \mathcal{I}_\ell$. We say that $\mathcal{N}$ is a G-convolutional neural network (or G-CNN) if the linear maps $\phi_1, \dots, \phi_L$ in $\mathcal{N}$ are generalized convolutions of the form $\phi_\ell(f_{\ell-1}) = f_{\ell-1} * \chi_\ell$ for filter functions χ_ℓ.*

Theorem 3. *Let G be a compact group and $\mathcal{N}$ be an $L+1$ layer feed-forward neural network in which the ℓth index set is identified with $\mathcal{I}_\ell = G/G_{x_\ell}$, where G_{x_ℓ} is the stabilizer of $x_\ell \in I_\ell$. Then $\mathcal{N}$ is equivariant to the action of G in the sense of definition 12 if and only if it is a G-CNN.*

We are now prepared to prove our main result using the tools outlined in this section.

2 Sufficient conditions for convolution structure of $Y = f(\eta, X)$

In this section, we show that equivariant $Y = f(\eta, X)$ being linear in its second argument almost surely, so that

$$f(\eta, a_1X_1 + a_2X_2) = a_1f(\eta, X_1) + a_2f(\eta, X_2),$$

implies that Y is a convolution of X,

$$Y \overset{\text{a.s.}}{=} X * \chi(\eta)$$

for some function $\chi(\eta)$. [2]

First, we define what it means for a function to be *linear*.

Definition 16. [3] *Let E and F be vector spaces on the same field $\mathcal{F}$. A map $T : E \to F$ is linear if for $x_1, x_2 \in E$ and $a_1, a_2 \in \mathcal{F}$, we have*

$$T(a_1x_1 + a_2x_2) = a_1T(x_1) + a_2T(x_2).$$

Now, the main contribution of this project follows.

Theorem 4. *Consider random vectors (X, Y) with index sets $\mathcal{I}_X$ and $\mathcal{I}_Y$ and a compact group G which acts on $\mathcal{I}_X$ and $\mathcal{I}_Y$. Moreover, suppose that $P_{X,Y}$ is G-equivariant. Lastly, suppose that the action of G on both $\mathcal{I}_X$ and $\mathcal{I}_Y$ is transitive. Then if, conditioned on X, $Y \overset{\text{a.s.}}{=} f(\eta, X)$ is a linear function almost surely (i.e. for some almost sure subset of the range of η), then $Y \overset{\text{a.s.}}{=} X * \chi(\eta)$ in the sense of definition 14. More generally, $(X, Y) \overset{\text{a.s.}}{=} (X, X * \chi(\eta))$.*

Proof. Let $F(X) = f(\eta, X)$ be some realization of Y so that $F(\cdot) = f(\eta, \cdot)$ is linear and $gY \overset{\text{a.s.}}{=} f(\eta, gX)$ (the set of all such realizations has probability one since $f(\eta, \cdot)$ is linear almost surely and $gY \overset{\text{a.s.}}{=} f(\eta, gX)$). We prove this theorem by constructing an appropriate MFF-NN as in definition 10, and then invoking theorem 3.

We may consider X to be a function of its index set so that $X(i)$ gives the value of X at component i for $i \in \mathcal{I}_X$. Likewise, we view Y as a function of its index set $\mathcal{I}_Y$ in the same way. Then the linear function F can be seen as a function from the space of functions on $\mathcal{I}_X$ to the space of functions on $\mathcal{I}_Y$. Lastly, we set the function σ to be simply the identity function. Then we define our MFF-NN $\mathcal{N}$ by the sequence of maps

$$X := f_0 \mapsto f_1 =: F(X) = Y$$

where

$$Y(i) = F(X)(i) = \sigma(F(X)(i)).$$

for $i \in \mathcal{I}_Y$. Clearly, this construction matches the description of a MFF-NN given in definition 10.

Now, define the action of group G on X to be the action of G on its index set, so that

$$gX := \mathbb{T}_g(X) := X(gi)$$

for $i \in \mathcal{I}_X$ and $g \in G$. Similarly, define the action of G on Y as the action of G on the its index set $\mathcal{I}_Y$, so that

$$gY := \mathbb{T}'_g(Y) = Y(gi)$$

for $i \in \mathcal{I}_Y$ and $g \in G$. By assumption, we have

$$\mathbb{T}'_g(Y) = gY = F(gX) = F(\mathbb{T}_g(X))$$

for all $g \in G$ since $P_{X,Y}$ is G-equivariant. Therefore, the layer of $\mathcal{N}$ is equivariant under the action of G.

Because the actions of G on both $\mathcal{I}_X$ and $\mathcal{I}_Y$ are transitive, we may identify the index sets of X and Y by quotient groups G/G_x and G/G_y, where G_x and G_y are the stabilizer subgroups of elements chosen to be the origins of $\mathcal{I}_X$ and $\mathcal{I}_Y$, respectively.

[2] Edit 14 December 2019: added homogeneity of scalar multiplication in definition of linearity

[3] Edit 14 December 2019: added homogeneity of scalar multiplication in definition of linearity, and clarified E and F should be vector spaces over the same field.

Therefore, by theorem 3, our MFF-NN $\mathcal{N}$ is a G-CNN so that

$$F(X) = X * \chi$$

for some filter function χ. Letting the value of η to vary over the subset of its range that gives $f(\eta, X)$ to be linear in X, it is easy to see that

$$f(\eta, X) = X * \chi(\eta).$$

Take $\Omega_0 \subseteq \Omega$ such that $f(\eta, X)$ is both linear and equivariant under the action of G. Because every $\omega \in \Omega_0$ gives that $Y(\omega) = X(\omega) * \chi(\eta(\omega))$ and $\mathcal{P}(\Omega_0) = 1$, we have

$$(X, Y) \overset{\text{a.s.}}{=} (X, X * \chi(\eta))$$

which finishes the proof. □

2.1 Theorem 4 in terms of matrix multiplication [4]

We can also formulate theorem 4 in the language of matrix multiplication. We have the following result in linear algebra:

Proposition 2. *Suppose V_1 and V_2 are finite dimensional vector spaces over the same field. A transformation $T : V_1 \to V_2$ is linear if and only if it can be represented as a matrix; that is, for any vector $v \in V_1$, we have*

$$T(v) = Mv$$

for some matrix M with column space in V_2.

Theorem 4 assumes that $f(\eta, X)$ is a linear mapping between the vector spaces in which X and Y lie. Therefore, proposition 2 implies that we can write

$$Y \overset{\text{a.s.}}{=} M(\eta)X$$

where $M(\eta)$ is a random matrix depending on η. This holds true even if $Y \overset{\text{a.s.}}{=} X * \chi(\eta)$, as we can express the filter $\chi(\eta)$ of a convolution as a matrix $M(\eta)$, so that

$$Y \overset{\text{a.s.}}{=} X * \chi(\eta) = M(\eta)X.$$

[4] Edit 14 December 2019: Added this section

3 Open questions and research directions

This research leads to several open questions:

1. We have shown that linearity of $f(\eta, \cdot)$ in the second argument along with equivariance of f are sufficient for $Y = f(\eta, X) = X * \chi(\eta)$ when the group G is compact. Can we expand these results for when the group G is not compact?

   A possible approach to answer this question is by restricting the possible forms of the filter χ.

2. We have shown that the filter function is a function of the noise-outsourced variable η so that $\chi = \chi(\eta)$. From an applications point of view, does adding noise to filters in a convolutional neural network increase its performance? For example, does adding noise to filters act as a form of regularization, similar to how drop-out layers in neural networks serve that purpose?

   Here are some possible ways to incorporate noise into the filter χ:

   - The value of $\eta \in [0, 1]$ serves as a scaling factor of the value of χ.
   - If the value of η is below a certain threshold, then some components of χ is set to 0, so that certain neurons in the previous layer are not included in the current layer.
   - The value of η serves as the probability that certain neurons in the previous layer are dropped, in similar vein to dropout layers.

A Exercises

1. Consider a probability measure μ on a group G that acts on the real numbers $\mathbb{R}$ equipped with the Borel σ-algebra $\mathcal{B}(\mathbb{R})$. For $x \in \mathbb{R}$, define the $\sigma(\mathcal{G})$-measurable function $h_x : G \to \mathbb{R}$ defined by $h_x(g) := f(\mathbb{T}_g(x))$ for some $\mathcal{B}(\mathbb{R})$-measurable function f. Compute $\int_G h_x(g) d\mu(g)$ under the following conditions.

   (a) f is G-invariant; that is, $f(\mathbb{T}_g(x)) = f(x)$ for all g.

   (b) f is G-equivariant, $G = \mathbb{R} \setminus \{0\}$, and $\mathbb{T}_g(x) = gx$.

2. Suppose $Y \perp\!\!\!\perp X$. Provide f so that $Y \overset{\text{d}}{=} f(\eta, X)$ conditioned on X for some $\eta \sim \text{Unif}[0, 1]$.

3. Consider a group $(G, \cdot)$ and a set X upon which it acts through its group action $\mathbb{T}$. Prove that the stabilizer of x defined by
$$G_x := \{g \in G : \mathbb{T}(g, x) = x\}$$
forms a group with the binary operator $\cdot$ borrowed from G.

B Solution to exercises

1. (a) We have

$$\begin{aligned}
\int_G h_x(g)d\mu(g) &= \int_G f(\mathbb{T}_g(x))d\mu(g) && \text{by definition of } h_x \\
&= \int_G f(x)d\mu(g) && \text{by } G\text{-invariance of } f \\
&= f(x)\int_G d\mu(g) && f(x) \text{ is a constant with respect to the integral} \\
&= f(x)\mu(G) \\
&= f(x)
\end{aligned}$$

   (b) Define Y to be a random variable whose distribution is μ. We have

$$\begin{aligned}
\int_G h_x(g)d\mu(g) &= \int_G f(\mathbb{T}_g(x))d\mu(g) \\
&= \int_G \mathbb{T}_g(f(x))d\mu(g) && \text{the action of } G \text{ on the range of } f \text{ is simply the original action } \mathbb{T}_g \\
&= \int_G gf(x)d\mu(g) \\
&= f(x)\int_G g\ d\mu(g) \\
&= f(x)\mathbb{E}[Y]
\end{aligned}$$

2. Because $X \perp\!\!\!\perp Y$ and we assume that $X \perp\!\!\!\perp \eta$, the quantile function Q of Y will do, so that $Y \overset{\text{d}}{=} Q(\eta)$.

3. We will show that $(G_x, \cdot)$ satisfies the properties of a group described in definition 1.

   (a) Consider $u, v \in G_x$. Notice that

$$\mathbb{T}(u \cdot v, x) = \mathbb{T}(u, \mathbb{T}(v, x)) = \mathbb{T}(u, x) = x.$$

   Therefore, $u \cdot v \in G_x$.

   (b) By definition of group action, $\mathbb{T}(e, x) = x$, where e is the identity. Therefore, $e \in G_x$.

   (c) Suppose for contradiction that $g \in G_x$ but $g^{-1} \notin G_x$. Then we have

$$\begin{aligned}
\mathbb{T}(g^{-1}, x) \neq x &\implies \mathbb{T}(g, \mathbb{T}(g^{-1}, x)) \neq \mathbb{T}(g, x) \\
&\implies \mathbb{T}(g \cdot g^{-1}, x) \neq x \\
&\implies \mathbb{T}(e, x) \neq x \\
&\implies x \neq x
\end{aligned}$$

   a contradiction.