Equivariant Representation Learning with Equivariant Convolutional Kernel Networks

Soutrik Roy Chowdhury [1] **Johan A.K. Suykens** [1]

Abstract

Convolutional Kernel Networks (CKNs) were proposed as multilayered representation learning models that are based on stacking multiple Reproducing Kernel Hilbert Spaces (RKHSs) in a hierarchical manner. CKN has been studied to understand the (near) group invariance and (geometric) deformation stability properties of deep convolutional representations by exploiting the geometry of corresponding RKHSs. The objective of this paper is two-fold: (1) Analyzing the construction of group equivariant Convolutional Kernel Networks (equiv-CKNs) that induce in the model symmetries like translation, rotation etc., (2) Understanding the deformation stability of equiv-CKNs that takes into account the geometry of inductive biases and that of RKHSs. Multiple kernel based construction of equivariant representations might be helpful in understanding the geometric model complexity of equivariant CNNs as well as shed lights on the construction practicalities of robust equivariant networks.

1. Introduction

In the past decade deep neural networks, especially convolutional neural networks (CNNs) (LeCun et al., 1989) have achieved impressive results for various predictive tasks, notably in the domains of computer vision (Krizhevsky et al., 2017) and natural language processing. Much success of CNNs in these domains relies on (1) the availability of large scaled labeled and structured data which allow the model to learn huge number of parameters without worrying too much of overfitting, and (2) the ability to model local information of signals (e.g., images) at multiple scales, while also representing the signals with some invariance through pooling operations. The latter property of CNNs have distinguished them from fully-connected networks (Li et al., 2021) in terms of sample efficiency, generalization ability and computational speed, much through its elegant model design. Still, understanding the exact mathematical nature of this invariance as well as the characteristics of the functional spaces where CNNs live are indeed open problems for which multiple constructions and analyses have been provided in past years.

One such construction is of group equivariant CNNs (Cohen & Welling, 2016a) where the translation equivariance of convolutional layers has been generalized to other kinds of symmetries, for e.g., rotations, reflections, etc., thus making CNNs equivariant to more general transformations, where such transformations and corresponding equivariant maps for learning layerwise features are encoded by the representation theory of finite symmetric groups, an important tool used by mathematicians and physicists for centuries. Despite different elegant constructions of group equivariant CNNs, for e.g. (Cohen & Welling, 2016b; Weiler et al., 2018; Weiler & Cesa, 2019) there exists only a few works, e.g., (Cohen et al., 2019b; Kondor & Trivedi, 2018) focusing on the theoretical analysis of such networks, which might be beneficial to understand the geometry of these inductive biases in the model that plays pivotal role in the enhanced expressive power of the equivariant convolutional networks.

Another construction is of Convolutional Kernel Networks (CKNs) (Mairal et al., 2014; Mairal, 2016) where local signal neighbourhoods are mapped to points in a reproducing kernel hilbert space (RKHS) through the kernel trick and then hierarchical representations are built by composing kernels with corresponding RKHSs (patch extraction + kernel mapping + pooling operations in each layer) which is equivalent to construction of a sequence of feature maps in conventional CNNs, but of infinite dimension. A wider functional space approach (Bietti & Mairal, 2019) of CKNs has been proposed for multi-dimensional signals which also admits multilayered and convolutional kernel structure. This functional space also contains a large class of CNNs with homogeneous activation functions, thus showing such CNNs can also enjoy same theoretical properties that of CKNs, therefore highlighting on the geometry of the functional

[1] Department of Electrical Engineering, ESAT-STADIUS, KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium. Correspondence to: Soutrik Roy Chowdhury <soutrik.roychowdhury@kuleuven.be>.

Presented at the 2^{nd} Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA.

spaces in which CNNs lie. Furthermore, an analysis of approximation and generalization capabilities of deep convolutional networks through the lens of CKNs has been performed in (Bietti, 2022). Despite such mathematical analysis, exploring the equivariance properties of CKNs as well as generalization capabilities and robustness of equiv-CKNs have not been performed in details.

In this paper[1] we first study how to make convolutional kernel layers and hence CKNs equivariant to actions by a locally compact group G. Following the notations of diffeomorphism stability (Mallat, 2012) we analyse the stability bounds of equiv-CKNs which depends upon the equivariant architecture of CKNs and corresponding RKHSs norms, thus providing a notion of robustness of equiv-CKNs. We then give an intuition on the (geometric) complexity of equivariant CNNs (equiv-CNNs) by giving a rough outline on how to construct equiv-CNNs in RKHSs, that might be helpful in studying stability and generalization properties of equiv-CNNs by bounding their corresponding RKHS norm.

2. Group Equivariant Convolutional Kernel Networks

The construction of a multilayered CKN involves transforming an input signal $x_0 \in L^2(\mathbb{R}^d, \mathcal{H}_0)$ (for e.g., $\mathcal{H}_0 = \mathbb{R}^{p_0}$, where for a 2D RGB image $p_0 = 3$ and $d = 1$ and $x_0(u)$ in $\mathbb{R}^2$ represents the RGB pixel value at location $u \in \mathbb{R}^2$) into a sequence of feature maps, x_k's in $L^2(\mathbb{R}^d, \mathcal{H}_k)$, by building a sequence of RKHSs $\mathcal{H}_k$'s, for each k, where a new feature map x_k is built from the previous one x_{k-1} by consecutive application of patch extraction P_k, kernel mapping M_k and linear pooling A_k operators, as shown in Figure 1. For a detailed construction of multilayered CKNs on continuous and discrete signal[2] domains we refer readers to (Bietti & Mairal, 2019; Mairal, 2016).

In (section 3.1, (Bietti & Mairal, 2019)) it is shown that CKNs are equivariant to the translations as the layers commute with the action of translations, much like its classical CNNs counterpart. Following the general notations of group equivariance in CNNs (Kondor & Trivedi, 2018) through the notion of locally compact group actions, it is possible to encode other kind of equivariance to group transformations (e.g., rotations, reflections) in CKN layers by constructing equivariant P_k's, M_k's and A_k's for each k that commutes with the action of a group of transformation G. We assume G is locally compact so that we can define a Haar measure μ on it.[3] The action of an element $g \in G$

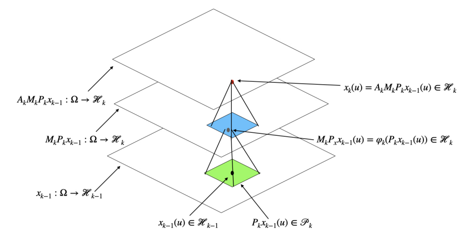

Figure 1. A schematic diagram of 1-layer of a CKN where one constructs k-th signal representation from the $k-1$-th one in a RKHS $\mathcal{H}_k$ through patch extraction, kernel mapping and pooling operators, as similarly shown in (Bietti & Mairal, 2019). Signal domain $\Omega = \mathbb{R}^d$ (in this figure $d = 2$) on which locally compact group G acts. One can construct a multilayered CKN by stacking these layers in a hierarchical manner and make the entire network equivariant by making each layers equivariant to the action of G.

is denoted by operator L_g where $L_g x(u) = x(g^{-1}u)$. We also assume that every element $x(u) \in \mathbb{R}^d$ can be reached with a transformation $u_\omega \in G$ from a neutral element, say $\hat{x}_0(u) \in \mathbb{R}^d$. One can then extend the original signal $\hat{x}$ by defining $x(u) = \hat{x}(u_\omega \cdot \hat{x}_0(u))$, as similarly shown in (Kondor & Trivedi, 2018; Bietti & Mairal, 2019). Then one has

$$\begin{aligned} L_g x(u_\omega) &= x(g^{-1}u_\omega) \\ &= \hat{x}((g^{-1}u_\omega) \cdot \hat{x}_0(u)) = \hat{x}(g^{-1} \cdot x(u)), \end{aligned} \tag{1}$$

where $\cdot$ denotes the group action and hence transformed signals preserve the structure of $\hat{x}$. With the input signals now defined on the locally compact group G, one can define layerwise equivariant patch extraction, kernel mapping and pooling operators at each layer k which are outlined below.

Patch extraction operator. Patch extraction operator $P_k : L^2(G, \mathcal{H}_{k-1}) \to L^2(G, \mathcal{P}_k)$ is defined for all $u \in G$ as

$$P_k x_{k-1}(u) := (x_{k-1}(uv))_{v \in S_k}, \tag{2}$$

where $S_k \subset G$ is a patch shape centered at the identity element of G and $\mathcal{P}_k := L^2(S_k, \mathcal{H}_{k-1})$ is a Hilbert space equipped with the norm $||x||^2 = \int_{S_k} ||x(u)||^2 d\mu_k(u)$, where $d\mu_k$ is the normalized Haar measure on S_k's. P_k commutes with L_g as one can show

$$\begin{aligned} P_k L_g x_{k-1}(u) &= (L_g x_{k-1}(uv))_{v \in S_k} = (x(g^{-1}uv))_{v \in S_k} \\ &= P_k x_{k-1}(g^{-1}u) = L_g P_k x_{k-1}(u) \end{aligned}$$

[1] Ongoing work.

[2] Note that though here in our construction signals are considered continuous for a better theoretical analysis, however for practical purposes one needs to discretize the feature maps.

[3] μ satisfies $\mu(gS) = \mu(S)$ for any Borel set $S \subseteq G$ and $g \in G$. Considering a Haar measure on G, which always exists for locally compact groups, the integration at pooling layers become invariant to group actions, as discussed briefly in appendix A.3 in (Cohen et al., 2019b).

Kernel mapping operator. Kernel operator $M_k : L^2(G, \mathcal{P}_k) \to L^2(G, \mathcal{H}_k)$, for all $u \in G$, is defined as

$$M_k P_k x_{k-1}(u) := \varphi_k(P_k x_{k-1}(u)), \tag{3}$$

where $\varphi_k : \mathcal{P}_k \to \mathcal{H}_k$ is the kernel mapping associated to a positive definite kernel K_k operating on the patches. Like (Mairal, 2016), we define the dot product kernel K_k as

$$K_k(x, x') = ||x||||x'||k_k\left(\frac{\langle x, x'\rangle}{||x||||x'||}\right), x, x' \neq 0,$$

which is positive definite because a Maclaurin expansion with only non-negative coefficients (Schölkopf & Smola, 2018) can be constructed from k_k. A choice of dot product kernels are listed in (Bietti & Mairal, 2019). As M_k is a pointwise operator, thus it commutes with L_g.

Pooling operator. Pooling operator $A_k : L^2(G, \mathcal{H}_k) \to L^2(G, \mathcal{H}_k)$, for all $u \in G$, is defined as

$$\begin{aligned} A_k x_k(u) &:= \int_G x_k(uv) h_k(v) d\mu(v) \\ &= \int_G x_k(v) h_k(u^{-1}v) d\mu(v), \end{aligned} \tag{4}$$

where h_k is the pooling filter at layer k[4] (for e.g., a Gaussian pooling filter), following similar construction from (Raj et al., 2017). Following it's definition it is easy to show that A_k commutes with L_g, i.e., one can show that $A_k L_g x_k(u) = L_g A_k x_k(u)$ for all $g \in G$ and therefore at each layer, all operators are equivariant to the action of G.

Note that the definitions of equivariant operators at each layers follow the similar construction of G-convolution with respect to a locally compact group (section 4, (Kondor & Trivedi, 2018)), thus we anticipate that we can also do the same analysis (for e.g., proposition 1, 2 and theorem 1 proved in (Kondor & Trivedi, 2018)), as discussed in Theorem B.5 in Appendix B. Here the subgroups H_k are the patches S_k, which are Borel sets, according to our assumptions.

A general theory of equiv-CNNs on homogeneous space is given through the notions of vector bundles, fiber space, induced representations, and equivariant kernels in (Cohen et al., 2019b) where equivariant maps between feature spaces are shown to be in one-to-one correspondence with equivariant convolutions, obtained by the space of equivariant kernels (convolution is all you need). As one can define vector bundles and fibers on Hilbert space (Bertram & Hilgert, 1998; Takesaki et al., 2003) we believe that similar notions of equivariant convolution maps can also be deducted for equiv-CKNs, though the latter already contains notion of equivariant kernels through the definitions of M_k's and A_k's. We will work on these in our follow-up studies.

[4]Note that Equation (4) is a type of Bochner integral when $\mathcal{H}$ is infinite dimensional. We provide further construction details of pooling filter as well as of kernel and patch operators in Appendix B.

3. Stability Analysis of Equivariant CKNs

Following our construction of equiv-CKNs in the previous section we now proceed to understand the stability of the equivariant kernel representations under the action of diffeomorphisms, which might be beneficial to get robustness of equiv-CKNs against adversarial examples. We follow the notion of deformation and stability from (Mallat, 2012) which is defined as a C^1-diffeomorphism $\tau : \mathbb{R}^d \to \mathbb{R}^d$ through a linear operator L_τ as $L_\tau x(u) = x(u - \tau(u))$ and we say that a representation $\Phi(\cdot)$ is stable under the actions of τ if there exist non-negative constants C_1 and C_2 such that

$$||\Phi(L_\tau x) - \Phi(x)|| \leq (C_1||\nabla\tau||_\infty + C_2||\tau||_\infty)||x||, \tag{5}$$

where $\nabla\tau$ is the Jacobian of τ and $||\cdot||$ is the L^2-operator norm and $||\nabla\tau||_\infty := sup_{u\in\mathbb{R}^d}||\nabla\tau(u)||$ and $||\tau||_\infty := sup_{u\in\mathbb{R}^d}|\tau(u)|$, where $|\cdot|$ is the standard Euclidean norm on $\mathbb{R}^d$. We also assume $||\nabla\tau||_\infty \leq 1/2$ in order to keep the deformation invertible and avoid degenerate situations, as assumed in (Mallat, 2012). Note that our representation $x_k(u)$ for each layer k can be stacked into a full representation of a N-layer CKN as $x_N(u) = \Phi_N(x) := A_N M_N P_N A_{N-1} M_{N-1} P_{N-1} \cdots A_1 M_1 P_1 x_0(u)$. We are interested in the stability of these convolutional kernel representations. For a semi-direct product group $G := \mathbb{R}^2 \rtimes H$ (Weiler & Cesa, 2019), where $H := SO(2)$, a stability bound with roto-translation patches is given in lemma 9 of (Bietti & Mairal, 2019) which matches the stability form given by Equation (5). For any compact group H the analysis still stands. For the sake of completeness here we state the stability bound of kernel representations for $G := \mathbb{R}^2 \rtimes H$, where each $g \in G$ is given by $g = (u, \hat{h})$, where $u \in \mathbb{R}^2$ and $\hat{h} \in H$ and the group action L_g on the signals are given by $L_g x(u) = x(g^{-1} \cdot u) = x((g^{-1} \cdot u, \hat{h}(\hat{h}')^{-1})) = x(g^{-1}(u, \hat{h}))$, where $\hat{h}'$ in an element of subgroup H.

Lemma 3.1. *If $||\nabla\tau||_\infty \leq 1/2$ and $sup_{c\in\hat{S}_k}|c| \leq \kappa\sigma_{k-1}$, where patch shape $S_k = \{(u, 0)\}_{u\in\hat{S}_k} \subset G$ with $\hat{S}_k \subset \mathbb{R}^2$, σ_{k-1}, the scale of pooling filter h_{k-1} at layer $k-1$, and κ is the patch size, and 0 is the identity element of the subgroup $H \subset G$. Then we have*

$$||[P_k A_{k-1}, L_\tau]|| \leq C_1||\nabla\tau||_\infty, \tag{6}$$

where C_1 depends upon h_{k-1} and κ and $L_\tau x((u, \hat{h})) = x((\tau(u), 0)^{-1}(u, \hat{h}))$. Similarly we have

$$||L_\tau A_N - A_N|| \leq \frac{C_2}{\sigma_N}||\tau||_\infty, \tag{7}$$

where $C_2 = 2^2 \cdot ||\nabla h_N||$ and ∇h_N is the gradient of the last pooling filter h_N.

Here while studying the bounds on operator norm, Equation (6) of Lemma 3.1 is stated on the norm of the commutators of operators, given by $[A, B] = AB - BA$. It shows that commutators are stable to diffeomorphism τ, as the norm is controlled by $||\nabla\tau||_\infty$, whereas the second norm in Equation (7) decays with the last pooling bandwidth σ_N. Note that for the semi-direct group G we restrict the diffeomorphism on $\mathbb{R}^2$ with the assumption that the elements of subgroup H remains unaffected by the deformation τ, which in practice may not be true. We anticipate that including H in deformation would give stability bounds that also depend upon the patch sizes controlled by the elements of H, more precisely through the irreducible representations of H.

Theorem 3.2 (Stability bound)**.** *Subsequently we have*

$$||\Phi_N(L_\tau x) - \Phi_N(x)|| \leq \left(C_1(1+N)||\nabla\tau||_\infty + \frac{C_2}{\sigma_N}||\tau||_\infty\right)||x||. \quad (8)$$

From Theorem 3.2 and Lemma 3.1 we understand that stability to deformation of a CKN representation depends linearly on the depth of network, the patch size (smaller the better) and pooling filter whereas C_2 controls the global invariance of network under deformation and is inversely proportional to last layer's pooling filter bandwidth, σ_N. One needs to have small C_2 in order to have global equivariant representation and indeed it's small as σ_N typically increases exponentially with the number of layers N. A further analysis of stability of equiv-CKNs and proofs of Lemma 3.1 and Theorem 3.2 are given in Appendix C. In the follow-up study we will analyse stability bounds for different G and H as listed in appendix D of (Cohen et al., 2019b).

4. Equiv-CNNs in RKHSs: An Outline

In this section we give a brief outline on how to construct an equivariant G-CNN f (Cohen & Welling, 2016a) recursively from intermediate functions $\hat{f}_k^i$ that lie in the RKHSs $\mathcal{H}_k$ which is of the form,

$$\hat{f}_k^i(x) = ||x||\sigma(\langle w_k^i, x\rangle / ||x||), \quad (9)$$

which closely resembles with the ideas from section 4 (lemma 11) of (Bietti & Mairal, 2019), primarily used to study embedding of CNNs[5] in RKHSs and thus extending theoretical results of CKNs to CNNs. Here w_k^i's are convolutional filters used to obtain intermediate feature maps $\hat{f}_k^i$'s followed by non-linear activation maps (σ's) and linear pooling, similarly as defined in Section 2. Note that to build a group equiv-CNN one needs equivariant filter bank (obtained through G-convolution), pointwise non-linearities and a G-pooling which closely resembles with that of pooling operator defined for equiv-CKNs. The construction (with some mild assumptions on homogeneous non-linearity) of G-CNN f then resembles with that of Equation (9). More details are given in Appendix D.

Following proposition 13 and proposition 14 of (Bietti & Mairal, 2019) one get upper bounds on the RKHS norm of classical CNNs f_σ which is given by the parameter of the final linear fully connected layer, the spectral norm of the convolutional filter parameters at each layers and the choice of the activation function. One can think of similar bounds for equiv-CNNs through it's RKHSs norm given by the final pooling layer (or the norm of the global pooling operator $A_c : L^2(G) \to L^2(\mathbb{R})$ defined for $x \in L^2(G)$ as $A_c x(u) = \int_G x(g^{-1}u) d\mu_c(g)$), equivariant filters and the choice of non-linearities. One can use spectral norms to study generalization, for e.g., done in (Bartlett et al., 2017), of equiv-CNNs. Similarly as one can do stability analysis of CNNs through the Lipschitz smoothness and given by the relation through Cauchy-Schwarz's inequality,

$$|f_\sigma(L_\tau x) - f_\sigma(x)| \leq ||f_\sigma||_{\mathcal{H}_N} ||\Phi_N(L_\tau x) - \Phi_N(x)||_{L^2(G)}, \quad (10)$$

where $||\cdot||_{\mathcal{H}}$ is the standard Hilbert norm, one can then extend the same for equiv-CNNs, outlined in Equation (9), and supported by Theorem 3.2.

5. Conclusion

We have shown how to construct a hierarchical kernel network for multilayered equivariant representation learning by constructing the equivariant feature maps in RKHSs. Then we studied the stability bounds of equiv-CKNs under some mild assumptions and through the Lipschitz stability which shows the stability with respect to a deformation depends upon the specific architecture of equiv-CKNs including the depth of the network and most importantly of the RKHS norm, which acts as an implicit regularizer in our model and controlling the norm leads to better stable model, as shown in (Bietti et al., 2019). Finally we outlined the possibility of embedding a group equiv-CNN into a RKHS and thus extending the studies of equivariant convolutional networks through the lens of equiv-CKNs that might provide novel insights on equivariant convolutions as well as on deep multilayered equivariant kernel networks, for e.g., shown in context of classical CNNs in (Anselmi et al., 2015).

Despite we follow the common framework of (Kondor & Trivedi, 2018) and expect such equiv-CKNs can also be defined on spherical domain (Cohen et al., 2018) it may not be possible to define the same framework on a general

[5] CNNs with homogeneous activation function σ's are considered. For e.g., smoothed-ReLU function.

manifold. One needs careful construction of gauge equivariant CKNs, following similar works on gauge equiv-CNNs (Cohen et al., 2019a; De Haan et al., 2020) which might be possible as one can define manifold on Hilbert spaces. This is a future work we are interested to work on.

We are also interested to do a thorough analysis of generalization capability of equivariant networks through analysing the generalization bounds of equiv-CKNs. A PAC-Bayesian generalization analysis has been performed recently on equivariant networks (Behboodi et al., 2022), whereas (Bietti, 2022) has studied generalization of 2-layers CKNs by bounding the excessive risk for the kernel ridge regression (KRR) estimator. Analyzing the generalization bounds of the equiv-CKNs with these approaches is indeed a promising direction of research.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Advanced Grant E-DUALITY (787960). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. Johan A.K. Suykens is also affiliated with Leuven.AI - KU Leuven Institute for AI, B-3001, Leuven, Belgium.

References

Anselmi, F., Rosasco, L., Tan, C., and Poggio, T. Deep convolutional networks are hierarchical kernel machines. *arXiv preprint arXiv:1508.01084*, 2015.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Behboodi, A., Cesa, G., and Cohen, T. S. A pac-bayesian generalization bound for equivariant networks. *Advances in Neural Information Processing Systems*, 35:5654–5668, 2022.

Bertram, W. and Hilgert, J. Reproducing kernels on vector bundles. *Lie Theory and Its Applications in Physics III*, pp. 43–58, 1998.

Bietti, A. Approximation and learning with deep convolutional models: a kernel perspective. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=lrocYB-0ST2.

Bietti, A. and Mairal, J. Group invariance, stability to deformations, and complexity of deep convolutional representations. *The Journal of Machine Learning Research*, 20(1):876–924, 2019.

Bietti, A., Mialon, G., Chen, D., and Mairal, J. A kernel perspective for regularizing deep neural networks. In *International Conference on Machine Learning*, pp. 664–674. PMLR, 2019.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016a.

Cohen, T., Weiler, M., Kicanaoglu, B., and Welling, M. Gauge equivariant convolutional networks and the icosahedral cnn. In *International conference on Machine learning*, pp. 1321–1330. PMLR, 2019a.

Cohen, T. S. and Welling, M. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016b.

Cohen, T. S., Geiger, M., Köhler, J., and Welling, M. Spherical CNNs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hkbd5xZRb.

Cohen, T. S., Geiger, M., and Weiler, M. A general theory of equivariant cnns on homogeneous spaces. *Advances in neural information processing systems*, 32, 2019b.

De Haan, P., Weiler, M., Cohen, T., and Welling, M. Gauge equivariant mesh cnns: Anisotropic convolutions on geometric graphs. *arXiv preprint arXiv:2003.05425*, 2020.

Kondor, R. and Trivedi, S. 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.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Lang, L. and Weiler, M. A wigner-eckart theorem for group equivariant convolution kernels. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=ajOrOhQOsYx.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551, dec 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.4.541. URL https://doi.org/10.1162/neco.1989.1.4.541.

Li, Z., Zhang, Y., and Arora, S. Why are convolutional nets more sample-efficient than fully-connected nets? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uCY5MuAxcxU.

Mairal, J. End-to-end kernel learning with supervised convolutional kernel networks. *Advances in neural information processing systems*, 29, 2016.

Mairal, J., Koniusz, P., Harchaoui, Z., and Schmid, C. Convolutional kernel networks. *Advances in neural information processing systems*, 27, 2014.

Mallat, S. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

Raj, A., Kumar, A., Mroueh, Y., Fletcher, T., and Schölkopf, B. Local group invariant representations via orbit embeddings. In *Artificial Intelligence and Statistics*, pp. 1225–1235. PMLR, 2017.

Schölkopf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 06 2018. ISBN 9780262256933. doi: 10.7551/mitpress/4175.001.0001. URL https://doi.org/10.7551/mitpress/4175.001.0001.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Takesaki, M. et al. *Theory of operator algebras II*, volume 125. Springer, 2003.

Weiler, M. and Cesa, G. General e (2)-equivariant steerable cnns. *Advances in Neural Information Processing Systems*, 32, 2019.

Weiler, M., Geiger, M., Welling, M., Boomsma, W., and Cohen, T. S. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. *Advances in Neural Information Processing Systems*, 31, 2018.

Zhang, K., Tsang, I. W., and Kwok, J. T. Improved nyström low-rank approximation and error analysis. In *Proceedings of the 25th international conference on Machine learning*, pp. 1232–1239, 2008.

Zhang, Y., Liang, P., and Wainwright, M. J. Convexified convolutional neural networks. In *International Conference on Machine Learning*, pp. 4044–4053. PMLR, 2017.

A. Some Useful Mathematical Tools

We state the classical result of characterizing a Reproducing Kernel Hilbert Space (RKHS) of functions defined from Hilbert space mappings.

Theorem A.1. *Let $\phi : \mathcal{X} \to H$ be a feature map to a Hilbert space H, and let $K(x, x') \coloneqq \langle \phi(x), \phi(x') \rangle_H$ for $x, x' \in \mathcal{X}$. Let $\mathcal{H}$ be the linear subspace defined by $\mathcal{H} \coloneqq \{f_w, w \in H\}$ such that $f_w : x \mapsto \langle w, \phi(x) \rangle_H$, and we consider the norm $||f_w||^2_{\mathcal{H}} \coloneqq inf_{w' \in H}\{||w'||^2_H$ such that $f_w = f_{w'}\}$. Then $\mathcal{H}$ is the RKHS associated to the kernel K.*

We now state another classical result, from harmonic analysis that is used to prove stability results of equiv-CKNs.

Lemma A.2 (Schur's test)**.** *Let $\mathcal{H}$ be a Hilbert space and Ω a subset of $\mathbb{R}^d$. Consider T an integral operator with kernel*[6] *$k : \Omega \times \Omega \to \mathbb{R}$ such that for all $u \in \Omega$ and $x \in L^2(\Omega, \mathcal{H})$,*

$$Tx(u) = \int_\Omega k(u, v)x(v)dv. \tag{11}$$

If $\int |K(u, v)|dv \leq C$ and $\int |K(u, v)|du \leq C$ for all $u \in \Omega$ and $v \in \Omega$ respectively, for some constant C, then for all $x \in L^2(\Omega, \mathcal{H})$, we have $Tx \in L^2(\Omega, \mathcal{H})$ and $||T|| \leq C$.

For an operator $T : L^2(\mathbb{R}^d, \mathcal{H}) \to L^2(\mathbb{R}^d, \mathcal{H}')$, the norm is defined as $||T|| \coloneqq sup_{||x||_{L^2(\mathbb{R}^d,\mathcal{H})} \leq 1} ||Tx||_{L^2(\mathbb{R}^d,\mathcal{H}')}$. One can extend this definition of operator norm on $L^2(G)$, as the latter is the base of our signals defined on the group G, rather than on $\mathbb{R}^d$. With the support of Haar measure on locally compact group G which supports the signal domain the structure of norm is similar to that of on $L^2(\mathbb{R}^d)$ (with a Lebesgue measure support).

B. Further Details on Group Equivariant CKNs on Euclidean Domain

Patch extraction operator P_k's, given by Equation (2) which is encoded in a Hilbert space $\mathcal{P}_k$, preserves the norm, i.e., $||P_k x_{k-1}|| = ||x_{k-1}||$, because of $\mathcal{P}_k$'s are supported by normalized Haar measure. Hence $P_k x_{k-1} \in L^2(G, \mathcal{P}_k)$.

Kernel mapping operator M_k's. Here we give a detailed description of the operator defined in Equation (3) and the choice of dot-product kernels. As defining a homogeneous dot-product kernel yields M_k's as point-wise operator and hence commutes well with the group action L_g's for $g \in G$, we stick to the definition of kernel mapping operators given by (Bietti & Mairal, 2019).

We define a function $k_k : [-1, +1] \to \mathbb{R}$ such that $k_k(u) = \sum_{i=0}^{\infty} b_i u^i$ such that $b_i \geq 0$ for all i and $k_k(1) = 1$ and $0 \leq k'_k(1) \leq 1$, where k'_k is the first order derivative of k_k. Then we define the kernel K_k on $\mathcal{P}_k$ as

$$K_k(x, x') \coloneqq ||x||||x'||k_k\left(\frac{\langle x, x' \rangle}{||x||||x'||}\right), \tag{12}$$

when $x, x' \in \mathcal{P}_k \backslash \{0\}$, and $K_k(x, x') = 0$ if either of x and x' is 0. Note that K_k is positive definite as k_k admits a Maulaurin series with only non-negative coefficients (Schölkopf & Smola, 2018). Then the kernel mapping $\varphi_k(\cdot)$, associated to the positive definite kernel K_k is denoted by $K_k(x, x') = \langle \varphi_k(x), \varphi_k(x') \rangle$.

Norm preservation of operator M_k**.** The constraint $k_k(1) = 1$ ensures that M_k preserves the norm, as, $||\varphi_k(x)|| = K_k(x, x)^{1/2} = ||x||$ leads us to $||M_k P_k x_{k-1}|| = ||P_k x_{k-1}||$ for any k, and therefore $M_k P_k x_{k-1} \in L^2(G, \mathcal{H}_k)$.

Non-expansiveness of $\varphi_k(\cdot)$**'s.** In order to study the stability results we need our kernel mapping non-expansive, i.e., $||\varphi_k(x) - \varphi_k(x')|| \leq ||x - x'||$, for $x, x' \in \mathcal{P}_k$, and the constraint on the derivative of k_k's, i.e., $0 \leq k'_k(1) \leq 1$ ensures that it is always going to hold. The following lemma states the non-expansivess of the kernel mapping.

Lemma B.1 (Lemma 1, (Bietti & Mairal, 2019))**.** *Let K_k be a positive-definite kernel given by Equation (12) which satisfies the constraints given by k_k's. Then the RKHS mapping $\varphi_k : \mathcal{P}_k \to \mathcal{H}_k$, for all $x, x' \in \mathcal{P}_k$ satisfies $||\varphi_k(x) - \varphi_k(x')|| \leq ||x - x'||$. Moreover $K_k(x, x') \geq \langle x, x' \rangle$, i.e., the kernel K_k's are lower bounded by the linear kernels.*

Proof. For the proof we make use of the fact from the Maclaurin expansion[7] of k_k's that

$$k_k(u) = k_k(1) - \int_u^1 k'_k(t)dt \geq k_k(1) - k'_k(1)(1 - u), \tag{13}$$

[6]This type of kernel is known as Schwartz kernel.

[7]We also assume that the series $\sum_i b_i$ and $\sum_i ib_i$'s are convergent.

for all $u \in [-1, +1]$. Then for $x, x' \neq 0$ we have

$$||\varphi_k(x) - \varphi_k(x')||^2 = ||x||^2 + ||x'||^2 - 2||x||||x'||k_k(u),$$

with $u = \langle x, x' \rangle / (||x||||x'||)$. Using the above inequality and the constraint $k_k(1) = 1$ we have

$$\begin{aligned} ||\varphi_k(x) - \varphi_k(x')||^2 &\leq ||x||^2 + ||x'||^2 - 2||x|||x'||(1 - k_k'(1) + k_k'(1)u) \\ &= (1 - k_k'(1))(||x||^2 + ||x'||^2 - 2||x|||x'||) + k_k'(1)(||x||^2 + ||x'||^2 - 2\langle x, x' \rangle) \\ &= (1 - k_k'(1))|||x|| - ||x'|||^2 + k_k'(1)||x - x'||^2 \\ &\leq ||x - x'||^2. \end{aligned}$$

For the last inequality we use the fact that $0 \leq k_k'(1) \leq 1$. □

Remark B.2. One can extend the above lemma for any Lipschitz continuous mapping with $\varphi_k(\cdot)$ being ρ-Lipschitz with $\rho = max(1, \sqrt{k_k'(1)})$, for any value of $k_k'(1)$. Then similarly the above inequality will hold and more generally we'll also have $||\varphi_k(x) - \varphi_k(x')||^2 \leq k_k'(1)||x - x'||^2$ when $k_k'(1) \geq 1$. This together with the above inequality gives us $||\varphi_k(x) - \varphi_k(x')||^2 \leq \rho^2||x - x'||^2$, and yields the result. However for the sake of simplicity we just avoid using Lipschitz continuous kernel mapping as otherwise the stability constants would also depend upon ρ which would increase exponentially with the number of layers that one wants to avoid.

For example, homogeneous Gaussian kernel defined as, $K_{RBF}(x, x') = exp(-\alpha||x - x'||^2)$ is non-expansive only when $\alpha \leq 1$ but is still Lipschitz for any values of α.

Pooling operator A_k's. From Equation (4) we have noted the involvement of a pooling filter h_k's. One typical example of such pooling filter is Gaussian pooling filter which is given by $h_{\sigma_k}(u) := \sigma_k^{-d} h_k(u/\sigma_k)$, where σ_k is the scale of the pooling filter and $h_k(u) = (2\pi)^{-d/2} exp(-|u|^2/2)$.

In the definition of A_k in Equation (4) the pooling filter h_k is typically localized around the identity element of G. By applying Schur's test on the operator A_k one obtains that $||A_k|| \leq 1$ and hence $x_k(u) \in L^2(G, \mathcal{H}_k)$.

Remark B.3. Unlike the operators P_k and M_k, A_k doesn't preserve the norm (which is in contrary to the setting of (Mallat, 2012)) as $||A_k x_k(u)|| \leq ||x_k(u)||$. As we are using a pooling filter with a scale of σ_k, therefore A_k's may reduce frequencies of signals that are larger than $1/\sigma_k$. However norm preservation is less relevant in the kernel based setting as discussed in (Bietti & Mairal, 2019), as if one picks a Gaussian kernel mapping on top of the last feature map instead of a linear layer as prediction layer then the final feature representation preserves stability as well as have a unit norm.

Remark B.4. One can also pool on subset $H \subseteq G$ by only integrating on H, much like the subgroup pooling described in (Cohen & Welling, 2016a) for group equiv-CNNs. This subsampling on a subgroup $H \subseteq G$, though gives the subsampled feature map H-equivariant but one can obtain the full group G-equivariance by performing the pooling on the entire H. Moreover from the first expression of A_k in Equation (4) it is easy to see that the pooling operator commutes with L_g.

Some notes on discretization and kernel approximation. Though for our theoretical analysis purposes we have defined signals on $L^2(G, \mathcal{H}_k)$ but for practical implementation one needs to discretize the signals as in practice, signals are discrete. For group equiv-CNNs it is nicely discussed in (Cohen & Welling, 2016b; Cohen et al., 2019b) through the notion of fiber space (bundles), making each discrete feature maps equivariant and hence the entire network equivariant, through the efficient implementation of G-equivariant layers. For our construction it is possible to sample each feature map $\Phi_k(x) := x_k(u)$ on a discrete set with no loss of information. For the classical CKNs an in-depth discussion on discretization is available through section 2.1 of (Bietti & Mairal, 2019) or by simply following the construction of hierarchical CKN layers from (Mairal, 2016).

In (Mairal, 2016) a finite dimensional subspace projection of RKHS mappings $\varphi_k(\cdot)$ are discussed through an adapted Nyström method (Zhang et al., 2008) which is essential in the construction of CKNs. However this is not a drawback as such finite dimensional approximation of RKHS mappings still live in the corresponding RKHSs as well as it won't hurt the stability results due to the non-expansiveness of the projection. However in this case some signal information is lost as through projection we can no longer maintain the norm preservence of the kernel mapping operator M_k.

Equivariant convolutional kernel representations. Note the term 'convolution' in equiv-CKNs come from the definition of pooling filter which resembles with the definition of classical convolutional mapping and in line with the generalized convolutional operator[8] defined on compact groups by (Kondor & Trivedi, 2018) which is given by $(f * h) = \int_G f(uv^{-1})h(v)d\mu(v)$,

[8]Unlike the classical convolution $u - v$ is replaced with uv^{-1}.

where f and h are functions defined on G and the integration is with respect to the Haar measure μ. Note how our pooling filter is in a convolution with the feature map $x_k(\cdot)$'s. We state the following theorem without a proof which establish the group equivariance of the CKNs, omitting some practicalities on which we'll work in details in our follow-up paper.

Theorem B.5 (Equivariance of an entire CKN). *Let G be a locally compact group and Φ_N be a (N+1)-layered CKN*[9]*, following the standard construction of a CKN. If the pooling operators A_k are in cross-correlation with the feature maps x_k for each $k \in 1, .., N$, i.e., $A_k(x_k) = x_k * h_k$, where h_k is the pooling filter associated to A_k's and the patches $S_k \subset G$ form the index sets $\chi_k = G/S_k$ on which one can define a generalized convolution defined in the previous paragraph following (Kondor & Trivedi, 2018), then the CKN Φ_N is equivariant with respect to locally compact group G's action on it's inputs. The converse also holds if the index sets χ_k covers the entire G.*

Note that, one can always write an equivariant map in an convolution-like integral (theorem 3.1, (Cohen et al., 2019b)) which also supports our construction of group equiv-CKNs on homogeneous space. A direct consequence of Theorem B.5 is the following.

Corollary B.6 (Equivariant kernels). *Equation (4) can always be written as cross-correlation between the feature map and the pooling filter. Moreover in equiv-CKNs, representation, $\Phi_N(x) \in L^2(G, \mathcal{H}_N)$ is equivariant (with respect to G) if and only if each φ_k's are in cross-correlation with an equivariant pooling filter.*

Proof. The proof is straight-forward and immediately follows from the definition of cross-correlation, i.e., $[h_k * x_k](u) := \int_G h_k(u^{-1}v)x_k(v)d\mu_k(v) = A_k x_k(u)$. For the second part, note that $A_k x_k(u)$ can be written as $A_k M_k P_k x_{k-1}(u)$ as one can see it from Figure 1. Then establishing link with kernel mapping φ_k with h_k's are staightforward and the equivariance followed from Theorem B.5. □

Remark B.7. Note that through the above corollary we get another equivalent notion of equivariant kernels, as described in (section 3.1 of (Cohen et al., 2019b)). However note that in equiv-CKNs the kernels are described by kernel mapping φ_k's which is given by the RKHS mapping, giving true flavour of kernel machine, which is missing in group equiv-CNNs. We note that more recently (Lang & Weiler, 2021) gives a full characterization of group equivariant kernels but it still misses the notion of RKHSs.

C. Stability Analysis Proofs

Before giving the proofs of Lemma 3.1 and Theorem 3.2 we first dive deep into the stability form and how it is controlled by the operator norm (and hence of the RKHSs norm) which are motivated by similar notion of diffeomorphism studied in (Mallat, 2012).

The assumption $sup_{c\in\hat{S}_k}|c| \leq \kappa\sigma_{k-1}$ is made to relate the scale of pooling operator at layer $k-1$ with the diameter of the patch S_k. As σ_k's increases exponentially with the layers k and characterizes resolution of each feature map, the assumption helps us to consider such patch sizes that are adapted to those resolutions, and helps us control the stability. Let us first state the bound on operator norms.

Proposition C.1 (Proposition 4 (Bietti & Mairal, 2019)). *For any $x \in L^2(\mathbb{R}^d, \mathcal{H}_0)$, we have*

$$||\Phi_N(L_\tau x) - \Phi_N(x)|| \leq \left(\sum_{k=1}^{N} |||[P_k A_{k-1}, L_\tau]||| + |||[A_N, L_\tau]||| + ||L_\tau A_N - A_N||\right) ||x||. \tag{14}$$

By expanding Φ_N's as shown in the multilayered construction of CKNs in Section 3 and using the facts of norm preservence of P_k and M_k's, non-expansiveness of M_k's and $||A_k|| \leq 1$ we can get the above result. Moreover one also uses the fact that kernel mapping M_k is defined point-wise and thus commutes with the deformation operator L_τ. The result holds even when x is defined on the locally compact group G, i.e., when $x \in L^2(G, \mathcal{H}_0)$.

C.1. Proof of Lemma 3.1.

Proof. Note that for all k we have

$$\begin{aligned} P_k x_{k-1}((u, \hat{h})) &= x((uv, \hat{h} \cdot 0))_{v\in\hat{h}\hat{S}_k} \\ &= x((uv, \hat{h}))_{v\in\hat{h}\hat{S}_k}, \end{aligned}$$

[9]The last layer is a prediction layer.

where $\hat{h}\hat{S}_k$ is in $S_k \subseteq G/H$, and by $\hat{h} \cdot 0$, we meant the group composition with the identity element.

Similarly we have $A_k x_k((u, \hat{h})) = \int_G x_k((v, \hat{h}'))h_k((u, \hat{h})^{-1}v)d\mu(v) = \int_{\mathbb{R}^2} x_k((v, \hat{h}))h_k(u^{-1}v)d\mu(v)$ which follows from the second term of Equation (4). Moreover as $G/H \simeq \mathbb{R}^2$, we can integrate over $G/H \simeq \mathbb{R}^2$ by using integral over G, i.e., $\int_{\mathbb{R}^2} f(x)dx = \int_G f(gH)dg$.

For a fixed $\hat{h} \in H$ we can obtain signal $\hat{x} := x(\cdot, \hat{h}) \in L^2(\mathbb{R}^2, \mathcal{H}_0)$ from the signal $x \in L^2(G, \mathcal{H}_0)$, and we have corresponding operators $\tilde{P}_k$, $\tilde{A}_k$ and $\tilde{L_\tau}$ now defined on $L^2(\mathbb{R}^2)$, with a transformed patch $\tilde{S}_k = \hat{h}\hat{S}_k$ for $\tilde{P}_k$.

Then for $x \in L^2(G, \mathcal{H}_0)$, we have,

$$\begin{aligned}
||[P_k A_{k-1}, L_\tau]x||^2_{L^2(G)} &= \int_G ||([P_k A_{k-1}, L_\tau]x)(\cdot, \hat{h})||^2_{L^2(\mathbb{R}^2)} d\mu(\hat{h}) \\
&= \int_{\mathbb{R}^2} ||[\tilde{P}_k \tilde{A}_{k-1}, \tilde{L}_\tau](\hat{x})||^2_{L^2(\mathbb{R}^2)} d\mu(\hat{h}) \\
&\leq \int_{\mathbb{R}^2} ||[\tilde{P}_k \tilde{A}_{k-1}, \tilde{L}_\tau]||^2 ||(\hat{x})||^2_{L^2(\mathbb{R}^2)} d\mu(\hat{h}) \\
&\leq \left(sup||[\tilde{P}_k \tilde{A}_{k-1}, \tilde{L}_\tau]||^2 \right) ||x||^2_{L^2(G)},
\end{aligned}$$

so that one has $||[P_k A_{k-1}, L_\tau]||_{L^2(G)} \leq sup||[\tilde{P}_k \tilde{A}_{k-1}, \tilde{L}_\tau]||_{L^2(\mathbb{R}^2)}$. As we have assumed that $sup_{c \in \hat{S}_k}|c| \leq \kappa\sigma_{k-1}$, so we can bound each of $||[\tilde{P}_k \tilde{A}_{k-1}, \tilde{L}_\tau]||$ as shown in section 3.1 of (Bietti & Mairal, 2019). We refer the readers to that section[10] for detailed understanding of deformation stability of classical CKNs by bounding the operator norms when signals are in $L^2(\mathbb{R}^2)$ which is possible as one can bound $||[\tilde{P}_k \tilde{A}_{k-1}, \tilde{L}_\tau]||$ with $sup_{c \in \hat{S}_k}||[L_c \tilde{A}_{k-1}, \tilde{L}_\tau]||$ and showing $[L_c \tilde{A}_{k-1}, \tilde{L}_\tau]$ is an integral operator, one can bound its norm via Schur's test. Equation (6) is then obtained by applying the bound derived for classical CKNs.

Similarly by applying lemma 2.11 from (Mallat, 2012) one obtains upper bound on $||L_\tau A_N - A_N||_{L^2(G)}$ by first restricting it on $||\tilde{L}_\tau \tilde{A}_N - \tilde{A}_N||_{L^2(\mathbb{R}^2)}$ and then applying the lemma 2.11 get the desired result, given by Equation (7). □

Proof of Theorem 3.2 immediately follows by combining Proposition C.1 with Equation (6) and Equation (7) which are extracted by bounding the corresponding operator norms.

D. Geometric Model Complexity of Deep Equivariant Convolutional Representations

If one can write a group equiv-CNN f in the form $f(x) = \langle f, \Phi(x) \rangle$, where $\Phi(\cdot)$ is the equivariant convolutional kernel representation, then one can extend the stability analysis of equiv-CKNs, $\Phi(\cdot)$'s to the stability analysis of equiv-CNNs. Moreover computing the RKHS norm of the equiv-CNNs one can also control generalization, so that controlling the RKHS norm serves as the geometric model complexity of equiv-CNNs, where the term 'geometric' refers to the equivariance of operators and the geometry of RKHSs.

Before outlining the construction of an equiv-CNNs in RKHSs, let's state a lemma from (Bietti & Mairal, 2019) which closely follows the results of (Zhang et al., 2017), linking the homogeneous activation function with RKHSs $\mathcal{H}_k$, which we believe also holds for group equiv-CNNs as the pointwise homogeneous activation maps σ are replaced with pointwise non-linearity maps ν, as described in (Cohen & Welling, 2016a).

Lemma D.1 (Lemma 11, (Bietti & Mairal, 2019))**.** *If the activation maps σ admits a polynomial expansion and we define our kernel K_k as given in Equation (12). Then for $g \in \mathcal{P}_k$, the RKHS $\mathcal{H}_k$ contains the function,*

$$f : x \mapsto ||x||\sigma(\langle g, x \rangle / ||x||), \tag{15}$$

which matches the form given by Equation (9).

For our construction of k_k's, the next corollary follows from the above lemma as well as from the Theorem A.1.

Corollary D.2. *The RKHSs $\mathcal{H}_k$ contain all linear functions of the form $x \mapsto \langle g, x \rangle$, with $g \in \mathcal{P}_k$.*

[10]Read appendix C.4. for proof of the lemma.

Note that RKHS of the kernel $K_N(x, x') = \langle \Phi(x), \Phi(x') \rangle$, defined at the prediction layer as final representation $\Phi(x) \in \mathcal{H}_{N+1}$ contains functions of the form $f : x \mapsto \langle w, \Phi(x) \rangle$, with $w \in \mathcal{H}_{N+1}$ and $||f|| \leq ||w||_{\mathcal{H}_{N+1}}$. This is a consequence of Theorem A.1, and also in line with the stated corollary, as in our construction $\mathcal{P}_k$'s are also RKHS.

D.1. Construction of group equiv-CNN f in the RKHS

One defines the k-th layer of equiv-CNN function f in $\mathcal{H}_k$ from the $(k-1)$-th layer as follows: For an input signal $x_0 \in L^2(G, \mathcal{H}_0 := \mathbb{R}^{p_0})$, we build a sequence of feature maps, $x_k \in L^2(G, \mathcal{H}_k := \mathbb{R}^{p_k})$ with p_k channels. We use the following intermediate functions $g_k^i \in \mathcal{P}_k$ and $f_k^i \in \mathcal{H}_k$, where $i = 1, ..., p_k$ and construct it from the $(k-1)$-th intermediate function inductively, where the intermediate functions are of form Equation (15).

$$g_k^i(u) = \sum_{h \in S_k} \sum_{j=1}^{p_{k-1}} w_k^{ij}(u^{-1}h) f_{k-1}^j(x(h))$$

$$f_k^i(x(u)) = ||x(u)|| \sigma\left(\langle g_k^i, x(u) \rangle / ||x(u)||\right),$$

for $x(u) \in \mathcal{P}_k \backslash \{0\}$, $u \in G$, and the filters $w_k^i(u) = (w_k^{ij}(u))_{j=1,\ldots p_{k-1}}$ are equivariant through the definition of the intermediates and also matches the notion of group equivariant correlation of (Cohen & Welling, 2016a).

With this construction one can show that the equivariant feature maps x_k are given are $x_k^i(u) = \langle f_k^i, M_k P_k x_{k-1}(u) \rangle$, where $u \in G$ and P_k and M_k's are our patch and kernel operators, respectively, used to define an equiv-CKN. With a final linear prediction layer one can immediately show that an equivariant CNN lies in a RKHS, supported by Corollary D.2. We will work on the detailed construction in our follow-up paper, discussing in depth the generalization bounds and sample complexity of equiv-CKNs.

D.2. Note on the norm of equiv-CNN f and generalization bounds

We have seen that how the operator norms control the stability of the CKNs and through Equation (10) we get the model complexity of group equiv-CNNs, where the RKHS norm of f also plays an important role in the stability of the model as well as understanding the generalization capabilities, and hence of the geometric model complexity of the equivariant convolutional networks.

One can study generalization bounds through Rademacher complexity and margin bounds, for e.g., as done in (Shalev-Shwartz & Ben-David, 2014), where one studies the upper bound on the Rademacher complexity of a function class $\mathcal{F}_\lambda$ with bounded RKHS norm, $\mathcal{F}_\lambda = \{f \in \mathcal{H}_K : ||f|| \leq \lambda\}$, for a dataset $\{x_1, x_2, ..., x_M\}$, given by,

$$Rad_M(\mathcal{F}_\lambda) \leq \frac{\lambda \sqrt{1/M \sum_{i=1}^{M} K(x_i, x_i)}}{\sqrt{M}}.$$

The bound remains valid when considering CNN functions of form f_σ, given by Equation (15), as such family of functions f_σ contains in the class of $\mathcal{F}_\lambda$. Generalization bound depends upon the model complexity parameter λ, sample size M and on the choice of the kernel at the prediction layer. However it doesn't explicitly yield the layer-wise architectural choices of CKNs. However in practice, learning with a tight constraint, like $||f|| \leq \lambda$, can be infeasible and thus one needs to replace λ with a similar bound with $||f_M||$ which can be directly obtained from the training data (Theorem 26.14,(Shalev-Shwartz & Ben-David, 2014)). This then involves the construction of equiv-CNNs in a RKHS, as seen in Appendix D.1. and the corresponding RKHS norm, together with the sample size gives the upper bound of Rademacher complexity. Hence this leads to a way of studying generalization bounds of group equiv-CNNs.