Extended Abstract Track

Barron's Theorem for Equivariant Networks

Editors: List of editors' names

Abstract

The incorporation of known symmetries in a learning task provides a powerful inductive bias, reducing the sample complexity of learning equivariant functions in both theory and practice. Group-symmetric architectures for equivariant deep learning are now widespread, as are accompanying universality results that verify their representational power. However, these symmetric approximation theorems suffer from the same major drawback as their original non-symmetric counterparts: namely, they may require impractically large networks. In this work, we demonstrate that for some commonly used groups, there exist smooth subclasses of functions — analogous to Barron classes of functions — which can be *efficiently* approximated using invariant architectures. In particular, for permutation subgroups, there exist invariant approximating architectures whose sizes, while dependent on the precise orbit structure of the function, are in many cases just as small as the non-invariant architectures given by Barron's Theorem. For the rotation group, we define an invariant approximating architecture with a new invariant nonlinearity, which may be of independent practical interest, that is similarly just as small as its non-invariant counterparts. Overall, we view this work as a first step towards capturing the *smoothness* of invariant functions in invariant universal approximation, thereby providing approximation results that are not only invariant, but *efficient*.

Keywords: Equivariance, invariance, symmetry, universality, approximation, Barron

1. Introduction

Symmetries arise in a wide range of learning tasks, with symmetry groups ranging from the discrete (such as permutations of sets or graph nodes) to the continuous (such as rotations of point clouds). In such settings, restricting the hypothesis class to only symmetric functions improves generalization bounds in linear and kernel learning (Elesedy and Zaidi, 2021; Mei et al., 2021), and empirically aids generalization in deep learning (Cohen and Welling, 2016; Kondor et al., 2018). Group-symmetric architectures come in many shapes and sizes, however, and one can ask how expressive various architectures truly are.

Yarotsky (2018) first demonstrated that shallow neural nets, based on classical results on polynomial invariants and equivariants (Hilbert, 1890, 1893), are universal approximators of invariant and equivariant continuous functions, respectively, for compact groups. Later works built on this framework for many different groups: for permutation subgroups, Maron et al. (2019b) defined tensor networks, and proved universality when the network has sufficiently high order tensor activations. Similarly, Maron et al. (2019a) defined invariant graph networks, which are universal for certain classes of graph functions (dependent on the network's tensor order). Bogatskiy et al. (2020) defined universal tensor-product networks equivariant to the Lorentz group, while Dym and Maron (2021) showed that commonly used point cloud architectures are universal. DeepSets, a permutation-invariant architecture for learning on sets, has also been shown to be universal (Zaheer et al., 2017).

© 2022 .

However, these results do not bound the *size* of the network needed to approximate a continuous equivariant function arbitrarily well. Indeed, the size of the neural network will often depend at least exponentially on the input dimension, or on abstract algebraic quantities such as the cardinality or degree of the generating set of polynomials. In fact, this caveat dates back to the original universality theorems by Cybenko (Cybenko, 1989) and Hornik (Hornik et al., 1990), who first demonstrated that shallow neural networks with sigmoidal activations are universal approximators. These early universality results relied on polynomial approximation, yielding exponential dependence on dimension. In his seminal work, Barron (1993) circumvented this drawback by defining a class of smooth functions which can be approximated *efficiently*, in terms of network size. Yet, an analogous result for invariant networks has not yet been shown. Our central question is therefore: *What classes of invariant functions can be* efficiently *approximated by invariant neural networks?*

Of course, one can trivially group-average any approximating architecture, such as that of Barron (1993), to obtain an invariant approximation with error at least as good (Yarotsky, 2018). However, the size of the network naively increases by a factor of $|G|$, which may be large or even infinite. Our main contribution is to demonstrate that this blow-up in size is not necessary for several commonly used groups and function classes.

Our Contributions For subgroups of the permutation group S_n (Theorem 3), we show that there exist invariant approximating architectures whose size very often improves on the group-averaging baseline, and in many non-trivial cases by a factor of $|G|$. For the rotation group (Theorem 7), we analytically simplify the group-averaging operation to obtain an invariant approximating architecture with a novel analytic nonlinearity. As a result, group-averaging need not increase the number of nodes in the network.

Preliminaries and Notation Let G be a group acting orthogonally on $\mathcal{X} \subseteq \mathbb{R}^n$. A function $f : \mathcal{X} \to \mathbb{R}$ is G-invariant if $f(\boldsymbol{x}) = f(g\boldsymbol{x})\ \forall \boldsymbol{x} \in \mathcal{X}, g \in G$. $\mathcal{X}$ is partitioned into possibly unequally-sized[1] orbits $\mathcal{O}$, where $[\boldsymbol{x}] = \{g\boldsymbol{x} : g \in G\} \in \mathcal{O}$. $\boldsymbol{x}_i$ is the ith entry of $\boldsymbol{x}$.

Barron's Theorem We review Barron's Theorem below; see Appendix A.1 for details.

Theorem 1 (Barron (1993)) *Let $f : \mathbb{R}^n \to \mathbb{R}$ have Fourier transform $\hat{f}$ and "Barron parameter" $C_f := \int_{w \in \mathbb{R}^n} \|w\| \cdot |\hat{f}(w)|$. Then, there exists a two-layer neural network with nonlinearity $\sigma(x) = \mathbb{1}(x \geq 0)$, and approximation error $||f - NN_{Barron}||_2^2$ at most $\frac{\Theta(C_f^2)}{k}$.*

The central idea of Barron's Theorem is to sample the Fourier basis functions[2] proportionally to f's Fourier coefficients, which produces a two-layer cosine[3] network. Maurey's Lemma (Pisier, 1981) (see Lemma 14) then bounds the network size versus approximation error tradeoff, with error governed by the expected norm of the sampled functions.

2. Subgroups of S_n

Let $G \leq S_n$, and throughout this subsection, let $f : [0,1]^n \to \mathbb{R}$ be a G-invariant function. f can be decomposed using discrete Fourier series: $f(x) = \sum_{w \in \mathbb{Z}^n} \hat{f}(w) e^{2\pi i \langle x, w \rangle} =$

1. For example, the orbit of $[1, 1, \ldots, 1]$ over S_n is itself, while the orbit of $[1, 2, \ldots, n]$ has size $|S_n| = n!$.
2. In Section 2, we make the minor switch to discrete (rather than continuous) Fourier series, which simplifies the analysis for invariant functions while retaining the key intuitions of Barron's Theorem.
3. An algebraic manipulation transforms the cosine nonlinearity to an indicator.

Extended Abstract Track

$\sum_{[\boldsymbol{w}]\in\mathcal{O}}\sum_{\boldsymbol{w}\in[\boldsymbol{w}]}|\hat{f}([\boldsymbol{w}])|v_{[\boldsymbol{w}]}$, where $\hat{f}(\boldsymbol{w}) = \int_{\boldsymbol{x}\in[0,1]^n} f(\boldsymbol{x})e^{2\pi i\langle \boldsymbol{x},\boldsymbol{w}\rangle}$, $v_{\boldsymbol{w}} := \cos(2\pi\theta([\boldsymbol{w}]) + 2\pi\langle \boldsymbol{x},\boldsymbol{w}\rangle)$, and the orbit $v_{[\boldsymbol{w}]} := \sum_{\boldsymbol{w}\in[\boldsymbol{w}]} v_{\boldsymbol{w}}$. We define an inner product over periodic functions, $\langle f,g\rangle = \int_{[0,1]^n} f(\boldsymbol{x})g(\boldsymbol{x})$. It is easy to check that $\hat{f}(\boldsymbol{w}) = \hat{f}(g\boldsymbol{w})\,\forall g\in G, \forall \boldsymbol{w}\in\mathbb{Z}^n$ (Appendix A.2). Using this Fourier invariance structure, we will see that Theorem 1 can be non-trivially extended for invariant networks by sampling *orbits*, instead of individual frequencies, in Fourier space. In fact, this can be done for any invariant function and group; naively, it is equivalent to group-averaging the architecture of Barron (1993). For permutation subgroups, however, one can show that fewer sampled orbits than group-averaging are needed to achieve the original network's approximation error, thereby circumventing the blow-up in size. The following technical condition greatly simplifies the analysis.

Definition 2 (SOO condition) *f satisfies the* signed orbit overlap (SOO) condition *if for all orbits $[\boldsymbol{w}]$ in Supp$(\hat{f})$, every element of the orbit has a unique absolute value: $\forall \boldsymbol{w}\in$ Supp$(\hat{f})$, $\forall g\in G$, $\exists\, i$ s.t. $|g\boldsymbol{w}|_i \neq |\boldsymbol{w}_i|$. E.g. for any $G\leq S_n$, Supp$(\hat{f})\subseteq\mathbb{N}^n$ suffices.*

Theorem 3 *Suppose a G-invariant periodic function f satisfies Definition 2, let $\mathcal{N} = \sum_{\boldsymbol{w}\in\mathbb{Z}^n}|\hat{f}(\boldsymbol{w})|$, and let m and M denote the minimum and maximum, respectively, orbit sizes $|[\boldsymbol{w}]|$ in Supp$(\hat{f})$. Then there exists a two-layer group-invariant network NN with cosine activations and kM hidden units such that*

$$\int_{[0,1]^n}|f(\boldsymbol{x})-NN(\boldsymbol{x})|^2 d\boldsymbol{x} \leq \frac{\mathcal{N}\cdot\sum_{[\boldsymbol{w}]\in\mathcal{O}}|\hat{f}([\boldsymbol{w}])|}{k} \leq \frac{\mathcal{N}^2}{mk} \tag{1}$$

The proof of Theorem 3 (see Appendix A.3) modifies that of Theorem 1 by sampling orbits instead of individual frequencies in Fourier space, to ensure invariance. Together, Definition 2 and the discrete Fourier series are useful, because they ensure that the cosine functions in a given orbit are orthogonal (Lemma 13). Then, $||v_{[\boldsymbol{w}]}||^2$ (which controls the approximation error from sampling) is equal to $\sum_{\boldsymbol{w}\in[\boldsymbol{w}]}||v_{\boldsymbol{w}}||^2$.

Remark 4 (Comparison to Barron) *A non-invariant version of Theorem 3 would obtain error $\frac{\mathcal{N}^2}{k}$ with k hidden units; then, naive group-averaging would obtain error $\frac{\mathcal{N}^2}{k}$ with kM units. Therefore, Theorem 3 has a factor of m improvement over group-averaging. Moreover, if $m=M$ (all orbits are equally sized), there is no increase in size relative to the non-invariant approximation.*

Although already an improvement for many functions and groups, Theorem 3 does not improve over group-averaging when $m=1$, and is more generally loose when $m << M$. A more precise version of Theorem 3 below (proven in Appendix A.4) articulates the exact interplay between f's orbit structure and approximation error.

Theorem 5 *In the setting of Theorem 3, let x_1 equal $\mathcal{N}\cdot v_{[\boldsymbol{w}]}$ with probability $\frac{|\hat{f}([\boldsymbol{w}])|\cdot|[\boldsymbol{w}]|}{\mathcal{N}}$, and s_1 equal the corresponding orbit size $|[\boldsymbol{w}]|$. Then there exists a two-layer group-invariant network NN with cosine activations, equal to a weighted sum of k orbits $v_{[\boldsymbol{w}]}$ and comprising H* total *hidden units, such that*

$$\int_{[0,1]^n}|f(x)-NN(x)|^2dx = (2-k)\mathbb{E}[s_1]||f||^2 - 2\langle f,\mathbb{E}[x_1s_1]\rangle + \mathcal{N}^2\mathbb{E}[s_1^2] + (k-1)\mathbb{E}[s_1]^2\mathcal{N}^2$$

Remark 6 *Theorem 5 depends on the new quantities $\mathbb{E}[s_1]$, $\mathbb{E}[s_1^2]$, and $\mathbb{E}[x_1 s_1]$, which capture how the mass of $\hat{f}$ is distributed among orbits of different sizes. If most or all of the mass of $\hat{f}$ is on equal-sized orbits, Theorem 5 essentially recovers Theorem 3 with $m = M$.*

3. Rotations

As noted in Section 1, one straightforward approach to obtaining universal approximation theorems for invariant/equivariant architectures is simply to group-average a generic approximation, such as that of Theorem 1, producing a symmetric architecture with no worse approximation error (Appendix A.5). However, this architecture is naively a factor of $|G|$ larger than the original architecture. In the previous section, we used a tighter sampling analysis to avoid this blow-up for permutation subgroups. In this section, we instead look closely at the group-averaged network architecture for $G = SO(n)$, and show that it can be computed with a network whose size is *equal* to that of the original architecture, by replacing the nonlinearities with a certain special function.

Theorem 7 *When averaged over the special orthogonal group in n dimensions, the nonlinear invariant function $\int_{g \in SO(n)} \mathbb{1}(\boldsymbol{w}^T g \boldsymbol{x} - b \geq 0)$ is expressible as follows:*

$$\bar{\mathbb{1}}_{\boldsymbol{w},b}(x) := \int_{g \in SO(n)} \mathbb{1}(\boldsymbol{w}^T g \boldsymbol{x} - b \geq 0) = \frac{1}{2} I_{\sin(\phi)^2}\left(\frac{n-1}{2}, \frac{1}{2}\right).$$

Here, $\phi = \cos^{-1}\left(\frac{b}{||\boldsymbol{w}|| \cdot ||\boldsymbol{x}||}\right)$ and $I_a(b, c)$ denotes the regularized incomplete beta function. As a result, any network (including that of Theorem 1) of the form $\boldsymbol{\alpha}^T \mathbb{1}(\mathbf{W}\boldsymbol{x} - \boldsymbol{b})$ can be averaged over $SO(n)$ using regularized incomplete beta function nonlinearities, with no further increase in network size.

The proof of Theorem 7 is in Appendix A.6. To our knowledge, no $SO(n)$-invariant architecture have used such a nonlinearity before. Computing analytic forms for invariant and equivariant network layers is itself an exciting direction for future work, both for stronger approximation results and as a practical tool in equivariant architecture design.

4. Conclusion

In this work, we provided evidence that invariant, universal, and small architectures should exist for broad classes of "nice" invariant functions. Barron's Theorem (Barron, 1993) first provided a particular quantification of the smoothness of a function, based on how well it can be approximated by sampling Fourier coefficients; here, we adapt this notion of smoothness for group-invariant functions, instead structuring our sampling in Fourier space to produce an approximating architecture that is itself invariant, too. However, this notion of smoothness arose as a consequence of the particular technical tools of Barron's Theorem. Is there a better, more inherent way of measuring the smoothness of invariant functions? For example, the orbits of a radially symmetric function correspond to $\mathbb{R}_+$, and it is clear that the function's approximability via neural nets is governed by its smoothness along the radial dimension: rigorously extending this intuition to other groups is an exciting direction for future work. Overall, we hope that the techniques and examples developed here serve as an invitation to further develop the theory of efficient invariant approximation.

Extended Abstract Track

References

Andrew Barron. Barron, a.e.: Universal approximation bounds for superpositions of a sigmoidal function. ieee trans. on information theory 39, 930-945. *Information Theory, IEEE Transactions on*, 39:930 – 945, 06 1993. doi: 10.1109/18.256500.

Alexander Bogatskiy, Brandon Anderson, Jan T. Offermann, Marwah Roussi, David W. Miller, and Risi Kondor. Lorentz group equivariant neural network for particle physics. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2990–2999. JMLR.org, 2016.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, December 1989. ISSN 0932-4194. doi: 10.1007/BF02551274. URL http://dx.doi.org/10.1007/BF02551274.

Nadav Dym and Haggai Maron. On the universality of rotation equivariant point cloud networks. *ArXiv*, abs/2010.02449, 2021.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. *arXiv preprint arXiv:2102.10333*, 2021.

D. Hilbert. Ueber die vollen invariantensysteme. *Mathematische Annalen*, 42:313–373, 1893. URL http://eudml.org/doc/157652.

David Hilbert. Über die theorie der algebraischen formen. *Math. Annalen*, 36:473–534, 1890.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3(5):551–560, 1990. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(90)90005-6. URL https://www.sciencedirect.com/science/article/pii/0893608090900056.

Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch–gordan nets: a fully fourier space spherical convolutional neural network. *Advances in Neural Information Processing Systems*, 31:10117–10126, 2018.

S Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics Statistics*, 4, 01 2011. doi: 10.3923/ajms.2011.66.70.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. 2019a.

Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the universality of invariant networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4363–4371. PMLR, 09–15 Jun 2019b. URL https://proceedings.mlr.press/v97/maron19a.html.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. 02 2021.

Gilles Pisier. Remarques sur un résultat non publié de b. maurey. 1981.

Dmitry Yarotsky. Universal approximations of invariant maps by neural networks. *Constructive Approximation*, 55:407–474, 2018.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf.

Appendix A. Proofs

A.1. Barron's Theorem (original)

Theorem 8 (Barron's Theorem, Barron (1993)) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function with Fourier transform $\hat{f}$ and "Barron parameter" $C_f := \left(\int_{w\in\mathbb{R}^n} \|w\| \cdot |\hat{f}(w)|\right)^2$. Let $\sigma(x)$ be a sigmoidal[4] nonlinearity. In particular, we will consider $\sigma(x) = \mathbb{1}(x \geq 0)$. Then, there exists a two-layer neural network $NN_{Barron}(x) = \sum_{i=1}^{k} \alpha_i \sigma(\langle w_i, x\rangle - b_i)$ with bounded approximation error:*

$$\int |f(x) - NN_{Barron}(x)|^2 \leq \frac{\Theta(C_f)}{k}$$

Proof Write f in terms of its Fourier coefficients as follows:

$$\begin{aligned} f(x) &= \mathrm{Re}\left(\int_w e^{2\pi\langle w,x\rangle}\hat{f}(w)dw\right) \\ &= \mathrm{Re}\left(\int_w e^{2\pi\langle w,x\rangle + 2\pi\measuredangle(w)|}\hat{f}(w)|dw\right) \\ &= \int_w \cos(2\pi\langle w,x\rangle + 2\pi\measuredangle(w))|\hat{f}(w)| \end{aligned}$$

4. A nonlinearity σ is sigmoidal if $\lim_{x\to-\infty}\sigma(x) = 0$ and $\lim_{x\to\infty}\sigma(x) = 1$.

Extended Abstract Track

Moreover, we can write $f(x) - f(0)$ as $\int_w (\cos(2\pi\langle w, x\rangle + 2\pi\measuredangle(w)) - \cos(2\pi\measuredangle(w)))|\hat{f}(w)|$, and algebraic manipulation yields

$$\cos(2\pi\langle w, x\rangle + 2\pi\measuredangle(w)) - \cos(2\pi\measuredangle(w)) \propto \int_0^{||w||} \mathbb{1}(w^T x - b \geq 0)\sin(2\pi b + 2\pi\measuredangle(w).$$

Finally, we apply Maurey's Lemma to argue that subsampling in w and b will not induce too much error:

Lemma 9 (Maurey's Lemma, Pisier (1981)) *Let x lie in a convex set $\mathcal{V}$. Then there exist k elements $v_1, \ldots, v_k \in \mathcal{V}$ and weights $\alpha_1, \ldots, \alpha_k$ such that*

$$||x - \sum_{i=1}^{k} \alpha_i v_i|| \leq \frac{\sup_{v\in\mathcal{V}} ||v||_2^2}{k}$$

Applying Maurey's Lemma to the set $\mathcal{V}$ indexed by w and $0 \leq b \leq ||w||$ of functions $\mathbb{1}(w^T x \geq b)$ yields the result. ■

A.2. Fourier structure of invariant and equivariant functions

In this subsection, we establish the Fourier structure of both invariant and equivariant functions. We begin with periodic functions, and consider the structure of their discrete Fourier series.

Lemma 10 *Let $G \leq S_n$, and let $f : [0,1]^n \to \mathbb{R}$ (or alternatively, let $f : \mathbb{R}^n \to \mathbb{R}$ be periodic on $[0,1]^n$. Then $\hat{f}(\boldsymbol{w}) = \hat{f}(g\boldsymbol{w})$ for all $g \in G$, $\boldsymbol{w} \in \mathbb{Z}^n$.*

Proof As before, the symmetry of f with respect to G is reflected in its Fourier coefficients as well:

$$\begin{aligned}
\hat{f}(gw) &= \int_{x\in[0,1]^n} f(x)e^{2\pi i\langle x, gw\rangle} \\
&= \int_{x\in[0,1]^n} f(x)e^{2\pi i\langle g^T x, w\rangle} \\
&= \int_{z\in[0,1]^n} f(gz)e^{2\pi i\langle z, w\rangle} \\
&= \int_{z\in[0,1]^n} f(z)e^{2\pi i\langle z, w\rangle} = \hat{f}(w)
\end{aligned}$$

Note that the domain of integration $[0,1]^n$ is invariant under permutations g, which allowed the change of variable $z = g^T x$. ■

Although we only use the discrete Fourier transform in the main body, it is worth noting for future work that this invariance structure in Fourier space holds for non-periodic functions as well.

Lemma 11 (Fourier characterization of invariant functions) *$f : \mathbb{R}^n \to \mathbb{R}$ is invariant to a group $G \leq O(n)$ iff the Fourier transform of f is also invariant, i.e. $\hat{f}(\boldsymbol{w}) = \hat{f}(g\boldsymbol{w})$ for all $\boldsymbol{w} \in \mathbb{R}^n$ and for all $g \in G$.*

Proof

$$\begin{aligned}
\mathcal{F}(f)(gw) &= \int_{x \in \mathbb{R}^n} e^{-2\pi i \langle gw, x \rangle} f(x) \\
&= \int_{x \in \mathbb{R}^n} e^{-2\pi i \langle w, g^T x \rangle} f(x) \\
&= \int_{z \in \mathbb{R}^n} e^{-2\pi i \langle w, z \rangle} f(z) \\
&= \mathcal{F}(f)(w)
\end{aligned}$$

■

Finally, vector-valued equivariant functions are also equivariant in Fourier space, in the following sense.

Lemma 12 (Fourier characterization of equivariant functions) *$f : \mathbb{R}^n \to \mathbb{R}^m$ is equivariant to a group $G \leq O(n)$ iff the entrywise Fourier transform $\hat{f} : \mathbb{R}^n \to \mathbb{R}^m$ is also equivariant, i.e. $g\hat{f}(\boldsymbol{w}) = \hat{f}(g\boldsymbol{w})$ for all $\boldsymbol{w} \in \mathbb{R}^n$ and for all $g \in G$.*

Proof

$$\begin{aligned}
\mathcal{F}(f)(gw)_s &= \int_{x \in \mathbb{R}^n} e^{-2\pi i \langle gw, x \rangle} f_s(x) \\
&= \int_{x \in \mathbb{R}^n} e^{-2\pi i \langle w, g^T x \rangle} f_s(x) \\
&= \int_{z \in \mathbb{R}^n} e^{-2\pi i \langle w, z \rangle} f_s(gz)
\end{aligned}$$

By the equivariance of f, the vector $f(gz)$ is equal to $gf(z)$. Therefore, we can interchange the entrywise integral and the group action to complete the proof:

$$\begin{aligned}
\mathcal{F}(f)(gw)_s &= \int_{z \in \mathbb{R}^n} e^{-2\pi i \langle w, z \rangle} f(gz) \\
&= \int_{z \in \mathbb{R}^n} e^{-2\pi i \langle w, z \rangle} gf(z) \\
&= g \int_{z \in \mathbb{R}^n} e^{-2\pi i \langle w, z \rangle} f(z) \\
&= g\mathcal{F}(f)(w)
\end{aligned}$$

■

Extended Abstract Track

A.3. Proof of Theorem 3

Consider the proof technique of Barron's Theorem applied in this setting, modified to preserve invariance. As before, let $\hat{f}(w) = |\hat{f}(w)|e^{2\pi i\theta(w)}$. We will often work with the orbits of $\mathbb{Z}^n$ with respect to G, denoted by $\mathcal{O}$. We denote individual orbits by $[w]$, where o is the orbit representative, i.e. $[w] = \{gw : g \in G\}$.

$$\begin{aligned} f(x) &= \sum_{w\in\mathbb{Z}^n} \hat{f}(w)e^{2\pi i\langle x,w\rangle} \\ &= \sum_{w\in\mathbb{Z}^n} |\hat{f}(w)|e^{2\pi i(\theta(w)+\langle x,w\rangle)} \\ &= \sum_{w\in\mathbb{Z}^n} |\hat{f}(w)|\cos(2\pi\theta(w) + 2\pi\langle x, w\rangle) \\ &= \sum_{[w]\in\mathcal{O}} \sum_{w\in[w]} |\hat{f}([w])|\cos(2\pi\theta([w]) + 2\pi\langle x, w\rangle) \end{aligned}$$

Lemma 13 (Cosine basis functions are orthogonal) *Let $w, z \in \mathbb{R}^n$ with $w_1, z_1 \in \mathbb{Z}$ and $w_1^2 \neq z_1^2$. Let x_1 indicate the first entry of a vector x, and $x_{2:}$ the rest of the entries. Then the following orthogonality relation holds for arbitrary functions $r, s : \mathbb{R}^{2n-2} \to \mathbb{R}$:*

$$\int_{x\in[0,1]^n} \cos(2\pi(x_1w_1 + r(x_{2:}, w_{2:}))) \cos(2\pi(x_1z_1 + s(x_{2:}, z_{2:}))) = 0$$

Proof We will heavily use the trigonometric identities $\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)$.

We will also rely on the standard trigonometric orthogonality statements, assuming n and m are integers with distinct absolute values:

$$\int_{x\in[0,1]} \sin(nx)\cos(mx) = 0$$
$$\int_{x\in[0,1]} \cos(2\pi nx)\cos(2\pi mx) = 0$$
$$\int_{x\in[0,1]} \sin(nx)\sin(mx) = 0$$

Recall the trigonometric identity $\cos\theta\cos\phi = \frac{\cos(\theta-\phi)+\cos(\theta+\phi)}{2}$. If $|n| \neq |m|$, then $\theta - \phi \neq 0$ and $\theta+\phi \neq 0$, and $\int_{x\in[0,1]} \cos(2\pi(n-m)x) = \int_{x\in[0,1]} \cos(2\pi(n+m)x) = 0$. Similar reasoning using the identities $\sin\theta\sin\phi = \frac{\cos(\theta-\phi)-\cos(\theta+\phi)}{2}$ and $\sin\theta\cos\phi = \frac{\sin(\theta-\phi)+\sin(\theta+\phi)}{2}$ yield the other two equalities.

To apply these one-dimensional identities in our setting, again apply the trigonometric identity for the cosine of a sum to decompose $\cos(2\pi(x_1w_1 + r(x_{2:}, w_{2:})))$, and likewise for

the analogous term with z. Therefore, we have the following:

$$\begin{aligned}
\cos(2\pi\langle x,w\rangle)\cos(2\pi\langle x,z\rangle) &= \cos(2\pi x_1 w_1 + 2\pi\langle x_{2:}, w_{2:}\rangle)\cos(2\pi x_1 z_1 + 2\pi\langle x_{2:}, z_{2:}\rangle) \\
&= (\cos(2\pi x_1 w_1)\cos(2\pi r(x_{2:}, w_{2:}) - \sin(2\pi x_1 w_1)\sin(2\pi r(x_{2:}, w_{2:}))\cdot \\
&\quad (\cos(2\pi x_1 z_1)\cos(2\pi s(x_{2:}, z_{2:}) - \sin(2\pi x_1 z_1)\sin(2\pi s(x_{2:}, z_{2:})) \\
&= \cos(2\pi x_1 w_1)\cos(2\pi r(x_{2:}, w_{2:})\cos(2\pi x_1 z_1)\cos(2\pi s(x_{2:}, z_{2:}))- \\
&\quad \cos(2\pi x_1 w_1)\cos(2\pi r(x_{2:}, w_{2:})\sin(2\pi x_1 z_1)\sin(2\pi s(x_{2:}, z_{2:}))- \\
&\quad \sin(2\pi x_1 w_1)\sin(2\pi r(x_{2:}, w_{2:}))\cos(2\pi x_1 z_1)\cos(2\pi s(x_{2:}, z_{2:}))+ \\
&\quad \sin(2\pi x_1 w_1)\sin(2\pi r(x_{2:}, w_{2:}))\sin(2\pi x_1 z_1)\sin(2\pi s(x_{2:}, z_{2:}))
\end{aligned}$$

Recall that the integral of a separable function is equal to the product of the integrals. Therefore, evaluating each term separately:

$$\begin{aligned}
&\int_{x\in[0,1]^n} \cos(2\pi x_1 w_1)\cos(2\pi r(x_{2:}, w_{2:})\cos(2\pi x_1 z_1)\cos(2\pi s(x_{2:}, z_{2:}) \\
=&\Big(\underbrace{\int_{x_1\in[0,1]} \cos(2\pi x_1 w_1)\cos(2\pi x_1 z_1)}_{=0}\Big)\Big(\int_{x_{2:}\in[0,1]^{n-1}} \cos(2\pi r(x_{2:}, w_{2:})\cos(2\pi s(x_{2:}, z_{2:})\Big) \\
&\int_{x\in[0,1]^n} \sin(2\pi x_1 w_1)\sin(2\pi r(x_{2:}, w_{2:})\cos(2\pi x_1 z_1)\cos(2\pi s(x_{2:}, z_{2:}) \\
=&\Big(\underbrace{\int_{x_1\in[0,1]} \sin(2\pi x_1 w_1)\cos(2\pi x_1 z_1)}_{=0}\Big)\Big(\int_{x_{2:}\in[0,1]^{n-1}} \sin(2\pi r(x_{2:}, w_{2:})\cos(2\pi s(x_{2:}, z_{2:})\Big) \\
&\int_{x\in[0,1]^n} \cos(2\pi x_1 w_1)\cos(2\pi r(x_{2:}, w_{2:})\sin(2\pi x_1 z_1)\sin(2\pi s(x_{2:}, z_{2:}) \\
=&\Big(\underbrace{\int_{x_1\in[0,1]} \cos(2\pi x_1 w_1)\sin(2\pi x_1 z_1)}_{=0}\Big)\Big(\int_{x_{2:}\in[0,1]^{n-1}} \cos(2\pi r(x_{2:}, w_{2:})\sin(2\pi s(x_{2:}, z_{2:})\Big) \\
=&\Big(\underbrace{\int_{x_1\in[0,1]} \sin(2\pi x_1 w_1)\sin(2\pi x_1 z_1)}_{=0}\Big)\Big(\int_{x_{2:}\in[0,1]^{n-1}} \sin(2\pi r(x_{2:}, w_{2:})\sin(2\pi s(x_{2:}, z_{2:})\Big).
\end{aligned}$$

■

Extended Abstract Track

Recall that we decomposed the symmetric periodic function f as

$$\begin{aligned} f(x) &= \sum_{w\in\mathbb{Z}^n} \hat{f}(w)e^{2\pi i\langle x,w\rangle} \\ &= \sum_{w\in\mathbb{Z}^n} |\hat{f}(w)|e^{2\pi i(\theta(w)+\langle x,w\rangle)} \\ &= \sum_{w\in\mathbb{Z}^n} |\hat{f}(w)|\cos(2\pi\theta(w)+2\pi\langle x,w\rangle) \\ &= \sum_{[w]\in\mathcal{O}} \sum_{w\in[w]} |\hat{f}([w])|\cos(2\pi\theta([w])+2\pi\langle x,w\rangle) \end{aligned}$$

As in the original Barron's Theorem, we can sample frequencies of f. However, sampling from $w \in \mathbb{Z}^n$ will yield a function that is not necessarily invariant. Instead, if we sample *orbits* of w, we will obtain an invariant approximation to f. Concretely, sample functions of the form $v_{[\boldsymbol{w}]} := \sum_{\boldsymbol{w}\in[\boldsymbol{w}]} \cos(2\pi\theta([w])+2\pi\langle x,w\rangle)$, which are indexed by $[w] \in \mathcal{O}$, proportionally to $|\hat{f}([w])|$. For any $w_1, w_2 \in \mathbb{Z}^n$, by Lemma 13,

$$\int_{x\in[0,1]^n} \cos(2\pi\theta(w_1)+2\pi\langle x,gw_1\rangle)\cos(2\pi\theta(w_2)+2\pi\langle x,hw_2\rangle) = 0 \tag{2}$$

so long as there is some index i such that $|(gw_1)_i|$ and $|(gw_2)_i|$ are distinct. This will not be the case for all invariant functions; the existence of such an i for any two Fourier modes with non-zero corresponding frequency is a structural condition that we impose on f. It is satisfied, for example, if $\hat{f}$ is only supported on frequencies $w \in \mathbb{Z}^n_+$.

Under this structural orbit condition, we may approximate $\hat{f}$ by sampling orbits over G in the frequency domain. By 2, these orbits consist of sums of orthogonal functions. Here, the inner product between two real-valued functions is defined by

$$\Big\langle f_1(x), f_2(x)\Big\rangle = \int_{x\in[0,1]^n} f_1(x)f_2(x)$$

As a result of the orthogonality properties above and the Pythagorean theorem, we have

$$||v_{[w]}||_2^2 = ||\sum_{w\in[w]} \cos(2\pi\theta([w])+2\pi\langle x,w\rangle)||_2^2 \tag{3}$$

$$= \sum_{w\in[w]} ||\cos(2\pi\theta([w])+2\pi\langle x,w\rangle)||_2^2 \tag{4}$$

Consider now the upper bound provided by Maurey's Lemma:

Lemma 14 *Let x lie in a convex set $\mathcal{V}$. Then there exist k elements $v_1,\ldots,v_k \in \mathcal{V}$ and weights $\alpha_1,\ldots,\alpha_k$ such that*

$$||x - \sum_{i=1}^k \alpha_i v_i||^2 \le \frac{\mathbb{E}[||x_1||^2]}{k},$$

where x is drawn from $\mathcal{V}$ proportionally to the convex coefficients of x.

Proof Since $x \in \mathcal{V}$, there exists a random variable u taking values in $\mathcal{V}$ such that $\mathbb{E}[u] = x$. Consider the squared error from approximating $\mathbb{E}[u] = x$ with k random draws of variables $x_1, \ldots, x_k \sim u$:

$$\begin{aligned}
\mathbb{E}[\|x - \frac{1}{k}\sum_{i=1}^{k} x_i\|^2] &= \frac{\sum_{i=1}^{k} \mathbb{E}[\|x_i - x\|^2]}{k^2} \\
&= \frac{\mathbb{E}[\|x_1 - x\|^2]}{k} \\
&= \frac{\mathbb{E}[||x_1||^2] + ||x||^2 - 2x^T\mathbb{E}[x_1]}{k} \\
&= \frac{\mathbb{E}[||x_1||^2] - ||x||^2}{k} \\
&\leq \frac{\mathbb{E}[||x_1||^2]}{k}
\end{aligned}$$

Therefore, there exists some drawing of the variables $x_1, \ldots, x_k$ satisfying this bound; the lemma follows. ■

Applied as in the original Barron's Theorem, Maurey's Lemma would construct $\mathcal{V}$ as $\{\mathcal{N} \cdot v_w\}$, where $v_w := \cos(2\pi\theta(w) + 2\pi\langle x, w\rangle)$ and $\mathcal{N}$ is a scaling factor, $\mathcal{N} = \sum_{\boldsymbol{w} \in \mathbb{Z}^n} |\hat{f}(w)| = \sum_{[\boldsymbol{w}] \in \mathcal{C}} |\hat{f}(\boldsymbol{w})| \cdot |[\boldsymbol{w}]|$.[5] Then, $f = \sum_{\boldsymbol{w} \in \mathbb{Z}^n} |\hat{f}(w)| v_w = \sum_{\boldsymbol{w} \in \mathbb{Z}^n} \frac{|\hat{f}(w)|}{\mathcal{N}} (\mathcal{N} v_w)$ is a convex combination of elements of $\mathcal{V}$. The numerator of the squared ℓ_2 error bound given by Maurey's Lemma is then:

$$\begin{aligned}
\mathbb{E}[||x||^2] &= \sum_{\boldsymbol{w} \in \mathbb{Z}^n} \frac{|\hat{f}(w)|}{\mathcal{N}} ||\mathcal{N} v_w||_2^2 \\
&= \sum_{\boldsymbol{w} \in \mathbb{Z}^n} \mathcal{N} |\hat{f}(w)| \cdot ||v_w||_2^2 \\
&= \sum_{[\boldsymbol{w}] \in \mathcal{O}} \mathcal{N} |\hat{f}([\boldsymbol{w}])| \sum_{\boldsymbol{w} \in [\boldsymbol{w}]} ||v_{\boldsymbol{w}}||_2^2 \\
&= \sum_{[\boldsymbol{w}] \in \mathcal{O}} \mathcal{N} |\hat{f}([\boldsymbol{w}])| \cdot |[\boldsymbol{w}]| \qquad (5) \\
&= \mathcal{N}^2
\end{aligned}$$

We now apply Maurey's Lemma with $\mathcal{V} = \{\mathcal{N} \cdot v_{[\boldsymbol{w}]} : [\boldsymbol{w}] \in \mathcal{O}\}$, where

$$f = \sum_{\boldsymbol{w} \in \mathbb{Z}^n} \frac{|\hat{f}(\boldsymbol{w})|}{\mathcal{N}} (\mathcal{N} v_w) = \sum_{[\boldsymbol{w}] \in \mathcal{O}} \frac{|\hat{f}([\boldsymbol{w}])| \cdot |[\boldsymbol{w}]|}{\mathcal{N}} \Big(\sum_{\boldsymbol{w} \in [\boldsymbol{w}]} \frac{\mathcal{N}}{|[\boldsymbol{w}]|} v_{\boldsymbol{w}} \Big).$$

5. Note that v_w is an individual frequency, while $v_{[w]}$ denotes a summed orbit of frequencies.

Extended Abstract Track

Note that $\frac{|\hat{f}([\boldsymbol{w}])|\cdot|[\boldsymbol{w}]|}{\mathcal{N}}$ is a measure: $\sum_{[\boldsymbol{w}]\in\mathcal{O}} \frac{|\hat{f}([\boldsymbol{w}])|\cdot|[\boldsymbol{w}]|}{\mathcal{N}} = \sum_{[\boldsymbol{w}]\in\mathcal{O}} \sum_{w\in[\boldsymbol{w}]} \frac{|\hat{f}([\boldsymbol{w}])|}{\mathcal{N}} = 1$. The term $\mathbb{E}[||x||_2^2]$ is given by:

$$
\begin{aligned}
\mathbb{E}[||x||^2] &= \sum_{[\boldsymbol{w}]\in\mathcal{O}} \frac{|\hat{f}([\boldsymbol{w}])|\cdot|[\boldsymbol{w}]|}{\mathcal{N}} \Big|\Big| \sum_{\boldsymbol{w}\in[\boldsymbol{w}]} \frac{\mathcal{N}}{|[\boldsymbol{w}]|} v_{\boldsymbol{w}} \Big|\Big|_2^2 \\
&= \sum_{[\boldsymbol{w}]\in\mathcal{O}} \frac{\mathcal{N}|\hat{f}([\boldsymbol{w}])|}{|[\boldsymbol{w}]|} \Big|\Big| \sum_{\boldsymbol{w}\in[\boldsymbol{w}]} v_{\boldsymbol{w}} \Big|\Big|_2^2 \\
&= \sum_{[\boldsymbol{w}]\in\mathcal{O}} \frac{\mathcal{N}|\hat{f}([\boldsymbol{w}])|}{|[\boldsymbol{w}]|} \sum_{\boldsymbol{w}\in[\boldsymbol{w}]} ||v_{\boldsymbol{w}}||_2^2 \\
&= \sum_{[\boldsymbol{w}]\in\mathcal{O}} \frac{\mathcal{N}|\hat{f}([\boldsymbol{w}])|}{|[\boldsymbol{w}]|} \cdot |[\boldsymbol{w}]| \\
&= \sum_{[\boldsymbol{w}]\in\mathcal{O}} \mathcal{N}|\hat{f}([\boldsymbol{w}])| \qquad (6)
\end{aligned}
$$

Comparing equations 6 and 5, it is clear that the improvement in the number of samples depends on the group G, and the support of f in Fourier space. For example, if $\hat{f}$ is supported only on orbits of equal size s, where $1 \leq s \leq |G|$, then 5 — what we'll call "original" Barron's theorem — is a factor of s larger than 6. In other words, the error induced by sampling k functions ($v_{\boldsymbol{w}}$ or $v_{[\boldsymbol{w}]}$) is a factor of s smaller when sampling $v_{[\boldsymbol{w}]}$, rather than $v_{\boldsymbol{w}}$. However, note that as neural network modules, $v_{[\boldsymbol{w}]}$ is always a factor of $|[\boldsymbol{w}]| \leq |G|$ larger than $v_{\boldsymbol{w}}$: in particular, $v_{[\boldsymbol{w}]} = \sum_{\boldsymbol{w}\in[\boldsymbol{w}]} \frac{\mathcal{N}}{|[\boldsymbol{w}]|} v_{\boldsymbol{w}}$. (For particular groups, it may be possible to implement this summation over an orbit more efficiently, but a naive implementation multiplies the network size by $|[\boldsymbol{w}]|$.) Therefore, a function (supported only on orbits of size s in Fourier space) that can be approximated to error ϵ using k individual frequencies, as in Barron's Theorem, can be approximated with an *invariant* neural network with sk nodes (or k orbit-averaged nodes) to error $\frac{\epsilon}{s}$. Note that for $k > s$, this implies that Barron's Theorem can be proven for *invariant* architectures of equal size (at least, when the function is supported on equal-sized orbits in Fourier space). In contrast, consider naively averaging the output of Barron's Theorem over the group G produces an invariant architecture, but at a factor of $|G|$ increase in network size.

A.4. Proof of Theorem 5

When the function is supported on unequally sized orbits in Fourier space, the previous analysis does not provide satisfactory control of the size of the network. This is because the probabilistic method used in Maurey's Lemma only proves the existence of k functions $v_{[\boldsymbol{w}]}$, but their sizes may vary depending on the orbit size $|[\boldsymbol{w}]|$. Instead, as a more precise analysis, we can evaluate the following quantity, which is the product of the approximation error and the network size. Here, let s_i denote the size of the orbit selected by the random

variable x_i.

$$\begin{aligned}
\mathbb{E}\left[\left|\left|x-\frac{1}{k}\sum_{i=1}^{k}x_i\right|\right|^2\Big(\sum_{i=1}^{k}s_i\Big)\right] &= \frac{1}{k^2}\mathbb{E}\left[\Big(\sum_{i,j=1}^{k}\langle x-x_i, x-x_j\rangle\Big)\Big(\sum_{\ell=1}^{k}s_\ell\Big)\right] \\
&= \frac{1}{k^2}\sum_{i,j=1}^{k}\mathbb{E}\left[\langle x-x_i, x-x_j\rangle\Big(\sum_{\ell=1}^{k}s_\ell\Big)\right] \\
&= \frac{1}{k^2}\sum_{i,j=1}^{k}\mathbb{E}\left[\langle x-x_i, x-x_j\rangle\Big(s_i+s_j+\sum_{\ell\neq i,j}s_\ell\Big)\right] \\
&= \frac{1}{k^2}\sum_{i,j=1}^{k}\mathbb{E}\left[\langle x-x_i, x-x_j\rangle\Big(s_i+s_j+\sum_{\ell\neq i,j}s_\ell\Big)\right] \\
&= \frac{1}{k^2}\sum_{i,j=1}^{k}\mathbb{E}\left[\langle x-x_i, x-x_j\rangle\Big(s_i+s_j\Big)\right] \\
&\quad + \mathbb{E}\left[\Big(\sum_{\ell\neq i,j}s_\ell\Big)\right]\mathbb{E}\left[\langle x-x_i, x-x_j\rangle\right] \qquad (7)
\end{aligned}$$

Note that $\mathbb{E}[\langle x-x_i, x-x_j\rangle] = ||x||^2 - 2||x||^2 + ||x||^2 = 0$ for $i \neq j$. Similarly, when $i \neq j$:

$$\begin{aligned}
\mathbb{E}[s_i\langle x-x_i, x-x_j\rangle] &= \mathbb{E}[s_i]||x||^2 - \mathbb{E}[s_i]x^T\mathbb{E}[x_j] - x^T\mathbb{E}[s_ix_i] + \mathbb{E}[x_j^T]\mathbb{E}[s_ix_i] \\
&= \mathbb{E}[s_i]||x||^2 - \mathbb{E}[s_i]||x||^2 - x^T\mathbb{E}[s_ix_i] + x^T\mathbb{E}[s_ix_i] \qquad = 0
\end{aligned}$$

Therefore, (7) vanishes when $i \neq j$, and defining the constant value $S_{k-1} := \mathbb{E}[\sum_{\ell\neq i}s_\ell]$, we can simplify (7) to:

$$\begin{aligned}
&\frac{1}{k^2}\sum_{i,j=1}^{k}\mathbb{E}\left[\langle x-x_i, x-x_j\rangle\Big(s_i+s_j\Big)\right] + \mathbb{E}\left[\Big(\sum_{\ell\neq i,j}s_\ell\Big)\right]\mathbb{E}\left[\langle x-x_i, x-x_j\rangle\right] \\
=&\frac{1}{k^2}\sum_{i=1}^{k}\mathbb{E}[||x-x_i||^2s_i] + S_{k-1}\mathbb{E}[||x-x_i||^2] \\
=&\frac{1}{k^2}\sum_{i=1}^{k}\mathbb{E}[s_i]||x||^2 - 2\langle x, \mathbb{E}[x_is_i]\rangle + \mathbb{E}[s_i||x_i||^2] + S_{k-1}||x||^2 - 2S_{k-1}||x||^2 + S_{k-1}\mathbb{E}[||x_i||^2] \\
=&\frac{1}{k^2}\sum_{i=1}^{k}\mathbb{E}[s_i]||x||^2 - 2\langle x, \mathbb{E}[x_is_i]\rangle + \mathbb{E}[s_i||x_i||^2] - S_{k-1}||x||^2 + S_{k-1}\mathbb{E}[||x_i||^2] \\
=&\frac{1}{k}\left[\mathbb{E}[s_1]||x||^2 - 2\langle x, \mathbb{E}[x_1s_1]\rangle + \mathbb{E}[s_1||x_1||^2] - S_{k-1}||x||^2 + S_{k-1}\mathbb{E}[||x_1||^2]\right]
\end{aligned}$$

Finally, note that

$$\mathbb{E}[s_1] = \sum_{[\boldsymbol{w}]\in\mathcal{O}}\frac{|\hat{f}([\boldsymbol{w}])|\cdot|[\boldsymbol{w}]|^2}{\mathcal{N}}, \qquad (8)$$

Extended Abstract Track

and moreover,

$$\begin{aligned}\mathbb{E}[||x_1||^2] &= \sum_{[\boldsymbol{w}]\in\mathcal{O}} \frac{|\hat{f}([\boldsymbol{w}])| \cdot |[\boldsymbol{w}]|}{\mathcal{N}} ||\mathcal{N} v_{[\boldsymbol{w}]}||^2 \\ &= \mathcal{N} \sum_{[\boldsymbol{w}]\in\mathcal{O}} |\hat{f}([\boldsymbol{w}])| \cdot |[\boldsymbol{w}]| ||v_{[\boldsymbol{w}]}||^2 \\ &= \mathcal{N} \sum_{[\boldsymbol{w}]\in\mathcal{O}} |\hat{f}([\boldsymbol{w}])| \cdot |[\boldsymbol{w}]|^2 \\ &= \mathcal{N}^2 \mathbb{E}[s_1].\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}[s_1||x_1||^2] &= \sum_{[\boldsymbol{w}]\in\mathcal{O}} \frac{|\hat{f}([\boldsymbol{w}])| \cdot |[\boldsymbol{w}]|^2}{\mathcal{N}} ||\mathcal{N} v_{[\boldsymbol{w}]}||^2 \\ &= \mathcal{N} \sum_{[\boldsymbol{w}]\in\mathcal{O}} |\hat{f}([\boldsymbol{w}])| \cdot |[\boldsymbol{w}]|^2 ||v_{[\boldsymbol{w}]}||^2 \\ &= \mathcal{N} \sum_{[\boldsymbol{w}]\in\mathcal{O}} |\hat{f}([\boldsymbol{w}])| \cdot |[\boldsymbol{w}]|^3 \\ &= \mathcal{N}^2 \mathbb{E}[s_1^2]\end{aligned}$$

Combining with $S_{k-1} = (k-1)\mathbb{E}[s_1]$, we finally obtain

$$\begin{aligned}&\mathbb{E}[s_1]||x||^2 - 2\langle x, \mathbb{E}[x_1 s_1]\rangle + \mathbb{E}[s_1||x_1||^2] - S_{k-1}||x||^2 + S_{k-1}\mathbb{E}[||x_1||^2] \\ =&\mathbb{E}[s_1]||x||^2 - 2\langle x, \mathbb{E}[x_1 s_1]\rangle + \mathbb{E}[s_1||x_1||^2] - (k-1)\mathbb{E}[s_1]||x||^2 + (k-1)\mathbb{E}[s_1]\mathcal{N}^2\mathbb{E}[s_1] \\ =&(2-k)\mathbb{E}[s_1]||x||^2 - 2\langle x, \mathbb{E}[x_1 s_1]\rangle + \mathcal{N}^2\mathbb{E}[s_1^2] + (k-1)\mathbb{E}[s_1]^2\mathcal{N}^2\end{aligned}$$

Note that, when s_1 is a constant, the first two terms simplify to a single negative term; in the upper bound from Maurey's lemma, this term is dropped.

A.5. Group-averaged Barron's theorem

We begin with the case of invariance, and have the following theorem.

Theorem 15 *Let G be a finite subgroup of $O(n)$ and let $\subseteq \mathbb{R}^n$ be a space*[6] *on which G acts. Furthermore, suppose $g =$ for all $g \in G$. Let $f :\to \mathbb{R}$ be a G-invariant function. Then there exists a two-layer group convolutional network NN_{G-CNN}, with sigmoidal nonlinearity, $\Theta(|G| \cdot k)$ units, and approximation error bounded as follows:*

$$\int_{x\in} |f(x) - NN_{G-CNN}(x)|^2 \leq \frac{C_f}{k}$$

Here, C_f is the Barron constant: $C_f = \int_{w\in\mathbb{R}^n} \|w\| \cdot |\hat{f}(w)|$, minimized over all extensions of f to $\mathbb{R}^n$.

6. If is a homogeneous space, then f is a trivial function, so we think of as a non-homogeneous space on which G acts.

Remark 16 *In contrast to existing invariant universal approximation results, such as those of Bogatskiy et al. (2020) and Yarotsky (2018), Theorem 15 provides an* explicit *and moderate bound on the size of the approximating networks in terms of the group size. It avoids both exponential dependence on dimension, which arises in any blackbox call to traditional approximation results, and any dependence on the number of invariants needed for Hilbert's theorem (both of which are central to Yarotsky (2018) and the works which build upon it.)*

We now prove Theorem 15.

Proof

Let be a non-homogeneous space on which G, a subgroup of the orthogonal group, acts, and suppose $g =$ for all $g \in G$. Let $f :\to \mathbb{R}$. Moreover, suppose f is invariant, i.e. $f(gx) = f(x)$ for all $g \in G$, $x \in$. By Barron's Theorem, for any ϵ there exists a depth two ReLU net

$$NN_{Barron}(x) = \boldsymbol{\alpha}^T \sigma(Wx - b)$$

with $h = O(\frac{C_f}{\epsilon})$ hidden units (i.e. $\boldsymbol{\alpha} \in \mathbb{R}^h$) such that

$$\int |f(x) - NN_{Barron}(x)|^2 dx \leq \epsilon$$

where $C_f = \left(\int ||w|| \cdot |\hat{f}(w)| dw \right)^2$. Call the ith row of the matrix W w_i^T. Then define the following group-convolutional network:

$$NN_{G-CNN}(x) = \frac{1}{|G|} \sum_i \boldsymbol{\alpha}_i \sum_{g \in G} \sigma((w_i * x)(g) - b_i)$$

NN_{G-CNN} is an h-channel group convolutional network with a sigmoidal nonlinearity, which pools over both group elements and channels.

$$\begin{aligned} NN_{G-CNN}(x) &= \frac{1}{|G|} \sum_i \boldsymbol{\alpha}_i \sum_{g \in G} \sigma((w_i * x)(g) - b_i) \\ &= \frac{1}{|G|} \sum_{g \in G} \sum_i \boldsymbol{\alpha}_i \sigma(w_i^T gx - b_i) \\ &= \frac{1}{|G|} \sum_{g \in G} \boldsymbol{\alpha}^T \sigma(Wgx - b) \\ &= \frac{1}{|G|} \sum_{g \in G} NN_{Barron}(gx) \end{aligned}$$

Extended Abstract Track

For any function f, let $f^g(x) = f(gx)$. Although $NN_{Barron}(gx) \neq NN_{Barron}(x)$, we can extend its error guarantee to NN_{G-CNN} as follows:

$$\begin{aligned}\left(\int |f(x) - NN_{G-CNN}(x)|^2\right)^{1/2} dx &= ||f - NN_{G-CNN}|| \\ &= ||\frac{1}{|G|}\sum_{g\in G} f^g - \frac{1}{|G|}\sum_{g\in G} NN^g_{Barron}|| \\ &\leq \frac{1}{|G|}\sum_{g\in G} ||f^g - NN^g_{Barron}|| \\ &= \frac{1}{|G|}\sum_{g\in G} ||f - NN_{Barron}|| \\ &\leq \sqrt{\epsilon}\end{aligned}$$

Here, we used Jensen's inequality and the integration change of variable $x \mapsto gx$, enabled by the orthogonality of g and the assumption $= g$ for all g. ■

By appropriately modifying the final linear layer of the NN_{G-CNN} architecture described above, we obtain the following analogous theorem for *equivariant* functions.

Theorem 17 *Let G be a finite subgroup of $O(n)$ acting orthogonally on $\mathbb{R}^n$, as well as acting on $\mathbb{R}^m$ via a linear representation ρ. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a G-equivariant function*[7]*. Then there exists a two-layer group convolutional network NN_{G-CNN}, with sigmoidal nonlinearity, with $\Theta(|G| \cdot k)$ units and approximation error bounded as follows:*

$$\int_{x\in\mathbb{R}^n} \|f(x) - NN_{G-CNN}(x)\|^2 \leq \frac{D^2}{k}\sum_s C_f^s$$

Here, D is a uniform bound on the operator norm of the matrix $\rho(g)$ over all $g \in G$, and C_f^s is the Barron parameter of f_s.

Proof The architecture of Theorem 15 averages the approximation given by Barron's Theorem over the group. For the equivariant case, it is then natural to approximate each output dimension by Barron's Theorem, and then to average in an equivariant way. This will contribute an additional layer, to apply the representation ρ.

Formally, begin by approximating the sth output of f as NN^s_{Barron}, defined by

$$NN_{Barron}(x)^s = (\alpha^s)^T\sigma(W^s x - b^s) = \sum_i \alpha_i^s \sigma(\langle w_i^s, x\rangle - b^s).$$

Letting C_f^s denote the Barron parameter of f_s, NN^s_{Barron} has $\Theta(k)$ hidden nodes and squared L_2 approximation error upper bounded by $\Theta(\frac{C_f^s}{k})$. Then, define NN_{G-CNN} as follows:

$$NN_{G-CNN}(x) := \frac{1}{|G|}\sum_{g\in G} \rho(g^{-1}) NN_{Barron}(gx)$$

7. As in Theorem 15, we could again state this theorem for subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$ on which G acts, but do not do so for simplicity.

One can verify that NN_{G-CNN} is equivariant:

$$\begin{aligned} NN_{G-CNN}(hx) &= \frac{1}{|G|}\sum_{g\in G}\rho(g^{-1})NN_{Barron}(ghx) \\ &= \frac{1}{|G|}\sum_{k\in G}\rho(hk^{-1})NN_{Barron}(kx) \\ &= \frac{1}{|G|}\sum_{k\in G}\rho(h)\rho(k^{-1})NN_{Barron}(kx) \\ &= \rho(h)NN_{G-CNN}(x) \end{aligned}$$

Moreover, its error is bounded as follows:

$$\begin{aligned} \left(\int_{x\in\mathbb{R}^n}\|f(x)-NN_{G-CNN}(x)\|^2\right)^{1/2} &= \left(\int_{x\in\mathbb{R}^n}\left\|f(x)-\frac{1}{|G|}\sum_{g\in G}\rho(g^{-1})NN_{Barron}(gx)\right\|^2\right)^{1/2} \\ &= \left(\int_{x\in\mathbb{R}^n}\left\|\frac{1}{|G|}\sum_{g\in G}\Big(\rho(g^{-1})f(gx)-\rho(g^{-1})NN_{Barron}(gx)\Big)\right\|^2\right)^{1/2} \\ &\le \frac{1}{|G|}\sum_{g\in G}\left(\int_{x\in\mathbb{R}^n}\left\|\rho(g^{-1})f(gx)-\rho(g^{-1})NN_{Barron}(gx)\right\|^2\right)^{1/2} \\ &\le \frac{D}{|G|}\sum_{g\in G}\left(\int_{x\in\mathbb{R}^n}\left\|f(gx)-NN_{Barron}(gx)\right\|^2\right)^{1/2} \\ &= \frac{D}{|G|}\sum_{g\in G}\left(\int_{x\in\mathbb{R}^n}\left\|f(x)-NN_{Barron}(x)\right\|^2\right)^{1/2} \\ &\le D\left(\sum_{s=1}^{m}\int_{x\in\mathbb{R}^n}\left|f^s(x)-NN^s_{Barron}(x)\right|^2\right)^{1/2} \\ &\le D\left(\sum_{s=1}^{m}\frac{C_f^s}{k}\right)^{1/2} \end{aligned}$$

■

Remark 18 *Theorem 15 and Theorem 17 are stated for finite groups, but the proofs clearly hold for infinite G, with each summation replaced by an integral over G with respect to the Haar measure.*

As a result of Theorem 15, the approximation error of NN_{G-CNN} is equal to that of NN_{Barron}, but it is representable as a group-convolutional network with h channels. This demonstrates not only that Barron's Theorem extends to show universality of group-convolutional networks for invariant functions, but more generally that the orbit-averaged version of any two-layer net is representable as a group-convolutional net.

Extended Abstract Track

Unfortunately, the architecture requires convolving over G, which either contributes a factor of $|G|$ additional neurons for a large finite group, or is not (obviously) computable and requires special treatment, such as approximating the convolution in group Fourier space for infinite groups. However, another way of interpreting this architecture is as replacing its first nonlinear layer, $\sigma(Wx-b)$, with a group-averaged version $\frac{1}{|G|}\sum_{g\in G}\sigma(Wgx-b)$. Theorem 7 (proven in Appendix A.6) demonstrates a special case in which this expression can be computed analytically, thereby reducing the effective number of neurons in Theorem 15 from $|G|\cdot k$ to k.

A.6. Proof of Theorem 7

The previous section demonstrated that any invariant function f with Barron parameter C_f is representable by an *invariant* neural net, whose size scales with $|G|$ but is not exponential in the input dimension. The group-averaged architecture can be understood as an ordinary two layer architecture with final pooling layer, but which crucially replaces the first layer — usually composed of a linear layer with a pointwise nonlinearity — with a group-averaged nonlinear function $\bar{\sigma}_{w,b}(x) = \int_{g\in G}\sigma(Wgx-b)$. It is this dependence on $|G|$ which contributes adversarially to the network size; if G is large, such as S_n with $|S_n| = n!$, or even infinite, such as $G = SO(n)$, this integral is intractable. It would therefore be of both theoretical and practical interest to devise schemes for computing $\bar{\sigma}_{w,b}(x)$ for common groups G in an automatically differentiable and efficient manner.

Failure of Polynomials To alleviate this issue, one might hope to approximate the nonlinearity (cosine, σ, indicator, or general sigmoidal) with a polynomial over the range of inputs. This is because bounded-degree polynomials can be easily and analytically averaged over compact groups as follows. First, write a polynomial of degree d in terms of its homogeneous parts (using $\cdot$ to denote inner products between higher-dimensional tensors):

$$p(x) = \sum_{i=0}^{d} C_i \cdot x^{\otimes i}.$$

Then, integration over the group takes a simple form:

$$\begin{aligned}\int_g p(gx) &= \int_g \sum_{i=0}^{d} C_i \cdot (gx)^{\otimes i} \\ &= \sum_{i=0}^{d} C_i \cdot \Big(\int_g g^{\otimes i}\Big) \cdot x^{\otimes i}\end{aligned}$$

Note then that $\int_g g^{\otimes i}$ is simply the Reynolds operator corresponding to the ith tensor product representation of the group. Using the Clebsch-Gordan decomposition, it is possible to decompose $g^{\otimes i}$ into irreducible representations, which are easy to integrate analytically (in particular, as an orthonormal basis, they always integrate to 0 with the exception of the trivial irreducible representation). Unfortunately, it is crucial for universal approximation results to use *non-polynomial* non-linearities. By approximating a non-polynomial nonlinearity with a finite-degree polynomial, we impede universal approximation and, in

fact, can approximate multidimensional invariant functions no better than a multivariate group-averaged Taylor expansion.

We return, then, to the problem of exactly computing $\bar{\sigma}_{w,b}(x)$. This will depend greatly on σ and G, and in fact, may not always be possible. However, we show that for $\sigma(x) = \mathbb{1}(x \geq 0)$ and $G = SO(n)$, the function $\bar{\sigma}_{W,b}$ is easily expressible in closed form.

Theorem 19 *When averaged over the special orthogonal group in n dimensions, the non-linear invariant function $\bar{\mathbb{1}}_{w,b}(x)$ is expressible as follows:*

$$\bar{\mathbb{1}}_{w,b}(x) := \int_{g \in SO(n)} \mathbb{1}(w^T gx - b \geq 0) = \frac{1}{2} I_{\sin(\phi)^2}(\frac{n-1}{2}, \frac{1}{2}).$$

Here, $\phi = \cos^{-1}(\frac{b}{||w||\cdot||x||})$ and $I_a(b,c)$ denotes the regularized incomplete beta function.

Proof Consider a fixed x, w, and b. Define an equivalence relation on $SO(n)$ with respect to x: $g \sim h$ if $gx = hx$. Therefore, if $g \sim h$, then $\mathbb{1}(w^T gx - b) = \mathbb{1}(w^T hx - b)$. The equivalence classes, denoted $[g]$, of $SO(n)$, can then be associated with vectors of length equal to $||x||$. Moreover, it is possible to integrate over a given equivalence class $[g]$ using the Haar measure on $SO(n-1)$. (Intuitively, any rotation in $[g]$ simply must rotate x to a specified n-dimensional vector, and then can perform any $n-1$ dimensional rotation in the orthogonal subspace to this vector.) Denote this set by $B(0, ||x||)$. Then, the following equality holds:

$$\begin{aligned}\int_{g \in SO(n)} \mathbb{1}(w^T gx - b) &= \int_{v \in B(0,||x||)} \int_{g \in SO(n) \text{ s.t. } gx = v} \mathbb{1}(w^T gx - b \geq 0) \\ &= \int_{v \in B(0,||x||)} \mathbb{1}(w^T v - b \geq 0)\end{aligned}$$

Here, the integration over $B(0, ||x||)$ is with respect to the uniform probability measure on the sphere $B(0, ||x||)$ in n dimensions. Moreover, the final expression has a simple geometrical interpretation, which will give rise to the stated analytic form. In particular, note the following:

$$\begin{aligned}\mathbb{1}(w^T v - b) = 1 &\iff w^T v \geq b \\ &\iff \cos \measuredangle(w, v) \geq \frac{b}{||w|| \cdot ||v||}\end{aligned}$$

Therefore, the desired integral is simply equal to the fraction of the surface area of $B(0, ||x||)$ lying in the spherical cap of angle $\phi := \cos^{-1}(\frac{b}{||w||\cdot||x||})$. Using the analytic form of the surface area of high-dimensional spherical caps, expressed neatly in terms of the incomplete beta function by Li (2011), yields the following expression for this integral:

$$\frac{\pi^{n/2} r^{n-1}}{\Gamma(\frac{n}{2})} I_{\sin(\phi)^2}(\frac{n-1}{2}, \frac{1}{2})$$

Finally, recall that the surface area of an r-radius sphere in n dimensions is given by $\frac{2\pi^{n/2} r^{n-1}}{\Gamma(n/2)}$. As a result, the fraction of $B(0, \|x\|)$ lying in the relevant cap is given by

$$\frac{1}{2} I_{\sin(\phi)^2}(\frac{n-1}{2}, \frac{1}{2})$$

Extended Abstract Track

■

Unfortunately, the regularized incomplete beta functions are not themselves easily expressible as small neural networks with standard nonlinearities. However, this quantity is easily computable using predefined special functions in many programming languages. If computed in an automatically differentiable manner, this makes practical the equivariant architecture of the previous section for $SO(n)$, an infinite group — albeit with a specialized nonlinearity.