Capacity of Group-invariant Linear Readouts from Equivariant Representations: How Many Objects can be Linearly Classified Under All Possible Views?

Matthew Farrell* MSFARRELL@SEAS.HARVARD.EDU
Blake Bordelon* BLAKE_BORDELON@SEAS.HARVARD.EDU
Harvard University

Shubhendu Trivedi SHUBHENDU@CSAIL.MIT.EDU
Massachusetts Institute of Technology

Cengiz Pehlevan CPEHLEVAN@SEAS.HARVARD.EDU
Harvard University

Editors: Sophia Sanborn, Christian Shewmake, Simone Azeglio, Arianna Di Bernardo, Nina Miolane

1. Introduction

The ability to robustly categorize objects under conditions and transformations that preserve the object categories is essential to animal intelligence, and to pursuits of practical importance such as improving computer vision systems. One way to achieve such robustness is equivariance. When such transformations are restricted to be an algebraic group, the resulting equivariant representations have found significant success in machine learning starting with classical convolutional neural networks (CNNs) (Denker et al., 1989; LeCun et al., 1989) and recently being generalized by the influential work of Cohen and Welling (2016).

While it is clear that equivariance imposes a strong constraint on the geometry of representations, the implications of such constraints on model expressivity are not well understood. In this work we take a step toward addressing this gap.

Our particular contributions are the following:

- We extend Cover's function counting theorem (a measure of expressivity) to equivariant representations, finding that expressivity scales with the dimension of the subspace fixed by the group action.

- We demonstrate the applicability of our result to G-convolutional network layers – including pooling layers – through theory and verify through simulation.

This abstract is based on recently published work (Farrell et al., 2022).

* These authors contributed equally

2. Problem formulation

Suppose $\boldsymbol{x}$ abstractly represents an object and let $\boldsymbol{r}(\boldsymbol{x}) \in \mathbb{R}^N$ be some feature map of $\boldsymbol{x}$ to an N-dimensional space (such as an intermediate layer of a deep neural network). We consider transformations of this object such that they form a group G in the algebraic sense of the word. We denote the abstract transformation of $\boldsymbol{x}$ by element $g \in G$ as $g\boldsymbol{x}$. Groups G may be represented by invertible matrices, which act on a vector space V (which themselves form the group $GL(V)$ of invertible linear transformations on V). We are interested in feature maps $\boldsymbol{r}$ which satisfy the following group equivariance condition:

$$\boldsymbol{r}(g\boldsymbol{x}) = \pi(g)\boldsymbol{r}(\boldsymbol{x}),$$

where $\pi : G \to GL(\mathbb{R}^N)$ is a linear **representation** of G which acts on feature map $\boldsymbol{r}(\boldsymbol{x})$. Note that many representations of G are possible, including the trivial representation: $\pi(g) = \boldsymbol{I}$ for all g.

We are interested in perceptual object manifolds generated by the actions of G. Each of the P manifolds can be written as a set of points $\{\pi(g)\boldsymbol{r}^\mu : g \in G\}$ where $\mu \in [P] \equiv \{1, 2, \dots, P\}$; that is, these manifolds are orbits of the point $\boldsymbol{r}^\mu \equiv \boldsymbol{r}(\boldsymbol{x}^\mu)$ under the action of π. We will refer to such manifolds as **π-manifolds**.

Each of these π-manifolds represents a single object under the transformation encoded by π; hence, each of the points in a π-manifold is assigned the same class label. To measure the expressivity of this representation, we consider a perceptron endowed with a set of linear readout weights $\boldsymbol{w}$ that attempts to determine the correct class of every point in every manifold. The condition for realizing (i.e. linearly separating) the dichotomy $\{y^\mu\}_\mu$ can be written as $y^\mu \boldsymbol{w}^\top \pi(g)\boldsymbol{r}^\mu > 0$ for all $g \in G$ and $\mu \in [P]$, where $y^\mu = +1$ if the μ^{th} manifold belongs to the first class and $y^\mu = -1$ if the μ^{th} manifold belongs to the second class. The **perceptron capacity** is the fraction of dichotomies that can be linearly separated; that is, separated by a hyperplane.

The proofs for results in the main text are in Appendix A.2.

3. Separability of π-manifolds

We begin with a lemma which states that classifying the P π-manifolds can be reduced to the problem of classifying their P centroids. Let $\pi : G \to \mathrm{GL}(\mathbb{R}^N)$ be an arbitrary linear representation of a compact group G.[1] We denote the average of π over G with respect to the Haar measure by $\langle \pi(g) \rangle_{g \in G}$; for finite G this is simply $\frac{1}{|G|} \sum_{g \in G} \pi(g)$ where $|G|$ is the order (i.e. number of elements) of G. For ease of notation we will generally write $\langle \pi \rangle \equiv \langle \pi(g) \rangle_{g \in G}$ when the group G being averaged over is clear.

Lemma 1 *A dataset $\{(\pi(g)\boldsymbol{r}^\mu, y^\mu)\}_{g \in G, \mu \in [P]}$ consisting of P π-manifolds with labels y^μ is linearly separable if and only if the dataset $\{(\langle \pi \rangle \boldsymbol{r}^\mu, y^\mu)\}_{\mu \in [P]}$ consisting of the P centroids $\langle \pi \rangle \boldsymbol{r}^\mu$ with the same labels is linearly separable.*

1. Note that our results extend to more general vector spaces than $\mathbb{R}^N$, under the condition that the group be semi-simple.

3.1. Relationship with Cover's Theorem

The fraction f of linearly separable dichotomies on a dataset of size P for datapoints in general position[2] in N dimensions was computed by Cover (1965) and takes the form $f(P,N) = 2^{1-P}\sum_{k=0}^{N-1}\binom{P-1}{k}$ where we take $\binom{n}{m} = 0$ for $m > n$. For fixed N, $f(P,N)$ is a nonincreasing sigmoidal function of P with inflection point at $P = 2N$. We take this inflection point to be the **capacity** of the system, noting that as $N \to \infty$ the sigmoid sharpens into a step function.

Theorem 2 *Suppose the points $\langle\pi\rangle \boldsymbol{r}^\mu$ for $\mu \in [P]$ lie in general position in the subspace $V_0 = \text{range}(\langle\pi\rangle) = \{\langle\pi\rangle \boldsymbol{x} : \boldsymbol{x} \in V\}$. Then V_0 is the fixed point subspace of π, and the fraction of linearly separable dichotomies on the P π-manifolds $\{\pi(g)\boldsymbol{r}^\mu : g \in G\}$ is $f(P, N_0)$, where $N_0 = \dim V_0$.*

The **fixed point subspace** is the subspace $W = \{w \in V | gw = w \ ,\forall g \in G\}$. The condition that the points $\langle\pi\rangle \boldsymbol{r}^\mu$ be in general position essentially means that there is no prescribed special relationship between the $\boldsymbol{r}^\mu$ and between the $\boldsymbol{r}^\mu$ and $\langle\pi\rangle$. Taking the $\boldsymbol{r}^\mu$ to be drawn from a full-rank Gaussian distribution is sufficient to satisfy this condition.

3.2. The regular representation of Z_m

First we illustrate the theory in the case of the cyclic group $G = Z_m$ on m elements. This group is isomorphic to the group of integers $\{0, 1, ..., m-1\}$ under addition modulo m, and this is the form of the group that we will consider. An example of this group acting on an object is an image that is shifted pixel-wise to the left and right, with periodic boundaries. Suppose $\pi : Z_m \to \text{GL}(V)$ is the representation of Z_m consisting of the cyclic shift permutation matrices (this is called the **regular representation** of Z_m). In this case $V = \mathbb{R}^m$ and $\pi(g)$ is the matrix that cyclically shifts the entries of a length-m vector g places. For instance, if $m = 3$ and $\boldsymbol{v} = (1, 2, 3)$ then $\pi(2)\boldsymbol{v} = (2, 3, 1)$. The average of the regular representation matrix is $\langle\pi\rangle = \frac{1}{|G|}\mathbf{1}_m\mathbf{1}_m^\top$, indicating that $\langle\pi\rangle$ projects data along $\mathbf{1}_m$. See Appendix A.3 for more illustrative examples of group representations and their capacities.

4. G-Equivariant Neural Networks

The proposed theory can shed light on the feature spaces induced by G-CNNs. Consider a single convolutional layer feature map for a finite group G with the following activation:

$$a_{i,k}(\boldsymbol{x}) = \phi(\boldsymbol{w}_i^\top g_k^{-1}\boldsymbol{x}) \ , \ g_k \in G \ , \ i \in \{1, ..., N\} \tag{1}$$

for some nonlinear function ϕ. For each filter i, and under certain choices of π, the feature map $\boldsymbol{a}_i(\boldsymbol{x}) \in \mathbb{R}^{|G|}$ exhibits the equivariance property $\boldsymbol{a}_i(g_k\boldsymbol{x}) = \pi(g_k)\boldsymbol{a}_i(\boldsymbol{x})$. We will let $\boldsymbol{a}(x) \in \mathbb{R}^{|G|N}$ denote a flattened vector for this feature map.

2. A set of P points is in **general position** in N-space if every subset of N or fewer points is linearly independent. This says that the points are "generic" in the sense that there aren't any prescribed special linear relationships between them beyond lying in an N-dimensional space. Points drawn from a Gaussian distribution with full-rank covariance are in general position with probability one.

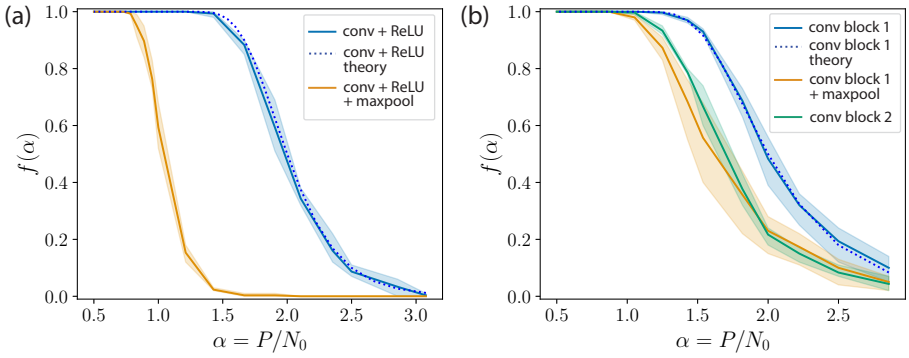

Figure 1: Capacity of GCNN representations. Solid lines denote the empirically measured fraction $f(\alpha)$ of 100 random dichotomies for which a logistic regression classifier finds a separating hyperplane, where $\alpha = P/N_0$. Dotted lines denote theoretical predictions. Shaded regions depict 95% confidence intervals over random choice of inputs, as well as network weights in (a) and (c). (a) $f(\alpha)$ of a random periodic convolutional layer after ReLU (blue line) and followed by 2x2 max pool (orange line), with $P = 40$ and $N_0 = \#$ output channels. (b) $f(\alpha)$ of VGG-11 pretrained on CIFAR-10 after a periodic convolution, batchnorm, and ReLU (blue line), followed by a 2x2 maxpool (orange line), and then another set of convolution, batchnorm, and ReLU (green line), with $P = 20$ and $N_0 = \#$ output channels.

In a traditional periodic convolutional layer applied to inputs of width W and length L, the feature map of a single filter $\boldsymbol{a}_i(\boldsymbol{x}) \in \mathbb{R}^{|G|}$ is equivariant with the regular representation of the group $G = Z_W \times Z_L$ (the representation that cyclically shifts the entries of $W \times L$ matrices). Here the order of the group is $|G| = WL$. Crucially, our theory shows that this representation contributes exactly one dimension to the fixed point subspace per filter (see Appendix A.6.1). Since the dimension of the entire collection of N feature maps $\boldsymbol{a}(x)$ is $N|G|$ for finite groups G, one might naively expect capacity to be $P \sim 2N|G|$ for large N. However, Theorem 2 shows that for G-invariant classification, only the fixed point subspace contribute to the classifier capacity. Since the dimension of the fixed point subspace present in the representation is equal to the number of filters N, we have $P \sim 2N$ (recall that $P = 2N$ is the inflection point of $f(P, N)$).

We show in Figure 1 that our prediction for $f(P, N)$ matches that empirically measured by training logistic regression linear classifiers on the representation. The convolutions in these networks are modified to have periodic boundary conditions while keeping the filters the same – see Appendix A.6.1 and Figure A.2 for more information and the result of using non-periodic convolutions, which impact the capacity but not the overall scaling with N_0.

4.1. Pooling Operations

In CNNs, local pooling is typically applied to the feature maps which result from each convolution layer. The impact of 2x2 maxpooling in CNNs is shown in Fig. 1, which reveals that maxpooling reduces capacity. In Appendix A.4 we prove that maxpooling reduces

capacity and bound this reduction. We also prove that average pooling leaves the capacity unchanged.

4.2. Induced representations

Induced representations are a fundamental ingredient in the construction of general equivariant neural network architectures (Cohen et al., 2019). Here we state our result, and relegate a formal definition of induced representations and the proof of the result to Appendix A.6.4.

Proposition 3 *Let π be a representation of a finite group induced from representation ρ. Then the fraction of separable dichotomies of π-manifolds is equal to that of the ρ-manifolds.*

5. Discussion and Conclusion

Equivariance has emerged as a powerful framework to build and understand representations that reflect the structure of the world in useful ways. In this work we take the natural step of quantifying the expressivity of these representations through the well-established formalism of perceptron capacity. We find that the number of "degrees of freedom" available for solving the classification task is the dimension of the space that is fixed by the group action. This has the immediate implication that capacity scales with the number of output channels in standard CNN layers, a fact we illustrate in simulations. However, our results are also very general, extending to virtually any equivariant representation of practical interest – in particular, they are immediately applicable to GCNNs.

Reproducibility Statement

Code for the simulations can be found at
`https://github.com/msf235/group-invariant-perceptron-capacity`.
This code includes an environment.yml file that can be used to create a python environment identical to the one used by the authors.

Acknowledgements

MF and CP are supported by the Harvard Data Science Initiative. MF thanks the Swartz Foundation for support. BB acknowledges the support of the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard (award #1764269) and the Harvard Q-Bio Initiative. ST was partially supported by the NSF under grant No. DMS-1439786.

References

Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data*, volume 4. AMLBook New York, NY, USA:, 2012.

Brandon Anderson, Truong Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *ICML*, 2016.

Taco Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *ArXiv*, abs/1801.10130, 2018.

Taco S Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant CNNs on homogeneous spaces. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

T. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electron. Comput.*, 14:326–334, 1965.

John Denker, W. Gardner, Hans Graf, Donnie Henderson, R. Howard, W. Hubbard, L. D. Jackel, Henry Baird, and Isabelle Guyon. Neural network recognizer for hand-written zip code digits. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1989. URL https://proceedings.neurips.cc/paper/1988/file/a97da629b098b75c294dffdc3e463904-Paper.pdf.

Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so(3) equivariant representations with spherical cnns. In *ECCV*, 2018.

Matthew Farrell, Blake Bordelon, Shubhendu Trivedi, and Cengiz Pehlevan. Capacity of group-invariant linear readouts from equivariant representations: How many objects can be linearly classified under all possible views? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=_4GFbtOuWq-.

William Fulton and Joe Harris. *Representation Theory: A First Course*. Readings in Mathematics. Springer-Verlag New York, New York, 2004. ISBN 9780387974958.

E. Gardner. Maximum storage capacity in neural networks. *EPL*, 4:481–485, 1987.

Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I. Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, August 2005. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature03721.

Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch-gordan nets: a fully fourier space spherical convolutional neural network. In *NeurIPS*, 2018.

Yann André LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551, 1989.

Richard L. Liboff. *Primer for Point and Space Groups*. Undergraduate Texts in Contemporary Physics. Springer, New York, 2004. ISBN 9780387402482.

liukuang. pytorch-cifar. https://github.com/kuangliu/pytorch-cifar, 2017.

George W. Mackey. *Induced Representations of Locally Compact Groups and Applications*, pages 132–166. Springer Berlin Heidelberg, Berlin, Heidelberg, 1970. ISBN 978-3-642-48272-4. doi: 10.1007/978-3-642-48272-4_6. URL https://doi.org/10.1007/978-3-642-48272-4_6.

Jean-Pierre Serre. *Linear Representations of Finite Groups*. Graduate Texts in Mathematics. Springer, New York, 2014. ISBN 978-1-4684-9460-0 978-1-4684-9458-7.

Mariya Shcherbina and Brunello Tirozzi. Rigorous Solution of the Gardner Problem. *Communications in Mathematical Physics*, 234(3):383–422, March 2003. ISSN 0010-3616, 1432-0916. doi: 10.1007/s00220-002-0783-3. URL http://link.springer.com/10.1007/s00220-002-0783-3.

Sameet Sreenivasan and Ila Fiete. Grid cells generate an analog error-correcting code for singularly precise neural computation. *Nature Neuroscience*, 14(10):1330–1337, October 2011. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn.2901.

Eugene Wigner. *Gruppentheorie Und Ihre Anwendung Auf Die Quantenmechanik Der Atomspektren*. Vieweg+Teubner Verlag, Wiesbaden, first edition, 1931. ISBN 978-3-663-00642-8.

Appendix A. Appendix

Appendix A.1 a glossary of definitions and notation.

Appendix A.2 Proofs of Lemma 1 and Theorem 2.

Appendix A.3 Additional example applications of Theorem 2.

Appendix A.4 Proofs of local pooling results.

Appendix A.5 a derivation of the irreps for the regular representation of the cyclic group Z_m.

Appendix A.6 a description of the construction of the GCNNs used in the paper, the methods for empirically measuring the fraction of linearly separable dichotomies, a description of local pooling, and complete formal proofs of the fraction of linearly separable dichotomies for these network representations (including with local pooling). This appendix also includes more description of the induced representation and a formal proof of the fraction of linearly separable dichotomies.

Appendix A.6.1 also contains an additional figure, Figure A.2.

A.1. Notation and Glossary

- $\boldsymbol{x}$: an abstract notation for an input object
- $\boldsymbol{r}(\boldsymbol{x})$: a feature map of the input to an N dimensional vector space.
- π: N dimensional linear representation of group G. For each $g \in G$, $\pi(g) \in GL(\mathbb{R}^N)$ is an $N \times N$ invertible real matrix.
- Equivariance property: $\boldsymbol{r}(g\boldsymbol{x}) = \pi(g)\boldsymbol{r}(\boldsymbol{x})$ for all $g \in G$ and all $\boldsymbol{x}$.
- Invariant measure: a measure $\mu : G \to \mathbb{R}_+$ on G with $\mu(gS) = \mu(S) = \mu(Sg)$. For finite groups, the uniform distribution. For locally compact topological groups, the Haar measure.
- $\langle \cdot \rangle_{g \in G}$: an average over the invariant measure of G
- Irreducible representation (irrep): an irreducible representation ρ on vector space V satisfies $\rho(g)v \in V$ for all $v \in V, g \in G$.
- Character $\chi(g)$: the trace of the representation $\chi(g) = \text{Tr } \pi(g)$.
- Fixed point subspace: the subspace V_0 for which $\pi(g)v \in V_0$ for all $v \in V_0$.
- General position: a collection of P points in general position in an N dimensional vector space have the property that any subset of $k \leq N$ points are linearly independent. These points are generic in the sense that they satisfy no more linear relationships than they must.
- Dichotomy: a particular binary labeling $\{y^\mu\}$ of P points $\{\boldsymbol{x}^\mu\}$.

- $f(P, N)$: fraction of linearly separable dichotomies given by Cover's function counting theorem (Cover, 1965).
- VC dimension: the largest possible integer P such that there exist P points where all possible dichotomies $\{y^\mu\}$ can be realized by the model Abu-Mostafa et al. (2012).
- Capacity: the largest possible ratio $\alpha_c = P/N$ where P points in general position can be linearly separated by a N dimensional perceptron with probability 1 in an asymptotic limit where $P, N \to \infty$ with $P/N = O_{N,P}(1)$. The classical result is $\alpha_c = 2$ (Gardner, 1987; Shcherbina and Tirozzi, 2003).
- $\mathcal{P}(\cdot)$: a local pooling operation.

A.2. Proofs of Lemma 1 and Theorem 2

Lemma 1 *A dataset $\{(\pi(g)\boldsymbol{r}^\mu, y^\mu)\}_{g\in G,\mu\in[P]}$ consisting of P π-manifolds with labels y^μ is linearly separable if and only if the dataset $\{(\langle\pi\rangle\boldsymbol{r}^\mu, y^\mu)\}_{\mu\in[P]}$ consisting of the P centroids $\langle\pi\rangle\boldsymbol{r}^\mu$ with the same labels is linearly separable. Formally,*

$$\exists \boldsymbol{w}\ \forall g \in G, \mu \in [P] : y^\mu \boldsymbol{w}^\top \pi(g)\boldsymbol{r}^\mu > 0 \iff \exists \boldsymbol{w}\ \forall \mu \in [P] : y^\mu \boldsymbol{w}^\top \langle\pi\rangle \boldsymbol{r}^\mu > 0.$$

Proof The forward implication is obvious: if there exists a $\boldsymbol{w}$ which linearly separates the P manifolds according to an assignment of labels y^μ, that same $\boldsymbol{w}$ must necessarily separate the centroids of these manifolds. This can be seen by averaging each of the quantities $y^\mu \boldsymbol{w}^\top \pi(g)\boldsymbol{r}^\mu$ over $g \in G$. Since each of these quantities is positive, the average must be positive.

For the reverse implication, suppose $y^\mu \boldsymbol{w}^\top \langle\pi\rangle \boldsymbol{r}^\mu > 0$, and define $\tilde{\boldsymbol{w}} = \langle\pi\rangle^\top \boldsymbol{w}$. We will show that $\tilde{\boldsymbol{w}}$ separates the P π-manfolds since

$$\begin{aligned}
y^\mu \tilde{\boldsymbol{w}}^\top \pi(g)\boldsymbol{r}^\mu &= y^\mu \boldsymbol{w}^\top \langle\pi\rangle\pi(g)\boldsymbol{r}^\mu \quad \text{(Definition of } \tilde{\boldsymbol{w}}\text{)} \\
&= y^\mu \boldsymbol{w}^\top \langle\pi(g')\pi(g)\rangle_{g'\in G}\boldsymbol{r}^\mu \quad \text{(Definition of } \langle\pi\rangle \text{ and linearity of } \pi(g)\text{)} \\
&= y^\mu \boldsymbol{w}^\top \langle\pi\rangle\boldsymbol{r}^\mu \quad \text{(Invariance of the Haar Measure } \mu(Sg) = \mu(S) \text{ for set } S\text{)} \\
&> 0 \quad \text{(Assumption that } \boldsymbol{w} \text{ separates centroids)}
\end{aligned}$$

Thus, all that is required to show that $\tilde{\boldsymbol{w}}$ separates the π-orbits are basic properties of a group representation and invariance of the Haar measure to G-transformations. ■

Theorem 2 *Suppose the points $\langle\pi\rangle\boldsymbol{r}^\mu$ for $\mu \in [P]$ lie in general position in the subspace $V_0 = \text{range}(\langle\pi\rangle) = \{\langle\pi\rangle\boldsymbol{x} : \boldsymbol{x} \in V\}$. Then V_0 is the fixed point subspace of π, and the fraction of linearly separable dichotomies on the P π-manifolds $\{\pi(g)\boldsymbol{r}^\mu : g \in G\}$ is $f(P, N_0)$, where $N_0 = \dim V_0$. Equivalently, N_0 is the number of trivial irreducible representations that appear in the decomposition of π into irreducible representations.*

Proof By the theorem of complete reducibility (see Fulton and Harris (2004)), π admits a decomposition into a direct sum of irreducible representations (**irreps**) $\pi \cong \pi_{k_1} \oplus \pi_{k_2} \oplus ... \oplus \pi_{k_M}$ acting on vector space $V = V_1 \oplus V_2 \oplus ... V_M$, where $\cong$ denotes equality up to similarity transformation (see Serre (2014) for a definition of irreps). The indices k_j indicate the

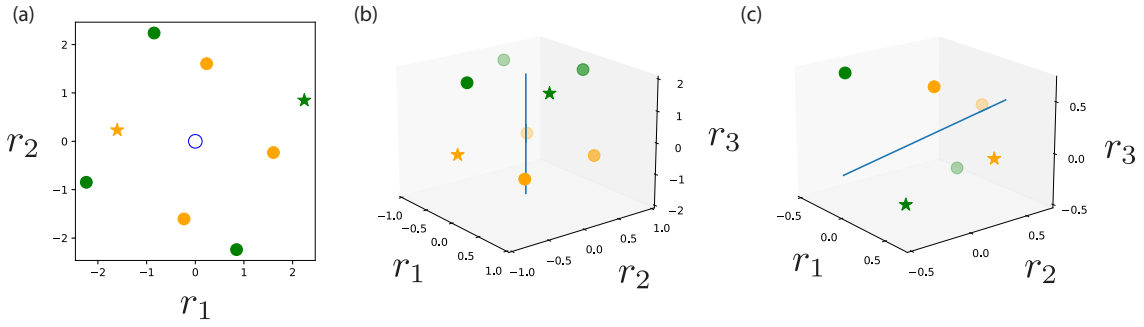

Figure A.1: π-manifolds for different π, illustrating that only the fixed point subspace contributes to capacity. In each panel two manifolds are plotted, with color denoting class label. Stars indicate the random points $\boldsymbol{r}^\mu$ for $\mu \in \{1, 2\}$ where the orbits begin, and closed circles denote the other points in the π-manifold. For (a) and (b) the group being represented is $G = Z_4$ and for (c) $G = Z_3$. (a) Here $\pi(g)$ is the 2×2 rotation matrix $\boldsymbol{R}(2\pi g/4)$. The open blue circle denotes the fixed point subspace $\{\boldsymbol{0}\}$. (b) Here $\pi(g)$ is the 3×3 block-diagonal matrix with the first 2×2 block being $\boldsymbol{R}(2\pi g/4)$ and second 1×1 block being 1. The blue line denotes the fixed point subspace $\text{span}\{(0, 0, 1)\}$. (c) Here $\pi(g)$ is the 3×3 matrix that cyclically shifts entries of length-3 vectors by g places. The blue line denotes the fixed point subspace $\text{span}\{(1, 1, 1)\}$.

type of irrep corresponding to invariant subspace V_j. The fixed point subspace V_0 is the direct sum of subspaces where trivial $k_j = 0$ irreps act: $V_0 = \bigoplus_{n:k_n=0} V_n$. By the Grand Orthogonality Theorem of irreps (see Liboff (2004)) all non-trivial irreps average to zero $\langle \pi_{k,ij}(g) \rangle_{g \in G} \propto \delta_{k,0}\delta_{i,j}$. Then, the matrix $\langle \pi \rangle$ simply projects the data to V_0. By Lemma 1 the fraction of separable dichotomies on the π-manifolds is the same as that of their centroids $\langle \pi \rangle \boldsymbol{r}^\mu$. Since, by assumption, the P points $\langle \pi \rangle \boldsymbol{r}^\mu$ lie in general position in V_0, the fraction of separable dichotomies is $f(P, \dim V_0)$ by Cover's Theorem. ■

A.3. Example Applications of Equivariant capacity

The 2×2 discrete rotation matrices $\boldsymbol{R}(\theta_g) \equiv \begin{bmatrix} \cos(\theta_g) & -\sin(\theta_g) \\ \sin(\theta_g) & \cos(\theta_g) \end{bmatrix}$ where $\theta_g = 2\pi g/m$ and $g \in Z_m$, are one possible representation of Z_m; in this case $V = \mathbb{R}^2$. This representation is irreducible and nontrivial, which implies that the dimension of the fixed point subspace is 0 (only the origin is mapped to itself by $\boldsymbol{R}$ for all g). Hence the fraction of linearly separable dichotomies of the π-manifolds by Theorem 2 is $f(P, 0)$. This result can be intuitively seen by plotting the orbits, as in Figure A.1a for $m = 4$. Here it is apparent that it is impossible to linearly separate two or more manifolds with different class labels, and that the nontrivial irrep $\boldsymbol{R}$ averages to the zero matrix.

The representation can be augmented by appending trivial irreps, defining $\pi : G \to \text{GL}(\mathbb{R}^N)$ by $\pi(g) = \boldsymbol{R}(\theta_g) \oplus \boldsymbol{I} \equiv \begin{bmatrix} \boldsymbol{R}(\theta_g) & 0 \\ 0 & \boldsymbol{I} \end{bmatrix}$ where $\boldsymbol{I}$ is an $(N-2) \times (N-2)$-dimensional

identity matrix. The number of trivial irreps is $N-2$, so that the capacity is $f(P, N-2)$. This is illustrated in Figure A.1b for the case $N=3$. Here we can also see that the trivial irrep, which acts on the subspace $\text{span}\{(0,0,1)\}$, is the only irrep in the decomposition of π that does not average to zero. This figure also makes intuitive the result of Lemma 1 that dichotomies are realizable on the π-manifolds if and only if the dichotomies are realizable on the centroids of the manifolds.

Suppose $\pi : Z_m \to \text{GL}(V)$ is the representation of Z_m consisting of the cyclic shift permutation matrices (this is called the **regular representation** of Z_m). In this case $V = \mathbb{R}^m$ and $\pi(g)$ is the matrix that cyclically shifts the entries of a length-m vector g places. For instance, if $m=3$ and $\boldsymbol{v} = (1,2,3)$ then $\pi(2)\boldsymbol{v} = (2,3,1)$.

In Appendix A.5 we derive the **irreducible representations** (irreps) of this representation, which consist of rotation matrices of different frequencies. There is one copy of the trivial irrep $\pi_0(g) \equiv 1$ corresponding with the fixed point subspace $\text{span}\{\mathbf{1}_m\}$ where $\mathbf{1}_m$ is the length-m all-ones vector. Hence the fraction of separable dichotomies is $f(P,1)$. This is illustrated in Figure A.1c in the case where $m=3$. The average of the regular representation matrix is $\langle \pi \rangle = \frac{1}{|G|}\mathbf{1}_m\mathbf{1}_m^\top$, indicating that $\langle \pi \rangle$ projects data along $\mathbf{1}_m$.

A.3.1. Direct sums of Regular Representations

For our last example we define a representation using the isomorphism $Z_m \cong Z_{m_1} \oplus Z_{m_2}$ for $m = m_1 m_2$ and m_1 and m_2 coprime[3]. Let $\pi^{(1)} : Z_{m_1} \to \text{GL}(\mathbb{R}^{m_1})$ and $\pi^{(2)} : Z_{m_2} \to \text{GL}(\mathbb{R}^{m_2})$ be the cyclic shift representations (i.e. the regular representations) of Z_{m_1} and Z_{m_2}, respectively. Consider the representation $\pi^{(1)} \oplus \pi^{(2)} : Z_m \to \text{GL}(\mathbb{R}^{m_1+m_2})$ defined by $(\pi^{(1)} \oplus \pi^{(2)})(g) \equiv \pi^{(1)}(g \mod m_1) \oplus \pi^{(2)}(g \mod m_2)$, the block-diagonal matrix with $\pi^{(1)}(g \mod m_1)$ being the first and $\pi^{(2)}(g \mod m_2)$ the second block.

There are two copies of the trivial representation in the decomposition of $\pi^{(1)} \oplus \pi^{(2)}$, corresponding to the one-dimensional subspaces $\text{span}\{(\mathbf{1}_{m_1}, \mathbf{0}_{m_2})\}$ and $\text{span}\{(\mathbf{0}_{m_1}, \mathbf{1}_{m_2})\}$, where $\mathbf{0}_{m_1}$ is the length-k vector of all zeros. Hence the fraction of separable dichotomies is $f(P,2)$. This reasoning extends simply to direct sums of arbitrary length ℓ, yielding a fraction of $f(P,\ell)$.[4] These representations are used to build a novel G-convolutional layer architecture with higher capacity than standard CNN layers in Section A.6.3.

These representations are analogous to certain formulations of grid cell representations as found in entorhinal cortex of rats (Hafting et al., 2005), which have desirable qualities in comparison to place cell representations (Sreenivasan and Fiete, 2011).[5] Precisely, a collection $\{Z_{m_k} \times Z_{m_k}\}_k$ of grid cell modules encodes a large 2-dimensional spatial domain $Z_m \times Z_m$ where $m = \prod_k m_k$.

A.3.2. SO(3): A Non-Abelian Lie Group

The special orthogonal group SO(3) on 3 dimensions (rotation group), the 3×3 orthogonal matrices with determinant $+1$, can also be analyzed within our theory. G-convolutional

3. Two numbers are coprime if they have no common prime factor.
4. Our results do not actually require that the m_k be coprime, but rather that none of the m_k divide one of the others. To see this, take $\tilde{m}$ and $\tilde{m}_k$, to be m and the m_k after being divided by all divisors common among them. Then $Z_{\tilde{m}} \cong \oplus_{k=1}^{\ell} Z_{\tilde{m}_k}$ and provided none of the $\tilde{m}_k$ are 1, one still gets a fraction of $f(P,\ell)$.
5. Place cells are analogous to standard convolutional layers.

neural networks that are equivariant to SO(3) rotations have become of high interest in the physical sciences and computer vision where the objects of interest often respect these symmetries (Anderson et al., 2019; Cohen et al., 2018; Esteves et al., 2018; Kondor et al., 2018) The irreducible representations have the form $\boldsymbol{B}_{k_m}$ where $\boldsymbol{B}_{k_m}$ are $(2k_m+1)\times(2k_m+1)$ block matrices, known as Wigner D-matrices (Wigner, 1931). The trivial irreps correspond to the one-dimensional irreps with $k = 0$. Thus, the SO(3)-invariant classification capacity merely counts the number of trivial irreps which have $N_0 = \sum_m \delta_{k_m,0}$. The capacity is again $f(P, N_0)$.

A.4. Pooling

In this section, we describe how our theory can be adapted for codes which contain such pooling operations. We will first assume that π is an N-dimensional representation of G. Let $\mathcal{P}(\boldsymbol{r}) : \mathbb{R}^N \to \mathbb{R}^{N/k}$ be a pooling operation which reduces the dimension of the feature map. The condition that a given dichotomy $\{y^\mu\}$ is linearly separable on a pooled code is

$$\exists \boldsymbol{w} \in \mathbb{R}^{N/k} \; \forall \mu \in [P], g \in G : y^\mu \boldsymbol{w}^\top \mathcal{P}(\pi(g)\boldsymbol{r}^\mu) > 0 \tag{2}$$

We will first analyze the capacity when $\mathcal{P}(\cdot)$ is a linear function (average pooling) before studying the more general case of local non-linear pooling on one and two-dimensional signals.

A.4.1. Local Average Pooling

In the case of average pooling, the pooling function $\mathcal{P}(\cdot)$ is a linear map, represented with matrix $\boldsymbol{P}$ which averages a collection of feature maps over local windows. Using an argument similar to Lemma 1 (Lemma 4 and Theorem 6 below), we prove that the capacity of a standard CNN is not changed by local average pooling: for a network with N filters, local average pooling preserves the one trivial dimension for each of the N filters. Consequently the fraction of separable dichotomies is $f(P, N)$.

A.4.2. Local Nonlinear Pooling

Often, nonlinear pooling operations are applied to downsample feature maps. For concreteness, we will focus on one-dimensional signals in this section and relegate the proofs for two-dimensional signals (images) to later in this section. Let $\boldsymbol{r}(\boldsymbol{x}) \in \mathbb{R}^{N\times D}$ represent a feature map with N filters and length-D signals. Consider a pooling operation $\mathcal{P}(\cdot)$ which maps the D pixels in each feature map into new vectors of size D/k for some integer k. Note that the pooled code is equivariant to the subgroup $H = \mathbb{Z}^{D/k}$, in the sense that $\mathcal{P}(\pi(h)\boldsymbol{r}) = \rho(h)\mathcal{P}(\boldsymbol{r})$ for any $h \in H$. The representation ρ is the regular representation of the subgroup H. We thus decompose G into cosets of size D/k: $g = jh$, where $j \in Z_k$ and $h \in Z_{D/k}$. The condition that a vector $\boldsymbol{w}$ separates the dataset is

$$\forall \mu \in [P], j \in Z_k, h \in H : y^\mu \boldsymbol{w}^\top \rho(h)\mathcal{P}\left(\pi(j)\boldsymbol{r}^\mu\right) > 0. \tag{3}$$

We see that there are effectively k points belonging to each of the P orbits in the pooled code. Since $\mathcal{P}(\cdot)$ is nonlinear, the averaging trick utilized in Lemma 1 is no longer available. However, we can obtain a lower bound on the capacity, $f(P, \lfloor N/k \rfloor)$, from a simple

extension of Cover's original proof technique as we show in Appendix A.6.2. Alternatively, an upper bound $f \leq f(P,N)$ persists since a $\boldsymbol{w}$ which satisfies Equation 3 must separate $\langle \rho(h) \rangle_{h\in H} \mathcal{P}(\boldsymbol{r}^\mu)$. This upper bound is tight when all k points $\langle \rho(h) \rangle_{h\in H} \mathcal{P}(\pi(j)\boldsymbol{r}^\mu)$ coincide for each μ, giving capacity $f(P,N)$. This is what occurs in the average pooling case where the upper bound $f \leq f(P,N)$ is tight. Further, if we are only interested in the H-invariant capacity problem, the fraction of separable dichotomies is $f(P,N)$, since ρ is a regular representation of H.

First we prove an extension of Lemma 1 to equivariant linear maps. This will be used to show that average pooling does not affect the capacity of the regular representation of Z_m.

Lemma 4 *Let π be a representation of the group G and suppose the matrix $\boldsymbol{P}$ is equivariant with respect to the restriction of π to a subgroup $H \subseteq G$, so that for all $h \in H$ $\boldsymbol{P}\pi(h) = \rho(h)\boldsymbol{P}$ for some representation ρ of H. Let R denote a set of representatives of G/H. Then we have the following equivalence.*

$$\begin{aligned}
&\exists \boldsymbol{w}\ \forall \mu \in [P], g \in G : y^\mu \boldsymbol{w}^\top \boldsymbol{P}\pi(g)\boldsymbol{r}^\mu > 0 \\
&\quad \iff \exists \boldsymbol{w}\ \forall \mu \in [P]\ \forall g' \in R\ :\ y^\mu \boldsymbol{w}^\top \boldsymbol{P}\langle \pi(h) \rangle_{h\in H} \pi(g')\boldsymbol{r}^\mu > 0.
\end{aligned}$$

Proof For the forward implication, we write the coset decomposition $g = hg'$ of g and average over H to find

$$\begin{aligned}
\forall g \in G : y^\mu \boldsymbol{w}^\top \boldsymbol{P}\pi(g)\boldsymbol{r}^\mu > 0 &\iff \forall h \in H, g' \in R : y^\mu \boldsymbol{w}^\top \boldsymbol{P}\pi(h)\pi(g')\boldsymbol{r}^\mu > 0 \\
&\implies \forall g' \in R : y^\mu \boldsymbol{w}^\top \boldsymbol{P}\langle \pi(h) \rangle_{h\in H} \pi(g')\boldsymbol{r}^\mu > 0.
\end{aligned}$$

For the backward implication, suppose $y^\mu \boldsymbol{w}^\top \boldsymbol{P}\langle \pi(h) \rangle_{h\in H} \pi(g')\boldsymbol{r}^\mu > 0$ for all representatives $g' \in R$, and define $\tilde{\boldsymbol{w}} = \langle \rho(h) \rangle_{h\in H}^\top \boldsymbol{w}$. For any $g \in G$, take a coset decomposition $g = hg'$ for $h \in H$ and $g' \in R$. We then have

$$\begin{aligned}
y^\mu \tilde{\boldsymbol{w}}^\top \boldsymbol{P}\pi(g)\boldsymbol{r}^\mu &= y^\mu \tilde{\boldsymbol{w}}^\top \boldsymbol{P}\pi(h)\pi(g')\boldsymbol{r}^\mu \ \text{(Coset decomposition)} \\
&= y^\mu \tilde{\boldsymbol{w}}^\top \rho(h)\boldsymbol{P}\pi(g')\boldsymbol{r}^\mu \ (\boldsymbol{P} \text{ is } H\text{-equivariant}) \\
&= y^\mu \boldsymbol{w}^\top \left\langle \rho(h') \right\rangle_{h'\in H} \rho(h)\boldsymbol{P}\pi(g')\boldsymbol{r}^\mu \ \text{(Definition of } \tilde{\boldsymbol{w}}) \\
&= y^\mu \boldsymbol{w}^\top \langle \rho(h) \rangle_{h\in H} \boldsymbol{P}\pi(g')\boldsymbol{r}^\mu \ \text{(Invariance of measure)} \\
&= y^\mu \boldsymbol{w}^\top \boldsymbol{P} \langle \pi(h) \rangle_{h\in H} \pi(g')\boldsymbol{r}^\mu \ (\boldsymbol{P} \text{ is linear and } H\text{-equivariant}) \\
&> 0 \ \text{(By assumption).}
\end{aligned} \tag{4}$$

The implication follows. ■

Lemma 5 *For the regular representation of $G = Z_D$, a local average pooling over windows of size k generates a matrix $\boldsymbol{P}$ which is equivariant with respect to the subgroup $H = Z_{D/k}$ with the property that*

$$\boldsymbol{P}\langle \pi(h) \rangle_{h\in H} = a\boldsymbol{P}\langle \pi(g) \rangle_{g\in G} \tag{5}$$

where $a > 0$ is a positive constant.

Proof First, note that the new pooled code is the regular representation of H since shifts of size mk in the original feature map corresponds to shifts of length m in the pooled code. Thus $\boldsymbol{P}$ is equivariant to $H = Z_{D/k}$. Next we note the following two facts

$$\boldsymbol{P} \langle \pi(h) \rangle_{h \in H} = a' \mathbf{1}_{D/k} \mathbf{1}_D^\top \tag{6}$$

$$\boldsymbol{P} \langle \pi(h) \rangle_{g \in G} = a'' \mathbf{1}_{D/k} \mathbf{1}_D^\top \tag{7}$$

a' and a'' are positive constants and $\mathbf{1}_D$ and $\mathbf{1}_{D/k}$ are the D and D/k dimensional vector of all ones, respectively. Thus, we have that $\boldsymbol{P}\langle \pi(h) \rangle_{h \in H} = a \boldsymbol{P} \langle \pi(g) \rangle_{g \in G}$ where a is a positive constant. ■

Theorem 6 *The fraction of linearly separable dichotomies of a CNN pooling layer with N filters after average pooling from feature maps with the size of the input image $W \times L$ to pooled feature maps of size $W/k \times L/k$ is $f(P, N)$, i.e. no capacity is lost due to local average pooling.*

Proof The CNN layer before pooling is a regular representation π of the full group $G = Z_W \times Z_L$, applied to each of the M input channels via a direct a sum $\bigoplus_{j=1}^M \pi$. The layer after pooling is a regular representation ρ of the subgroup $H = Z_{W/k} \times Z_{L/k}$, also applied to the output channels via a direct sum $\bigoplus_{j=1}^N \rho$. Let R be a set of representatives of G/H. Since $\boldsymbol{P}$ is equivariant to π and ρ over H we have by the previous two lemmas that

$$\begin{aligned}
\forall g \in G : y^\mu \boldsymbol{w}^\top \boldsymbol{P} \pi(g) \boldsymbol{r}^\mu > 0 &\iff \forall g' \in R : y^\mu \boldsymbol{w}^\top \boldsymbol{P} \langle \pi(h) \rangle_{h \in H} \pi(g') \boldsymbol{r}^\mu > 0 \\
&\iff \forall g' \in R : y^\mu \boldsymbol{w}^\top \boldsymbol{P} \langle \pi(g) \rangle_{g \in G} \pi(g') \boldsymbol{r}^\mu > 0 \\
&\iff y^\mu \boldsymbol{w}^\top \boldsymbol{P} \langle \pi(g) \rangle_{g \in G} \boldsymbol{r}^\mu > 0 \\
&\iff y^\mu \boldsymbol{w}^\top \boldsymbol{P} \langle \pi(h) \rangle_{h \in H} \boldsymbol{r}^\mu > 0 \\
&\iff y^\mu \boldsymbol{w}^\top \langle \rho(h) \rangle_{h \in H} \boldsymbol{P} \boldsymbol{r}^\mu > 0
\end{aligned}$$

Thus the capacity is determined by the rank of $\langle \rho(h) \rangle_{h \in H}$, assuming the $\langle \rho(h) \rangle_{h \in H} \boldsymbol{P} \boldsymbol{r}^\mu$ are in general position. Since each ρ is a copy of the regular representation for H, the rank of $\langle \bigoplus_{j=1}^N \rho(h) \rangle_{h \in H}$ is merely N. Thus the fraction of linearly separable dichotomies is $f(P, N)$, the same as the capacity before pooling. ■

Now we prove local pooling operations in a standard CNN preserve a regular representation of a subgroup of the cyclic group.

Lemma 7 *Suppose P is a local pooling operation on two-dimensional signals (CNN feature maps), and that π is the regular representation of a group $G = Z_W \times Z_L$ on code $\boldsymbol{a}(\boldsymbol{x})$. A pooled feature map $\boldsymbol{r} = \mathcal{P}(\boldsymbol{a})$ which acts on $k \times k$ windows of $\boldsymbol{a}$ is a regular representation of the subgroup $H = Z_{W/k} \times Z_{L/k}$.*

Proof Suppose an equivariant feature map $\boldsymbol{a}(\boldsymbol{x}) \in \mathbb{R}^{W \times L \times N}$ has corresponding regular representation of the group $G = Z_W \times Z_L$ for each of the N filters. Consider any local

pooling operation $\mathcal{P}(\cdot)$ (such as average or maximum) which acts on $k \times k$ patches where k divides both W and L.

$$r_{ij,h}(x) = \mathcal{P}\left(\{a_{i',j',h}(x) \mid i' \in [ik, (i+1)k], j' \in [jk, (j+1)k]\}\right) \tag{8}$$

Note that for $k > 1$, $\boldsymbol{r}(x)$ is no longer equivariant to G since the representation does not satisfy the homomorphism property for shifts with length ℓ not divisible by k. However, the new code is equivariant to a subgroup $H = Z_{W/k} \times Z_{L/k}$, namely vertical and horizontal shifts with length divisible by k. Let $\boldsymbol{\pi}^{\boldsymbol{x}}_{nk,mk}$ represent a vertical shift of the image $\boldsymbol{x}$ by nk pixels and horizontal shift by mk pixels. Note that $a_{ij,h}(\boldsymbol{\pi}^{\boldsymbol{x}}_{nk,mk}\boldsymbol{x}) = a_{i+nk,j+mk}(\boldsymbol{x})$ since $\boldsymbol{a}(\boldsymbol{x})$ is equivariant. Then, the h-th pooled feature map transforms as

$$\begin{aligned}
r_{ij,h}(\boldsymbol{\pi}^{\boldsymbol{x}}_{nk,mk}\boldsymbol{x}) &= \mathcal{P}\left(\{a_{i',j',h}(\boldsymbol{\pi}^{\boldsymbol{x}}_{nk,mk}x) \mid i' \in [ik, (i+1)k], j' \in [jk, (j+1)k]\}\right) \\
&= \mathcal{P}\left(\{a_{i'+nk,j'+mk,h}(\boldsymbol{x}) \mid i' \in [ik, (i+1)k], j' \in [jk, (j+1)k]\}\right) \\
&\qquad (\boldsymbol{a} \text{ is Equivariant to } \pi^x) \\
&= \mathcal{P}\left(\{a_{i',j',h}(\boldsymbol{x}) \mid i' \in [(n+i)k, (n+i+1)k], j' \in [(m+j)k, (m+j+1)k]\}\right), \\
&\qquad (k \text{ divides } W, H) \\
&= r_{i+n,j+m,h}(\boldsymbol{x}) \ , \ (\text{Definition of } \boldsymbol{r})
\end{aligned}$$

We thus find that the code is a regular representation of the subgroup of the cyclic translations $H = \{(nk, mk)\}_{n\in[H/k],m\in[W/k]}$. This new group G' has dimension $|H| = \frac{1}{k^2}|G|$. ■

A.5. Irreps for the cyclic group

Here we compute the irreducible representations (irreps) of representations $\pi : Z_m \to \mathrm{GL}(V)$ of the cyclic group Z_m over a real vector space V. See Serre (2014) for a definition of irrep, and for a derivation of the irreps when $\boldsymbol{V}$ is a complex vector space. The irreps can be used to determine the dimensionality of V_0, as shown in the proof of Theorem 2 (Appendix A.2). To find the irreps, one can use the form for the eigenvalues and eigenvectors for circulant matrices, since all the $\pi(g)$ are circulant. This results in the simultaneous diagonalization

$$\pi(g) = \boldsymbol{V}\left(1 \oplus \boldsymbol{R}(2\pi g/m) \oplus \boldsymbol{R}(4\pi g/m) \oplus \cdots \oplus \boldsymbol{R}((m-1)\pi g/m)\right)\boldsymbol{V}^\top$$

where $\boldsymbol{V}$ is the real-valued version of the discrete Fourier transform matrix (the columns are proportional to cosines and sines of varying frequencies, along with a column proportional to $\mathbf{1}_m$).

Note that the $2x2$ rotation matrices $\boldsymbol{R}(2\pi gk/m)$ are irreps for $k \neq m/2$ and $k \neq 0$, since there is no one-dimensional subspace of $\mathbb{R}^2$ that is invariant to $\boldsymbol{R}(2\pi gk/m)$ for all g. The exception for $k = m/2$ if m is even, gives $\boldsymbol{R}(2\pi gk/m) = (-1)^g\boldsymbol{I}$, which corresponds to rotation of 180 degrees. The subspace $\mathrm{span}\{(1, 0)\}$ is invariant to this action, so that $\boldsymbol{R}(2\pi gk/m)$ is not an irrep. This representation can thus be reduced to an action on a one-dimensional subspace, represented by $(-1)^g$. The case $k = 0$ gives the trivial representation.

A.6. G-equivariant convolutional layers

A.6.1. Standard convolutional layers

A convolutional layer consists of a set of N $k \times k'$ filters F_i that are convolved (technically cross-correlated) with a stack of M $W \times L$ input tensors. Here M is the number of input channels and N the number of output channels. The convolution runs each filter (i.e. takes the dot product at all possible positions) over each of the $W \times L$ input tensors, and the result is averaged across the M input channels to produce the output of one output channel. In the positions where the filters approach the edges of the input tensor, different choices can be made about how to handle these edge conditions. The standard choice is to pad the edges with some number of zeros depending on the desired shape of the output tensor and run the convolution out to the end of the padded image. Another possible choice is to loop the edges of the input tensor together, so that the filter is applied to the other side of the input tensor as it runs off the edge. This periodic boundary condition allows us to write the convolution formally in terms of group actions, and to apply our theory directly. When convolutions are not periodic, the resulting capacity increases somewhat but still follows the P/N_0 scaling of the periodic convolutions (Figure A.2).

For the random convolutional layers of Figure 1a, the input tensors are size 10×10 and the number of input channels are $M = 3$, as for standard color images. Each entry of these tensors is normally distributed with mean 0. The filters are also of size 10×10 with periodic boundary conditions, and are initialized according to a normal Xavier distribution with parameters that are the default for Pytorch 1.9. The convolution is implemented via the Pytorch 1.9 implementation of Conv2d with padding_mode="circular" and padding=0 in the case of periodic boundary conditions. The bias term of the convolution is set to zero and the convolution is followed by a ReLU nonlinearity (blue line). The resulting figures do not change appreciably for different choices of input tensor size, number of input channels, or size of filters (though note that the nonlinearity is essential for satisfying the general position condition of Theorem 2; otherwise, the capacity would be determined by the number of input channels rather than number of output channels). The output of this convolution is then fed through a 2×2 max pooling layer (orange line in Figure 1a), provided by Pytorch 1.9's MaxPool2d.

The pretrained VGG-11 layers used in Figure 1b and Figure A.2 are taken from liukuang (2017). The first convolutional block (blue line) consists of 3×3 pretrained filters applied to CIFAR-10 image tensors randomly selected from the validation set and normalized in the same way they are normalized during training (see liukuang (2017) for details). These images are of size 32×32 and have $M = 3$ input channels, followed by a batch normalization layer in evaluation mode (fixed parameters), and then followed by a ReLU nonlinearity. The boundary conditions of these convolutions are set to be periodic in Figure 1b and nonperiodic with a zero padding sizes of 1 in Figure A.2, and the bias term is set to zero. The batch normalization is an element-wise operation and so equivariant to the representations we consider – thus this operation is not expected and is not observed to affect the perceptron capacity. This convolutional block is then followed by a 2×2 max pooling layer (orange line). Finally, another convolutional block of 3×3 filters, batch normalization, and ReLU nonlinearity are applied (green line).

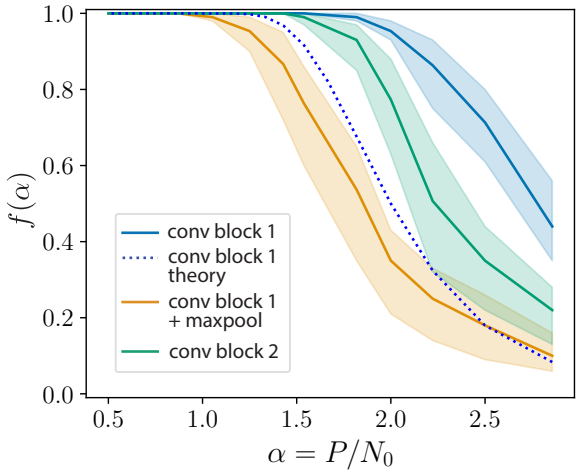

Figure A.2: The fraction of realizable dichotomies of non-periodic convolutional layers is higher than periodic convolutional layers, but still obeys the same scaling. Details are exactly as in Figure 1b, but using non-periodic convolutions with a zero padding of size 1. Here the theory line refers to the theory for periodic convolutions.

The fraction of linearly separable dichotomies is measured empirically by using the scikit-learn LogisticRegression implementation of logistic regression, with a tolerance value of tol=1e-18 and an inverse regularization value of C=1e8. The maximum number of iterations is set to 500. An intercept (i.e. bias) is not used for this fit.

To formally prove that the fraction of separable dichotomies is $f(P, N)$ for standard periodic convolutional layers, first note that the convolution is equivariant with respect to cyclic permutation of the inputs and of the outputs. The representation for cyclically permutating the output tensor can be written $\bigoplus_{k=1}^{N} \pi$ where π is the representation that cyclically permutes the entries of $W \times L$ matrices. Since each copy of π contains one trivial irrep in its decomposition into a direct sum of irreps, the direct sum $\bigoplus_{k=1}^{N} \pi$ contains N trivial irreps in its decomposition. The final step to use Theorem 2 is to argue that the centroids of the manifolds are in general position. Since a nonlinearity (ReLU) is applied to the output of the convolution, and since there is no particular structure in the convolutional filters beyond possibly sparsity, we can generically expect this to be the case.

A.6.2. Lower Bound and Upper Bound on Capacity for Nonlinear Pooling

Theorem 8 *Suppose a code which is equivariant to a finite group G is pooled to a new code which is equivariant to a finite subgroup $H \subseteq G$. Suppose the number of trivial dimensions in the original G-equivariant code is N_0. Then the fraction of linearly separable dichotomies on the G-invariant problem for the pooled code is at least $f(P, \lfloor N_0/k \rfloor)$ where $k = |G/H|$. Similarly the fraction is at most $f(P, N_0)$.*

Proof The pooled code, by assumption, has the property $\mathcal{P}(\pi(hg')\boldsymbol{r}) = \rho(h)\mathcal{P}(\pi(g')\boldsymbol{r})$ for any $h \in H$ and $g' \in R$, where R is a set of representatives of G/H. The G-invariant separability condition amounts to the proposition

$$\exists \boldsymbol{w} \forall \mu \in [P] \ \forall h \in H, g' \in R : y^\mu \boldsymbol{w}^\top \rho(h) \mathcal{P}(\pi(g')\boldsymbol{r}^\mu) > 0 \tag{9}$$

$$\iff \exists \boldsymbol{w} \forall \mu \in [P] \ \forall g' \in R : y^\mu \boldsymbol{w}^\top \langle \rho(h) \rangle_{h \in H} \mathcal{P}(\pi(g')\boldsymbol{r}^\mu) > 0; \tag{10}$$

in other words, a solution on the right hand side affords a solution over all of the manifolds generated in the input space. We see that, this requires considering if this particular dichotomy is linearly separble on the Pk anchor points $\langle \rho(h) \rangle_{h \in H} \mathcal{P}(\pi(g')\boldsymbol{r}^\mu)$. The simplest strategy to obtain an upper bound is to consider what happens when a single new manifold is added. We see that when a single new base point $\boldsymbol{r}$ is added it corresponds to k new points in the orbit $\langle \rho \rangle \mathcal{P}(\pi(g')\boldsymbol{r})$ for all $g' \in R$. Suppose that $\langle \rho(h) \rangle_{h \in H}$ has rank N_0. Let $C(P, N_0)$ represent the number of linearly separable dichotomies for P G-orbits in N_0 trivial dimensions. Upon the addition of the k new points ($P \to P + 1$), we find that some of the pre-existing separable dichotomies give a new separable dichotomy. This can be guaranteed to occur when a $\boldsymbol{w}$ separates the old dichotomy and has $\boldsymbol{w}^\top \langle \rho \rangle \cdot \mathcal{P}(\pi(g')\boldsymbol{r}) = 0$ for the new anchor point $\boldsymbol{r}$ (but this condition is not *necessary* for a new dichotomy to be separable). This condition means that the original dichotomy is separable in the $N_0 - k$ dimensional subspace $\{\boldsymbol{w} : \boldsymbol{w} \cdot \mathcal{P}(\pi(g')\boldsymbol{r}) = 0\}$. By making infinitesimal adjustment to this $\boldsymbol{w}$ the correct label on this new orbit can be achieved without altering the labels on any other dichotomy. Since this argument gives a sufficient but not necessary condition to generate a new dichotomy, we obtain the following inequality

$$C(P + 1, N_0) \geq C(P, N_0) + C(P, N_0 - k). \tag{11}$$

Solving this recursion gives the capacity $f_{Pooled}(P, N_0) \geq f_{Cover}(P, \lfloor N_0/k \rfloor)$. The greatest capacity occurs in the special case where $\langle \rho \rangle \mathcal{P}(\pi(g')\boldsymbol{r}) = \langle \rho \rangle \mathcal{P}(\boldsymbol{r})$. In this case, the usual counting theorem applies giving a fraction of separable dichotomies of $f(P, N_0)$. This is achieved, for instance in average pooling as we showed in Theorem 6. ■

A.6.3. Direct sum equivariant convolutional layers

Here we describe how to build a convolutional layer architecture that is equivariant with respect to the regular representation in the input space and the direct sum representations introduced in Section A.3.1 in the output space. For the following we assume that m_1 and m_2 are coprime, though see the footnote in Section A.3.1 for a discussion of how to loosen this requirement.

The input data is a $W \times L \times M$ tensor where M is the number of input channels. The first step is to simply take the output of a standard convolution (in our simulations we also apply a ReLU nonlinearity) applied to this input with periodic boundary conditions, resulting in a $W \times L \times N$ tensor where N is the number of output channels. The next step is to, for each of the output channels, take an average (or maximum) between entries spaced m_1 entries apart horizontally or vertically in the matrix, resulting in an $m_1 \times m_1$ matrix. In our simulations we took averages rather than maximums. This is repeated for the other number m_2, resulting in an $m_2 \times m_2$ matrix. This is repeated for every output

channel, resulting in N matrices of size $m_1 \times m_1$ and N matrices of size $m_2 \times m_2$. Finally, the resulting matrices are flattened and appended into an $(m_1^2 + m_2^2) \times N$ matrix, and the result is passed through a nonlinearity (ReLU).

As the input tensor is cyclically permuted according to a regular representation π of $Z_{m_1 m_2}$, the output of this equivariant convolutional layer permutes according to the representation $\pi^{(1)} \oplus \pi^{(2)}$ where $\pi^{(1)}$ is the regular representation of Z_{m_1} and $\pi^{(2)}$ is the regular representation of Z_{m_2}.

The proof that the fraction of separable dichotomies is given by $f(P, 2N)$ follows the same proof as for the standard periodic convolutions in Appendix A.6.1. Instead of a direct sum $\bigoplus_{k=1}^{N} \pi$ we get a direct sum $\bigoplus_{k=1}^{N} (\pi^{(1)} \oplus \pi^{(2)})$. Each of the $\pi^{(1)} \oplus \pi^{(2)}$ contain two trivial irreps in its decomposition, so that the final fraction is $f(P, 2N)$.

A.6.4. Induced representations

First we state the definition of an induced representation. Let H be a subgroup of a finite group G and let $\rho : H \to \mathrm{GL}(W)$ be a representation of H. Let V be the vector space of functions $f : G \to W$ such that $f(gh) = \rho(h) f(g)$ for all $h \in H$ and $g \in G$. We now define the induced representation $\pi : G \to \mathrm{GL}(V)$ to be the representation which satisfies $(\pi(g')f)(g) = f(gg')$.

For intuition, note that every element of g can be written $g = rh$ where r is a representative for a coset in G/H and $h \in H$. This is because the cosets G/H partition G and the action of H stays within a coset; hence r selects out the coset, and h goes to the desired element of the coset: $g = rh$. With this decomposition, the action of π is then $\pi(rh)f(g) = \rho(h)f(gr)$. Hence the r component under π has the effect of permuting to the new coset that gr belongs in, and the h component under π then has the effect $\rho(h)$ on the resulting vector $f(gr)$ that we originally specified. This is the most natural way to get a representation of G from a representation of H. In the case of finite groups, one can think of the r component as permuting a set of isomorphic copies of V, each copy corresponding to a different coset.

To compute the capacity of induced representations, and so prove Proposition 3, we use Frobenius reciprocity of characters. Recall that the character θ of a representation $\pi : G \to \mathrm{GL}(V)$ is the map $\theta : G \to \mathbb{C}$ induced by the trace: $\theta(g) = \mathrm{Tr}(\pi(g))$. Now let θ be the character of $\rho : H \to \mathrm{GL}(V)$ and let θ^G be the character of the induced representation. Then $\langle \theta^G(g) \rangle_{g \in G} = \langle \theta(h) \rangle_{h \in H}$ by Frobenius reciprocity of characters (Mackey, 1970). The average of the character is the number of trivial representations contained in the decomposition of the representation (see Serre (2014)). Hence the capacity of the induced representation is equal to the capacity of ρ. The existence of extensions beyond finite groups is not clear to the authors, but we welcome information if such exists.