References

[1] Yaser S. Abu-Mostafa. Hints and the VC Dimension. *Neural Computation*, 5(2):278–288, 03 1993.

[2] 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.

[3] Fabio Anselmi, Joel Z. Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso A. Poggio. Unsupervised learning of invariant representations in hierarchical architectures. *CoRR*, abs/1311.4158, 2013.

[4] Fabio Anselmi, Lorenzo Rosasco, and Tomaso Poggio. On invariance and selectivity in representation learning. *Information and Inference: A Journal of the IMA*, 5(2):134–158, 05 2016.

[5] Minkyung Baek, Frank Dimaio, Ivan V. Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy M. DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina Aike van Dijk, Ana C Ebrecht, Diederik Johannes Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373:871 – 876, 2021.

[6] Peter Bartlett. Covering numbers, chaining & dudley's entropy integral (lecture 24). class notes on statistical learning theory, 2006.

[7] Arash Behboodi, Gabriele Cesa, and Taco S Cohen. A pac-bayesian generalization bound for equivariant networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 5654–5668. Curran Associates, Inc., 2022.

[8] Erik J. Bekkers. B-spline cnns on lie groups. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[9] Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning under geometric stability. *Advances in neural information processing systems*, 34:18673–18684, 2021.

[10] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*, pages 169–207, 2004.

[11] Anadi Chaman and Ivan Dokmanic. Truly shift-invariant convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3773–3783, 2021.

[12] Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation. *The Journal of Machine Learning Research*, 21(1):9885–9955, 2020.

[13] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *ICML*, 2019.

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

[15] Taco S. Cohen. Learning transformation groups and their invariants. Master's thesis, University of Amsterdam, 2013.

[16] 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.

[17] Stephan Eismann, Raphael J. L. Townshend, Nathaniel Thomas, Milind Jagota, Bowen Jing, and Ron O. Dror. Hierarchical, rotation-equivariant neural networks to select structural models of protein complexes. *Proteins: Structure*, 89:493 – 501, 2020.

[18] Bryn Elesedy. Provably strict generalisation benefit for invariance in kernel methods. *Advances in Neural Information Processing Systems*, 34:17273–17283, 2021.

[19] Bryn Elesedy. Group symmetry in PAC learning. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022.

[20] Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 2959–2969. PMLR, 2021.

[21] Marc Finzi, Gregory Benton, and Andrew G Wilson. Residual pathway priors for soft equivariance constraints. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 30037–30049. Curran Associates, Inc., 2021.

[22] Marc Finzi, Max Welling, and Andrew Gordon Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. *ArXiv*, abs/2104.09459, 2021.

[23] Jonathan Gordon, David Lopez-Paz, Marco Baroni, and Diane Bouchacourt. Permutation equivariant models for compositional generalization in language. In *ICLR*, 2020.

[24] B. Haasdonk, A. Vossen, and H. Burkhardt. Invariance in kernel methods by haar-integration kernels. In Heikki Kalviainen, Jussi Parkkinen, and Arto Kaarna, editors, *Image Analysis*, pages 841–851, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[25] Osman Semih Kayhan and Jan C van Gemert. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14274–14285, 2020.

[26] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. *ArXiv*, abs/2003.03123, 2020.

[27] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *ICML*, 2018.

[28] Aryeh Kontorovich and Roi Weiss. Maximum margin multiclass nearest neighbors. In *International conference on machine learning*, pages 892–900. PMLR, 2014.

[29] Robert Krauthgamer and James R Lee. Navigating nets: Simple algorithms for proximity search. In *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 798–807. Citeseer, 2004.

[30] 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.

[31] Clare Lyle, Mark van der Wilk, Marta Z. Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. *ArXiv*, abs/2005.00178, 2020.

[32] Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. *ArXiv*, abs/1812.09902, 2019.

[33] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3351–3418. PMLR, 15–19 Aug 2021.

[34] Marvin Misky and Seymour Papert. *Perceptrons, Expanded Edition: An Introduction to Computational Geometry*. The MIT Press, Cambridge, MA, 1987.

[35] Behnam Neyshabur, Srinadh Bhojanapalli, David A. McAllester, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *ArXiv*, abs/1707.09564, 2017.

[36] Siamak Ravanbakhsh, Jeff G. Schneider, and Barnabás Póczos. Equivariance through parameter-sharing. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2892–2901. PMLR, 2017.

[37] Marco Reisert and Hans Burkhardt. Learning equivariant functions with matrix valued kernels. *Journal of Machine Learning Research*, 8(15):385–408, 2007.

[38] David W Romero and Suhas Lohit. Learning partial equivariances from data. *Advances in Neural Information Processing Systems*, 35:36466–36478, 2022.

[39] Akiyoshi Sannai, Masaaki Imaizumi, and Makoto Kawano. Improved generalization bounds of group invariant / equivariant deep networks via quotient feature spaces. In Cassio P. de Campos, Marloes H. Maathuis, and Erik Quaeghebeur, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, volume 161 of *Proceedings of Machine Learning Research*, pages 771–780. AUAI Press, 2021.

[40] Victor Garcia Satorras, E. Hoogeboom, F. Fuchs, I. Posner, and M. Welling. E(n) equivariant normalizing flows for molecule generation in 3d. *ArXiv*, abs/2105.09016, 2021.

[41] Bernhard Schölkopf, Chris Burges, and Vladimir Vapnik. Incorporating invariances in support vector learning machines. In Christoph von der Malsburg, Werner von Seelen, Jan C. Vorbrüggen, and Bernhard Sendhoff, editors, *Artificial Neural Networks — ICANN 96*, pages 47–52, Berlin, Heidelberg, 1996. Springer Berlin Heidelberg.

[42] H. Schulz-Mirbach. On the existence of complete invariant feature spaces in pattern recognition. In *Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol.II. Conference B: Pattern Recognition Methodology and Systems*, pages 178–182, 1992.

[43] Han Shao, Omar Montasser, and Avrim Blum. A theory of PAC learnability under transformation invariances. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[44] John. Shawe-Taylor. Building symmetries into feedforward networks. In *1989 First IEE International Conference on Artificial Neural Networks, (Conf. Publ. No. 313)*, pages 158–162, 1989.

[45] John Shawe-Taylor. Threshold network learning in the presence of equivalences. In *NIPS*, 1991.

[46] John Shawe-Taylor. Symmetries and discriminability in feedforward network architectures. *IEEE Transactions on Neural Networks*, 4(5):816–826, 1993.

[47] John Shawe-Taylor. Introducing invariance: a principled approach to weight sharing. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 345–349 vol.1, 1994.

[48] John Shawe-Taylor. Sample sizes for threshold networks with equivalences. *Information and Computation*, 118(1):65–72, 1995.

[49] Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel Rodrigues. Generalization error of invariant classifiers. In *Artificial Intelligence and Statistics*, pages 1094–1103. PMLR, 2017.

[50] I. Sosnovik, A. Moskalev, and A. Smeulders. Scale equivariance improves siamese tracking. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2764–2773, 2021.

[51] Behrooz Tahmasebi and Stefanie Jegelka. The exact sample complexity gain from invariances for kernel regression on manifolds, 2023.

[52] Vladimir M. Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. *Selected Works of AN Kolmogorov: Volume III: Information Theory and the Theory of Algorithms*, pages 86–170, 1993.

[53] Raphael J. L. Townshend, Stephan Eismann, Andrew M. Watkins, Ramya Rangan, Maria Karelina, Rhiju Das, and Ron O. Dror. Geometric deep learning of rna structure. *Science*, 373:1047 – 1051, 2021.

[54] Tycho F. A. van der Ouderaa, David W. Romero, and Mark van der Wilk. Relaxing equivariance constraints with non-stationary continuous filters. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, Dec 2022.

[55] Andrea Vedaldi, Matthew Blaschko, and Andrew Zisserman. Learning equivariant structured output svm regressors. In *2011 International Conference on Computer Vision*, pages 959–966, 2011.

[56] Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with lipschitz functions. *J. Mach. Learn. Res.*, 5(Jun):669–695, 2004.

[57] Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *International Conference on Machine Learning*, pages 23078–23091. PMLR, 2022.

[58] Maurice Weiler, Patrick Forré, Erik P. Verlinde, and Max Welling. Coordinate independent convolutional networks - isometry and gauge equivariant convolutions on riemannian manifolds. *ArXiv*, abs/2106.06020, 2021.

[59] Jennifer C. White and Ryan Cotterell. Equivariant transduction through invariant alignment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4651–4663, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics.

[60] Marysia Winkels and Taco Cohen. Pulmonary nodule detection in ct scans with equivariant cnns. *Medical image analysis*, 55:15–26, 2019.

[61] Jeffrey Wood and John Shawe-Taylor. Theory of symmetry network structure. Project report, 1993.

[62] Jeffrey Wood and John Shawe-Taylor. Representation theory and invariant neural networks. *Discrete Applied Mathematics*, 69(1):33–60, 1996.

[63] Jeffrey Wood and John Shawe-Taylor. A unifying framework for invariant pattern recognition. *Pattern Recognition Letters*, 17(14):1415–1422, 1996.

[64] Yinshuang Xu, Jiahui Lei, Edgar Dobriban, and Kostas Daniilidis. Unified fourier-based kernel and nonlinearity design for equivariant networks on homogeneous spaces. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 24596–24614. PMLR, 2022.

[65] Sicheng Zhu, Bang An, and Furong Huang. Understanding the generalization benefit of model invariance from a data perspective. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4328–4341. Curran Associates, Inc., 2021.

[66] Xupeng Zhu, Dian Wang, Ondrej Biza, Guanang Su, Robin Walters, and Robert Platt. Sample efficient grasp learning using equivariant models. *CoRR*, abs/2202.09468, 2022.

A Proofs

A.1 Proofs of results from Section 4

A.1.1 Proof of Proposition 4.1

Proof. Reformulating (4.1) we find that with probability at least $1-\delta$ it holds that:

$$\sup_{f\in\mathcal{F}} |(P-P_n)f| \leq \mathcal{R}(\mathcal{F}_Z) + \frac{\sqrt{2\|\mathcal{F}\|_\infty \log(2/\delta)}}{\sqrt{n}}. \tag{A.1}$$

Next, we compound the bounds for $\mathcal{R}(\mathcal{F}_Z)$. The optimum $\alpha > 0$ in the first line of (4.2) is the one for which $\frac{n}{9} = \ln \mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)$. Keeping in mind (4.6), we choose $\alpha = \mathsf{diam}(\mathcal{Z}) n^{-\frac{1}{\mathsf{ddim}\mathcal{Z}}}$ in (4.2), which gives (abbreviating $d = \mathsf{ddim}\mathcal{Z}, D = \mathsf{diam}\mathcal{Z}$)

$$\begin{aligned}\mathcal{R}(\mathcal{F}_Z) &\lesssim D/n^{1/d} + n^{-1/2}\int_{Dn^{-1/d}}^{\infty} (D/t)^{d/2} dt = D/n^{1/d} + (D/n)^{1/2}\int_{n^{-1/d}}^{\infty} \tau^{-d/2} d\tau \\ &= \frac{D}{n^{1/d}}\left(1 + \frac{2}{d-2}\right).\end{aligned}$$

Now the last equation and equation (A.1) yield the claim. □

A.1.2 Proof of Proposition 4.2

We start with a preliminary result under hypotheses of strict equivariance. In this case, we are able to use a change of variables to reduce the generalization error formula to an equivalent one depending only on a measurable choice of G-orbit representatives of elements from $\mathcal{Z}$:

Proposition A.1. *Let $\mathcal{F}$ be a set of G-invariant functions, and let $\mathcal{Z}^0 \subset \mathcal{Z}$ be a choice of G-orbit representatives for points in $\mathcal{Z}$, such that $\iota^0 : \mathcal{Z} \to \mathcal{Z}^0$ associating to each $z \in \mathcal{Z}$ its orbit representative z^0, is Borel measurable. Let $\mathcal{F}^0 := \{f|_{\mathcal{Z}^0} : \ f \in \mathcal{F}\}$ and denote by $\iota^0(\mathcal{D})$ the image measure of $\mathcal{D}$. Then for each $n \in \mathbb{N}$ if $\{Z_i\}_{i=1}^n$ are i.i.d. samples with $Z_i \sim \mathcal{D}$ and $Z_i^0 := \iota^0 \circ Z_i$, we have*

$$\mathsf{GenErr}(\mathcal{F}, \{Z_i\}, \mathcal{D}) = \mathsf{GenErr}^{(G)}(\mathcal{F}, \{Z_i\}, \mathcal{D}) = \mathsf{GenErr}(\mathcal{F}^0, \{Z_i^0\}, \iota^0(\mathcal{D})).$$

Proof. For the first equality, we use the definition of GenErr and the change of variable formula (3.1) and the fact that G-invariant functions f satisfy $f(Z) = \mathbb{E}_g[f(g \cdot Z)]$. For the second equality, note that by hypothesis, for each $f \in \mathcal{F}$ we have $f(z) = f(g \cdot z)$ for all $g \in G, z \in \mathcal{Z}$, in particular $f(z) = f(\iota^0(z))$ and we conclude by a change of variable by the map ι^0 in the expectations from the definition of $\mathsf{GenErr}^{(G)}$. □

Now the proof Proposition 4.2 combines the above idea with a simple extra step:

Proof of Proposition 4.2: The proof uses the triangular inequality. For $f \in \mathcal{F}$ and $g \in \mathsf{Stab}_\epsilon$, we have:

$$|Pf - P_n f| = \left|\mathbb{E}[f(Z)] - \frac{1}{n}\sum_{i=1}^n f(Z_i)\right| \leq \left|\mathbb{E}[g \cdot f(Z)] - \frac{1}{n}\sum_{i=1}^n g \cdot f(Z_i)\right| + 2\|g \cdot f - f\|_\infty.$$

By averaging over $g \in \mathsf{Stab}_\epsilon$, we obtain the inequality in the statement of the proposition. The equality follows by a change of variable via map ι_ϵ^0, exactly as in the strategy of Proposition A.1. □

A.1.3 Proof of Corollary 4.3 and of the more general result of Theorem A.3

In order to make the treatment better digestible, we first consider the intuitively simpler case of strict equivariance, and then describe how to extend it to the more general case of approximate and partial equivariance. In this case, if we restrict our equivariant functions to only the space of orbit representatives $\mathcal{Z}^0$, the dimension counts from classical generalization bounds of Proposition 4.1 improve as follows:

Corollary A.2. *Assume that $\mathcal{F}$ is composed of G-invariant functions and that $d_0 := \mathsf{ddim}(\mathcal{Z}^0) > 2$. Also, denote $D_0 := \mathsf{diam}(\mathcal{Z}^0)$. With the same notation as in Proposition A.1 and with the hypotheses of Proposition 4.1, for any probability distribution $\mathcal{D}$ over $\mathcal{Z}$, the following holds with probability at least $1-\delta$:*

$$\mathsf{GenErr}(\mathcal{F}, \{Z_i\}, \mathcal{D}) \lesssim \frac{d_0}{d_0 - 2}\left(\frac{D_0^{d_0}}{n}\right)^{1/d_0} + n^{-1/2}\sqrt{\|\mathcal{F}\|_\infty \log(2/\delta)}.$$

Proof. Due to Proposition A.1, we only need to bound $\mathsf{GenErr}(\mathcal{F}^0, \{Z_i^0\}, \iota^0(\mathcal{D}))$. Thus it suffices to apply Proposition 4.1 for the above function. We note that $\|\mathcal{F}^0\|_\infty \leq \|\mathcal{F}\|_\infty$ to conclude. □

The drawback of the above Corollary, is that it leaves open the question of *how to actually bound* the diameter and dimension of $\mathcal{Z}^0$, on which we do not have direct control. The next steps we take consist precisely in translating the information from properties of G to relevant properties of $\mathcal{Z}^0$.

A first, simpler, approach could be the following. Under the reasonable assumption that $\mathcal{Z}, \mathcal{Z}^0$ have diameter greater than 1, the leading term on the left in Corollary A.2 is n^{-1/d_0}.Thus the optimal choices of $\mathcal{Z}^0$ are those which minimize the doubling dimension $d_0 = \mathsf{ddim}(\mathcal{Z}^0)$ amongst sets of representatives of G-orbits. This is a weak regularity assumption, implying that we want $\mathcal{Z}^0$ to not oscillate wildly. The effect of G on coverings is evident in case G, $\mathcal{Z}^0$ are manifolds, and $\mathcal{Z} = \mathcal{Z}^0 \times G$ (see (A.3) for the strictly equivariant case, and the more general (A.5) for the general case). Since ddim coincides with topological dimension, we immediately have

$$d_0 = d - \mathsf{dim}(G).$$

Intuitively, the dimensionality of G can be understood as eliminating degrees of freedom from $\mathcal{Z}$, and it is this effect that improves generalization by $n^{-1/(d-\mathsf{dim}(G))} - n^{-1/d}$.

In order to include more general situations, we now describe a second, more in-depth approach. We take a step back and rather than addressing direct diameter and dimension bounds for $\mathcal{Z}^0$, we go "back to the source" of Proposition 4.1. We update the bounds on covering numbers of $\mathcal{Z}^0$, directly in terms of the G-action and of $\mathcal{Z}$. The ensuing framework is robust enough to later include, after a few adjustments, also the cases of partial and approximate equivariance. Here is our fundamental bound, which generalizes and extends [49, Thm.3].

Theorem A.3. *Assume that $\mathcal{Z}$ is a metric space with distance d and $S \subset G$ is a subset of a metric group G consisting of transformations $g : \mathcal{Z} \to \mathcal{Z}$ (with action denoted $g \cdot z := g(z)$) for which there exists a choice of S-orbit representatives $\mathcal{Z}_0 \subset \mathcal{Z}$ and a distance function d_G on S satisfying the following for $L, L' \in (0, +\infty]$ (with the conventions $1/+\infty := 0$ and $\mathcal{N}(X, 0) := +\infty, \mathcal{N}(X, +\infty) := 0$):*

1. *For all $z_0, z_0' \in \mathcal{Z}_0$ and all $g \in S$ it holds that $\frac{1}{L} d(z_0, z_0') \leq d(g \cdot z_0, g \cdot z_0') \leq L' d(z_0, z_0')$.*
2. *For all $g, g' \in S$ and $z_0, z_0' \in \mathcal{Z}_0$ it holds that $\frac{1}{L} d_G(g, g') \leq d(g \cdot z_0, g' \cdot z_0') \leq L' d_G(g, g')$.*

Then for each $\delta' > 0$ the following holds

$$\mathcal{N}(\mathcal{Z}_0, 2\delta' L)\mathcal{N}(S, 2\delta' L) \leq \mathcal{N}(\mathcal{Z}, \delta') \leq \mathcal{N}(\mathcal{Z}_0, \delta'/2L')\mathcal{N}(S, \delta'/2L'). \tag{A.2}$$

Before the proof, we observe how Corollary 4.3 can be recovered using the choice $L > 0, L' = +\infty$ in Theorem A.3:

Proof of Corollary 4.3: We apply Theorem A.3 to $S = \mathsf{Stab}_\epsilon$ and $\mathcal{Z}_0 = \mathcal{Z}_\epsilon^0$ as in the corollary statement. Then we take $\delta = 2L\delta'$ in the conclusion (A.2), and consider only the lower bound inequality, which directly gives the claim of the corollary. □

Proof of Theorem A.3: In this proof, we will denote a minimal α-ball cover of a metric space X by X_α.

Note that we are not assuming G to be a group, but due to lower bound in property 1. it follows that $z_0 \mapsto g \cdot z_0$ is injective (for $z_0 \neq z_0'$ we have $d(z_0, z_0') > 0$ and thus $g \cdot z_0 \neq g \cdot z_0'$), and when below we write "g^{-1}" this has to be interpreted as the inverse of the g-action, restricted to its image.

Further, note that the case when one of L, L' is $+\infty$, corresponds to removing the part of the assumptions (and of the conclusions) involving that value, thus we only consider the case of finite L' and finite L.

On fixing an arbitrary point $z \in \mathcal{Z}$, we can write $z = g \cdot z_0$ for a suitable $g \in G, z_0 \in \mathcal{Z}^0$. Let $\eta := \delta'/2L'$. For fixed covers $\mathcal{Z}^0_\eta, G_\eta$, there exists a point $z_0' \in \mathcal{Z}^0_\eta$ with $d(z_0', z_0) < \eta$ and $g' \in G_\eta$ with $d_G(g', g) < \eta$. Thus by property 1. we have $d(g \cdot z_0', z) < L'\eta = \epsilon/2$ and by property 2. we have $d(g' \cdot z_0', g \cdot z_0') < L'\eta = \delta'/2$. By the triangle inequality, $d(g' \cdot z_0', z) < \delta'$ and thus $G_\eta \cdot \mathcal{Z}^0_\eta$ is an δ'-cover of $\mathcal{Z}$. Thus we have

$$\mathcal{N}(\mathcal{Z}, \delta') \leq \#G_{\delta'/2L'} \, \#\mathcal{Z}^0_{\delta'/2L'},$$

optimizing over the cardinalities on the right hand side yields the second inequality in (A.2).

Now consider an δ'-cover $\mathcal{Z}_{\delta'}$ of $\mathcal{Z}$, and for $\eta = 2\delta L$ consider an η-cover G_η of G. We find that for each $z \in \mathcal{Z}_{\delta'}$, there exists at most one $g \in G_\eta$ such that $\mathsf{dist}(z, g \cdot \mathcal{Z}_0) < \eta/2L = \delta'$. Notice that if this were false, we could use the triangle inequality and contradict property 2. in the statement. For each $g \in G_\eta$ denote Z_g the set of such points $z \in \mathcal{Z}_{\delta'}$ such that there exists $x \in g \cdot \mathcal{Z}^0$, and assign exactly one such $x = x(z)$ to each z, forming a set X_g of all such $x(z)$. Any other point $x' \in g \cdot \mathcal{Z}^0$ such that $d(x', z) < \delta'$ then satisfies $d(x', x) < 2\delta'$ by triangle inequality, and thus X_g is a $2\delta'$-cover of $g \cdot \mathcal{Z}_0$. If for $g \cdot z_0 \in g \cdot \mathcal{X}_0$ the point $x \in g \cdot \mathcal{X}_0$ satisfies $d(g \cdot z_0, x) < 2\delta'$, then by property 1. in the statement we have $d(z_0, g^{-1} \cdot x) < 2\delta' L$, and thus $g^{-1} \cdot X_g$ is a $2\delta' L$-cover of $\mathcal{Z}_0$, having the same cardinality as Z_g. We then compute as follows, proving the first inequality in (A.2):

$$\mathcal{N}(\mathcal{Z}, \delta') \geq \sum_{g \in G_\eta} \#Z_g = \sum_{g \in G_\eta} \#(g^{-1} \cdot X_g) \geq \mathcal{N}(G, 2\delta' L)\mathcal{N}(\mathcal{Z}_0, 2\delta' L).$$

□

A.1.4 Proof of Theorem 4.4

As before, we focus again first on the exact equivariance case, where Theorem A.4 is the direct analogue to (or special case of) Theorem 4.4.

Under the hypotheses of Theorem A.3 on the G-action, we directly obtain the following, for the strictly equivariant case:

$$\mathcal{N}(\mathcal{Z}_0, \delta) \leq \frac{\mathcal{N}(\mathcal{Z}, \delta/2L)}{\mathcal{N}(G, \delta)}. \tag{A.3}$$

We next impose that for $\delta \lesssim \mathsf{diam}(G)$ the group G satisfies the natural "volume growth" assumption, where for compact groups $\mathsf{Vol}(G)$ is its $\mathsf{dim}(G)$-dimensional Hausdorff measure and $\mathsf{dim}(G)$ is the usual Hausdorff dimension, and for finite groups we use minimum separation notation $\delta_G > 0$ as defined in (4.4):

$$\textbf{Assumption:} \quad \mathcal{N}(G, \delta) \gtrsim \begin{cases} \#G/(\max\{\delta, \delta_G\})^{\mathsf{ddim}(G)}, & \text{if } G \text{ finite,} \\ \mathsf{Vol}(G)/\delta^{\mathsf{dim}G}, & \text{if } \mathsf{dim}G > 0. \end{cases} \tag{A.4}$$

Similarly to Proposition 4.1, we then get the following, in which the leading term in the bound has exponent figuring $d_0 = \mathsf{ddim}(\mathcal{Z}) - \mathsf{dim}(G)$. Recall that $\mathsf{dim}(G) = 0$ for finite groups, thus the distinction can be made directly in terms of the dimension of G.

Theorem A.4. *Let $\delta > 0$ be fixed. Assume that for a given ambient group G the group G satisfies assumption* (A.4) *and that its action satisfies the the assumptions 1. and 2. of Theorem A.3 for some finite $L > 0$ and $L' = +\infty$. We denote $d_G = \mathsf{ddim}(G)$ if G is a discrete group and $d_G = \mathsf{dim}(G)$ if G is compact and non-discrete, and $d = \mathsf{ddim}(\mathcal{Z})$. Furthher assume $d_0 := d - d_G > 2$. Furthermore, set $|G| := \mathsf{Vol}(G)$ if $\mathsf{dim}G > 0$ and $|G| := \#G$ for finite G. Then with the notations of Proposition 4.1 we have with probability $\geq 1 - \delta$*

$$\mathsf{GenErr}(\mathcal{F}, \{Z_i\}, \mathcal{D}) \lesssim n^{-1/2}\sqrt{\|\mathcal{F}\|_\infty \log(2/\delta)} + (E),$$

where

$$(E) := \begin{cases} \frac{d_0}{d_0-2}\delta_G^{-d_0/2+1}\left(\frac{(2L)^d D^d}{|G|n}\right)^{1/2} & \text{if } G \text{ is finite and } (2L)^d D^d < |G| n\ \delta_G^{d_0}, \\ \frac{d_0}{d_0-2}\left(\frac{(2L)^d D^d}{|G|n}\right)^{1/d_0} & \text{otherwise.} \end{cases}$$

Proof. We follow the same computation as Proposition 4.1, but use Proposition A.1 in order to reduce to restrictions of functions to $\mathcal{Z}^0$. In this case, using (A.3) and assumption (A.4), and with notation as in our statement, we will have:

$$\mathcal{N}(\mathcal{Z}_0, t) \lesssim \frac{(2L)^d D^d}{|G|} \max\{\delta_G, t\}^{-d_0},$$

where we have $\delta_G = 0$ for $\mathsf{dim}G > 0$. We set $C := \frac{(2L)^d D^d}{|G|}$ for simplicity of notation. In case $C\delta_G^{d_0} < 1$ (which includes the case $\mathsf{dim}G > 0$), we take $\alpha = (C/n)^{1/d_0}$ in the Dudley integral (4.2) and find

$$\mathcal{R}(\mathcal{F}_{Z^0}) \lesssim \alpha + n^{-1/2} \int_\alpha^\infty \sqrt{C t^{-d_0}} dt,$$

from which the proof goes exactly as in Proposition 4.1, with C replacing D^d, and we get the second option for the value of (E) as given in our statement. In case $C\delta_G^{d_0} < 1$ instead we take $\alpha = 0$ and our above bound for $\mathcal{N}(\mathcal{Z}_0, t)$ plugged into (4.2) (recalling the notation for C):

$$\mathcal{R}(\mathcal{F}_{Z^0}) \lesssim \int_0^\infty \sqrt{C \max\{\delta_G, t\}^{-d_0}} dt = \delta_G^{-d_0/2+1}\sqrt{C/n} + \sqrt{C/n} \int_{\delta_G}^\infty t^{-d_0/2} dt,$$

from which the second case of the value of (E) follows by direct computation. □

Now the proof of Theorem 4.4 proceeds in exactly the same manner as the above. Below we explain the required adaptations:

Proof of Theorem 4.4: The following updates are the principal adaptations required for the above proof of Theorem A.4:

- The role of G should be replaced by Stab_ϵ, except for the fact that parameters δ_G, d_G remain unchanged (i.e. we use their values corresponding to "ambient" group G rather than those for Stab_ϵ).
- The G-orbit representative set $\mathcal{Z}^0$ then should be replaced by representatives $\mathcal{Z}^0_\epsilon$ for orbits of Stab_ϵ.

With these changes, assumption (A.4) implies its more general version, assumption (4.7). Indeed, $|G|$ equals $\#G$ for finite G and $\mathsf{Vol}(G)$ for compact G, and $\delta_G > 0$ only in the first case. Furthermore, we have $\delta_{\mathsf{Stab}_\epsilon} \geq \delta_G$ as a direct consequence of $\mathsf{Stab}_\epsilon \subseteq G$.

We observe that Corollary 4.3 (which also is obtained from Theorem A.3 with the above two main substitutions) directly gives the version of Theorem (A.3) required to get the correct replacement of (A.3) under our initially declared two substitutions. We get:

$$\mathcal{N}(\mathcal{Z}^0_\epsilon, \delta) \leq \frac{\mathcal{N}(\mathcal{Z}, \delta/2L)}{\mathcal{N}(\mathsf{Stab}_\epsilon, \delta)}. \tag{A.5}$$

With the above changes, the proof follows by exactly the same steps as in the above proof of Theorem A.4. □

Remark A.5. Note that, as might be evident from the last proof, we could have introduced new more precise parameters to keep track of dimensionality and minimum separation for Stab_ϵ rather than formulating assumption (4.7) in terms of d_G, δ_G. This is justified for the aims of this work. Indeed, all the main situations of interest to us are those in which Stab_ϵ is a "large" subset of G, i.e. it has dimension d_G, and in all our examples for finite groups $\delta_G > 0$, the minimum separation for Stab_ϵ is within a small factor of δ_G itself.

A.2 Proof of Proposition 5.1

Proof. We have that $Z = (X, Y)$ is a data distribution in which by our assumption 2. preceding the proposition, we have that almost surely $Y = y^*(X)$ for a deterministic function y^*. With this notation, we may write

$$\mathbb{E}_Z[f(Z)] = \mathbb{E}_{X_\epsilon^0}\mathbb{E}_{g|X_\epsilon^0}[f(g \cdot X_\epsilon^0, y^*(g \cdot X_\epsilon^0)].$$

Recalling that we restrict to functions of the form $f(x, y) = \ell(\widetilde{f}(x), y) = d_{\mathcal{Y}}(\widetilde{f}(x), y)^2$, we first consider the precise equivariance case $\epsilon = 0$. In this case for $g \in \mathsf{Stab}_\epsilon(\mathcal{F})$ we find also $\widetilde{f}(g \cdot X_\epsilon^0) = g \cdot \widetilde{f}(X_\epsilon^0)$ and thus when optimizing over $\widetilde{f}$ we have to determine the optimal value of $y = \widetilde{f}(X_\epsilon^0)$ to be associate to each X_ϵ^0. Thus as a consequence of all the above, if $\widetilde{\mathcal{F}}$ would be the class of all precisely Stab_ϵ-equivariant measurable functions, we would get the following rewriting:

$$\mathsf{AppGap}(\mathcal{F}, \mathcal{D}) = \min_{\widetilde{f} \in \widetilde{\mathcal{F}}} \mathbb{E}_Z[d_{\mathcal{Y}}(\widetilde{f}(X), y^*(X)] = \mathbb{E}_{X_\epsilon^0} \min_{y \in \mathcal{Y}} \mathbb{E}_{g|X_\epsilon^0} \left[d_{\mathcal{Y}}(g \cdot y, y^*(g \cdot X_\epsilon^0)^2\right]. \quad \text{(A.6)}$$

For $\epsilon > 0$, for each fixed $X_\epsilon^0 = x_\epsilon^0$ we may further perturb the associated $y = \widetilde{f}(x_\epsilon^0)$ by at most ϵ in the direction of $y^*(X_\epsilon^0)$, while still obtaining a measurable function with approximation ℓ_∞-norm error bounded by ϵ, thus the above bound is improved to

$$\mathsf{AppErr}(\mathcal{F}, \mathcal{D}) \leq \mathbb{E}_{X_\epsilon^0} \min_y \mathbb{E}_{g|X_\epsilon^0} \left[\left(d_{\mathcal{Y}}(g \cdot y, y^*(g \cdot X_\epsilon^0)) - \epsilon\right)_+^2\right], \quad \text{(A.7)}$$

as desired. In case $\widetilde{\mathcal{F}}$ contains a strict subset of measurable invariant functions with error ϵ, we would only get an inequality instead of the first equality in (A.6) but we still have the same bound as in (A.7), and thus the proof is complete. □

A.3 Proof of Theorem 5.2

Proof. We use the isodiametric inequality (5.1) in G, applying it to $\mathsf{Stab}_\epsilon(\mathcal{F})$ for $\mathcal{F} \in \mathcal{C}_{\epsilon,\lambda}$. Then by taking $\mathsf{Stab}_\epsilon(\mathcal{F}) = X$ which is optimal for inequality (5.1) we can saturate the two bounds (modulo discretization errors for discrete G) and we get

$$\frac{|\mathsf{Stab}_\epsilon(\mathcal{F})|}{|G|} \simeq \lambda, \quad \mathsf{diam}(\mathsf{Stab}_\epsilon(\mathcal{F})) \simeq C_G \lambda^{1/\mathsf{ddim}(G)}.$$

We next use Lipschitz deformation bounds and find that for all $x \in \mathcal{X}$ we have

$$\begin{aligned} \mathsf{diam}\left\{y^*(g \cdot x) :\ g \in \mathsf{Stab}_\epsilon(\mathcal{F})\right\} &\leq \mathsf{diam}(\mathsf{Stab}_\epsilon(\mathcal{F}))\ \mathsf{Lip}(y^*)\ L' \\ &\leq C_G \lambda^{1/\mathsf{ddim}(G)} \epsilon(\mathcal{F}))\ \mathsf{Lip}(y^*)\ L'. \end{aligned}$$

Then we use Proposition 5.1 for $\mathcal{F}$ and observe that when g, X_ϵ^0 are random variables as in the proposition, in particular $g \in \mathsf{Stab}_\epsilon(\mathcal{F})$ and for each $X_\epsilon^0 = x_\epsilon^0$ we find the following estimate valid uniformly over $y \in \left\{y^*(g \cdot X_\epsilon^0) :\ g \in \mathsf{Stab}_\epsilon(\mathcal{F})\right\}$:

$$d_{\mathcal{Y}}(y, y^*(g \cdot x_\epsilon^0)) \leq C'_G \lambda^{1/\mathsf{ddim}(G)}\ \mathsf{Lip}(y^*)\ L'.$$

In a similar way, we also find

$$d_{\mathcal{Y}}(y, g \cdot y) \leq C'_G \lambda^{1/\mathsf{ddim}(G)}\ L'.$$

By triangle inequality, and using the assumption that $\mathsf{Lip}(y^*) \simeq 1$ it follows that

$$d_{\mathcal{Y}}(y, y^*(g \cdot x_\epsilon^0) \leq C'_G \lambda^{1/\mathsf{ddim}(G)}(1 + \mathsf{Lip}(y^*))L' \lesssim C'_G \lambda^{1/\mathsf{ddim}(G)}\ \mathsf{Lip}(y^*)\ L'.$$

Then we may perturb each y by ϵ in order to possibly diminish this value without violating the condition defining $\mathsf{Stab}_\epsilon(\mathcal{F})$, and with these choices we obtain the claim. □

A.4 Finding the optimal $\lambda = \lambda^*$ for the bound of Theorem 6.1

We note that λ^* minimizing $C\lambda^\alpha + C'\lambda^{-\beta}$ for $\alpha, \beta > 0$ is given by

$$\lambda^* = \left(\frac{\beta}{\alpha}\frac{C'}{C}\right)^{1/(\alpha+\beta)}.$$

recall that in our case have the following choices, for case 1 and case 2 in the theorem's statement.

$$\alpha_1 = \alpha_2 = 1/d_G, \qquad \beta_1 = 1/2, \beta_2 = 1/d_0,$$

and

$$\begin{aligned} C &= C_G \mathsf{Lip}(y^*) L', \\ C'_1 &\simeq \frac{(2LD)^{d/2}|G|^{1/2}}{\delta_G^{(d_0-2)/2}}, \\ C'_2 &\simeq (2LD)^{d/d_0}|G|^{1/d_0}, \end{aligned}$$

and thus the optimal choice of λ is

$$\begin{aligned} \text{in case } n\lambda \geq C_3, \quad \lambda^* &= \left(\frac{2}{d_G}\frac{(2LD)^{d/2}|G|^{1/2}}{\delta_G^{(d_0+2)/2}}\right)^{2d_G/(d_G+2)} n^{-d_G/(d_G+2)}, \\ \text{in case } n\lambda > C_3, \quad \lambda^* &= \left(\frac{d_0}{d_G}(2LD)^{d/d_0}|G|^{1/d_0}\right)^{d_0 d_G/(d_0+d_G)} n^{-d_G/(d_G+2)}. \end{aligned}$$

B Examples

We describe some concrete examples of partial and approximate equivariance using the language we used in section 3.2 while sourcing them from existing literature. But first, we expand a little on our equivariance error notation.

B.1 Equivariance error notation

Recall that the action of elements of an ambient group G over the product space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ may be written as follows: For coordinates $z = x \times y$ we may write $g \cdot z = (g \cdot x, g \cdot y)$, and thus for $\widetilde{f} : \mathcal{X} \to \mathcal{Y}$ and $f(x, y) = \ell(\widetilde{f}(x), y)$, we have the action

$$(g \cdot f)(z) := f(g \cdot z) = \ell(\widetilde{f}(g \cdot x), g \cdot y).$$

For the equivariance error of g, f, interpreted as "the error of f's approximate equivariance under the action by g", we get the following, which is valid in the common situations in which $\ell(y, y') \geq 0$ in general with $\ell(y, y) = 0$ for all $y \in \mathcal{Y}$:

$$\mathsf{ee}(f, g) := \|g \cdot f - f\|_\infty = \sup_{x,y} \left|\ell(\widetilde{f}(g \cdot x), g \cdot y) - \ell(\widetilde{f}(x), y)\right| \geq \sup_x \ell(\widetilde{f}(g \cdot x), g \cdot \widetilde{f}(x)), \quad \text{(B.1)}$$

where the last inequality follows by restricting the supremum from $\mathcal{X} \times \mathcal{Y}$ to the graph of $\widetilde{f}$, namely by imposing $y = \widetilde{f}(x)$.

In several recent works, the equivariance error is defined simply by comparing $g \cdot \widetilde{f}(x) and \widetilde{f}(g \cdot x)$, as in the rightmost term of (B.1), thus it is lower than the one found here. We provide a justification for our definition of the equivariance error:

- The loss ℓ is the integrative part of the model, thus a definition for equivariance error which does not include it will only detect partial information concerning the influence of symmetries.
- The notion of equivariance error defined via ℓ simplifies the comparison between $\widetilde{f}$ and data distributions $\mathcal{D}$.

B.2 Examples

B.2.1 Imperfect translation equivariance in classical CNNs

We consider here the most common examples of group equivariant convolutional networks (GCNNs), which are the usual Convolutional Neural Networks (CNNs) for computer vision tasks. We follow observations from [25] and [11], and connect the underlying ideas to Theorem 6.1.

Setting of the problem. We consider a usual CNN layer, keeping in mind a segmentation task, where both $\mathcal{X} = \mathcal{Y}$ represent spaces of images. More precisely, we think of images as pixel values at integer-coordinate positions taken from the index set $\mathbb{Z} \times \mathbb{Z}$. We also assume that the relevant information of each image only comes from a square of size $n \times n$ pixels, outside which the pixel values are taken to be 0. We consider the application of a single convolution kernel/filter, of $k \times k$ pixels (with k a small odd number). One typically applies padding by a layer of 0's of size $(k-1)/2$ on the perimeter of the $n \times n$ square, after which convolution with the kernel is computed on the $n \times n$ central pixels of the resulting $(n+k-1) \times (n+k-1)$ padded input image. The output relevant information is restricted to a $n \times n$ square, outside which pixel values are set to 0 again, via padding.

Metric on $\mathcal{X}$. As a natural choice of distance over $\mathcal{X}$ we may consider L^2-difference between pixel-value functions, or interpret pixel values as probability densities, and use Wasserstein distance, or consider other ad-hoc image metrics.

Group action: translations. The group acting on our "pixel-value functions" is the group of translations with elements from $\mathbb{Z} \times \mathbb{Z}$. We expect the following invariance for the segmentation function $f : \mathcal{X} \to \mathcal{X}$:

$$f(v \cdot x) = v \cdot f(x),$$

where $x \in \mathcal{X}$ represents an image with pixel values assigned to integer coordinates and $v \in \mathbb{Z}^2$ is a translation vector and (in two alternative notations) $v \cdot x = \tau_v(x)$ is the result of translating all pixel values of x by v.

Deformation properties of the action. If we take the previously mentioned distance functions on $\mathcal{X}$ and the usual distance induced from $\mathbb{R}^2$ over translation vectors v, it is easy to verify that the assumptions of Theorem A.3 about the action of translations hold, and the Lipschitz constants with respect to the metric on $\mathcal{X}$ only depend on the mismatch near the boundary, due to "zero pixels moving in" and to "interior pixels moving out" of our $n \times n$ square, and being truncated. The ensuing bounds only depend on the precise distance that we introduce use on $\mathcal{X}$.

More realistic actions. An alternative more realistic definition of $\mathbb{Z} \times \mathbb{Z}$-action consists of defining $v \cdot x$ as the truncation of $\tau_v(x)$ where, for pixels outside our "relevant" $n \times n$ square we set pixel value to 0 after the translation.

Problems near the boundary. Nevertheless, the updated translation action, will move pixel values of 0, coming from pixels outside the $n \times n$ square, and will create artificial zero pixel values inside the image $v \cdot x$, different than the values that would be present in real data.

Imperfect equivariance of data. Also, even in the latter more realistic alternative, the above translation equivariance is not respecting by segmentation input-output pairs coming from finite $n \times n$ images, since, independently of n, the boundary pixel positions translated by v, fall outside the original image.

Approximate stabilizer. In any case, we have to restrict the choices of v to integer-coordinate elements of a (subset of a) $n \times n$ square, containing only the translations that are relating "real" segmentations that appear within our $n \times n$ relevant window. It is thus

natural to restrict Stab_ϵ to only include vectors in a smaller subset of $\mathbb{Z} \times \mathbb{Z}$, of cardinality

$$|\mathsf{Stab}_\epsilon| \leq n^2.$$

The value of error ϵ may quantify the allowed error (or noisiness) for our model sets.

Reducing to finite G and computing λ. For the sake of computing λ in our Theorem 6.1, we can observe that input and output values outside the "padding perimeter" given by a $(n+k-1) \times (n+k-1)$ square, are irrelevant, and thus we may actually periodize the images and consider the images as subsets of a *padded torus* $(\mathbb{Z}/(n+k-1)\mathbb{Z})^2$, which we consider as acting on itself by translations. In this case $G \simeq (\mathbb{Z}/(n+k-1)\mathbb{Z})^2$, so that

$$|G| = (n+k-1)^2 \quad \text{and thus} \quad \lambda = \frac{|\mathsf{Stab}_\epsilon|}{(n+k-1)^2}.$$

Further extensions. In [25], it is argued that convolutional layers in classical CNNs are not fully translation-equivariant, and can encode positional information, due to the manner in which boundary padding is implemented. Possible solutions are increasing the padding to size k (so that the padded image is a square of size $(n+2k) \times (n+2k)$) or extending images by periodicity, via so-called "circular padding" (which transforms each image into a space equivalent to the torus $(\mathbb{Z}/n\mathbb{Z})^2$). In either case, the application of actions by translations by vectors that are too large compared to the image size of $n \times n$, will increase the mismatch between model equivariance and data equivariance.

Stride and downsampling. In [11], a different equivariance error for classical CNNs is studied, related to the use of stride > 1 in order to lower the output dimensions of CNN layer outputs. If for example we use stride 2 when defining layer operation $f : \mathcal{X} \to \mathcal{Y}$, then $\mathcal{Y}$ will have relevant pixel values only in an $n/2 \times n/2$-square, and we apply the $k \times k$ convolution kernel only at positions with coordinates in $(2\mathbb{Z}) \times (2\mathbb{Z})$ from image x. In this case we require that translations by group elements $v \in (2\mathbb{Z}) \times (2\mathbb{Z})$ on x have the effect of a translation by $v/2 \in \mathbb{Z}$ on the output. However for shifts in the input via vectors v that do not have two even coordinates, we may not have have an explicit corresponding action on the output, and in [11] a solution via adaptive polyphase sampling is proposed. A possibility for studying the best polyphase sampling strategy via almost equivariance, would be to include a bound for equivariance error $\epsilon > 0$ and consider the optimization problem of finding the polyphase approximation that minimizes theoretical or empirical quantifications of ϵ. As a benchmark (modelled on the case of infinite images without boundary effects) one could compare the above to the action via $\mathsf{Stab}_0 = (2\mathbb{Z}) \times (2\mathbb{Z})$ which has $\lambda = 1/4$ within the ambient group $G = \mathbb{Z} \times \mathbb{Z}$. Our Theorem 6.1 can be used to compare the effects of increasing or decreasing ϵ, λ, in terms of data symmetry.

B.2.2 Partial equivariance in GCNNs

In this section, we connect the main results from [38] to our setup. In [38], one of the main motivating examples was to consider rotations applied to a handwritten digit and revert them. The underlying group action was via $SO(2)$ and only rotations of angles between $[-60°, 60°]$ were permitted in one case, which allowed to not confound rotated digits "3" and "9" for example.

The above task can be formulated on a space of functions $f : \mathcal{X} \to \mathcal{Y}$ in which $\mathcal{X}$ represents the space of possible images and $\mathcal{Y}$ the labels. Elements $(Y_d, Y_\theta) \in \mathcal{Y}$ include a digit classification label Y_d and a rotation angle value Y_θ.

We consider actions by group $G = SO(2) = \{R_\phi : \ \phi \in \mathbb{R}/360\mathbb{Z}\}$, where R_ϕ is the rotation matrix by angle ϕ, and the group operation corresponds to summing the angles, modulo $360°$ (or in radians, modulo 2π). The action of R_ϕ over $\mathcal{X}$ would be by rotation as usual (R_θ sends image $x \in \mathcal{X}$ to $R_\theta \cdot x$, now rotated by θ), and over $\mathcal{Y}$ we consider the action by

$$R_\phi(Y_d, Y_\theta) = (Y_d, Y_\theta + \phi),$$

i.e. the restriction of the action on the digit label leaves it invariant and the restriction of the action on the angle label is non-trivial, giving a shift on the label.

The optimum labelling assigns to x a label $y^*(x)$ enjoying precise equivariance under the above definitions of the actions, and thus we are allowed to permit equivariance error $\epsilon = 0$. However as mentioned above, applying rotations outside the range $\theta \in [-60°, 60°]$ to the data would surely bring us outside the labelled data distribution, thus we are led to take

$$\epsilon = 0, \quad \mathsf{Stab}_0 = \{R_\theta : \ \theta \in [-60°, 60°]\}.$$

We then have that $|\mathsf{Stab}_0|/|SO(2)| = 1/3$, with respect to the natural Haar measure on rotations. It is natural to think of the set of Stab_0-action representatives of images $\mathcal{X}_0^0$ given as the "unrotated" images. If we take a digit image that is rotated, say by 20°, from its "base" version, and we apply a rotation of 50° to it (i.e. an element of Stab_0), then we reach the version of the image now excessively rotated by 70°. This means that without further modification, considering model symmetries with Stab_0 taken to be independent of the point, would automatically generate some error when tested on the data. While decreasing the threshold angle in the definition of Stab_0 from 60° would limit this effect, it will also decrease generalization error in the model. The study of point-dependent invariance sets Stab_ϵ is interesting in view of this example application, but it is outside the scope of the current approximation/generalization bounds and is left for future work.

B.2.3 Possible applications to partial equivariance in Reinforcement Learning

The use of approximate invariances for RL applications was considered in [21, Sec. 6] via soft equivariance constraints allowing better generalization for an agent moving in a geometric environment. While imposing approximate equivariance for memoryless G-action for groups such as $G = SO(2), \mathbb{Z}_2$ has produced positive results, it may be interesting, in analogy to the previous section, to include memory, and thus restrict the choices of group actions across time steps. Note that for a temporal evolution of T steps, the group action by G acting independently at each step would produce a T-interval action via the product group G^T, and allowing for a partial action via Stab_ϵ, with possibly increased fitness to evolving data. More precise time-dependence prescriptions and consequences within Q-learnig are left to future work.

C Discussion of [19]

In section 2 we mentioned [19] (reference [18] in the main file), which considers PAC-style bounds under model symmetry. [19] works with compact groups and argues how the learning problem in such a setup reduces to only working with a set of reduced orbit representatives, which leads to a generalization gain. This message of [19] is similar to ours, although we work with a more general setup. However, we noted that the main theorem in [19] has an error. Here, we briefly sketch the issue with the proof.

One of the main quantities of interest in the main theorem of [19] is $D_\tau(\mathcal{X}, \mathcal{H})$ (notation from their paper), which directly comes from the following bound, and is claimed to have a linear dependence on $Cov(\mathcal{X}, \rho, \delta)$. Again, for the sake of easier verification, we follow their notation. Crucially, note that [19] uses the notation Cov as analogous to our $\mathcal{N}$:

$$Cov(\mathcal{H}, \|\cdot\|_{L_\infty}, 2C\delta + \kappa) \leq Cov(\mathcal{X}, \rho, \delta) \sup_{x \in \mathcal{X}} Cov(\mathcal{H}(x), \|\cdot\|_\infty, \kappa)$$

However, the correct application of the Kolmogorov-Tikhomirov estimate shows that the reasoning in the proof should yield a dependence which is exponential in $Cov(\mathcal{X}, \rho, \delta)$, not linear. To see this, set $s = 2$ (sufficient for our purposes) in equation 238 in [52] (page 186). In other words, it is not possible to cover Lipschitz functions in infinity norm by only using constant functions.