Provably Strict Generalisation Benefit for Equivariant Models

Bryn Elesedy [1] Sheheryar Zaidi [2]

Abstract

It is widely believed that engineering a model to be invariant/equivariant improves generalisation. Despite the growing popularity of this approach, a precise characterisation of the generalisation benefit is lacking. By considering the simplest case of linear models, this paper provides the first *provably non-zero* improvement in generalisation for invariant/equivariant models when the target distribution is invariant/equivariant with respect to a compact group. Moreover, our work reveals an interesting relationship between generalisation, the number of training examples and properties of the group action. Our results rest on an observation of the structure of function spaces under averaging operators which, along with its consequences for feature averaging, may be of independent interest.

1. Introduction

There is significant and growing interest in models, especially neural networks, that are invariant or equivariant to the action of a group on their inputs. It is widely believed that these models enjoy improved generalisation when the group is correctly specified. The intuition being that if the salient aspects of a task are unchanged by some transformation, then more flexible models would need to learn (as opposed to being hard-coded) to ignore these transformations, requiring more samples to generalise. Work in this area has progressed quickly (Cohen & Welling, 2016; Cohen et al., 2018; 2019) and has found application in domains where the symmetry is known a priori, for instance in particle physics (Pfau et al., 2020).

In contrast with practical successes, our theoretical understanding of invariant/equivariant models is limited. Many previous works that have attempted to address the generalisation of invariant/equivariant models, such as Sokolic et al. (2017); Sannai & Imaizumi (2019), cover only the worst case performance of algorithms. These works use complexity measures to find tighter upper bounds on the test risk for invariant/equivariant models, but a strict benefit is not demonstrated. The VC dimension governs distribution independent generalisation, so these complexity measures can show no more than a separation between the VC dimensions of a model and its invariant/equivariant version. This would imply the existence of a distribution on which the model with smaller VC dimension will generalise better, but would not rule out that on many common distributions training an invariant/equivariant model using standard procedures provides no benefit. For instance, there could be many invariant/equivariant distributions on which SGD automatically favours parameters that result in (possibly approximately) invariant/equivariant predictors, regardless of architecture.

The results of this paper move to address this issue, by quantifying exactly the generalisation benefit of invariant/equivariant linear models. We do this in the case of the minimum L_2-norm least-squares solution, which reflects the implicit bias of gradient descent in overparameterised linear models. While the linear model provides a tractable first-step towards understanding more complex models such as neural networks, the underlying ideas of this paper are equally applicable to non-linear predictors. We emphasise this by providing new perspectives on feature averaging and suggestions for how to apply the ideas of this paper to find new methods for training invariant/equivariant neural networks.

1.1. Our Contributions

The main result of this paper is Theorem 13, which quantifies the generalisation benefit of equivariance in a linear model. We define the *generalisation gap* $\Delta(f, f')$ between predictors f and f' to be the difference in their test errors on a given task. A positive generalisation gap $\Delta(f, f') > 0$ means that f' has strictly smaller test error than f. Theorem 13 concerns

[1] Department of Computer Science, University of Oxford, Oxford, United Kingdom [2] Department of Statistics, University of Oxford, Oxford, United Kingdom. Correspondence to: Bryn Elesedy <bryn@robots.ox.ac.uk>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

$\Delta(f, f')$ in the case that $f : \mathbb{R}^d \to \mathbb{R}^k$ is the minimum-norm least-squares predictor and f' is its equivariant version. Let a compact group $\mathcal{G}$ act via orthogonal representations ϕ and ψ on inputs $X \sim \mathcal{N}(0, I_d)$ and outputs $Y = h(X) + \xi \in \mathbb{R}^k$ respectively, where $h : \mathbb{R}^d \to \mathbb{R}^k$ is an equivariant linear map. Let $(\chi_\psi|\chi_\phi) = \int_\mathcal{G} \operatorname{Tr}(\psi(g)) \operatorname{Tr}(\phi(g)) \, \mathrm{d}\lambda(g)$ denote the scalar product of the characters of the representations. The generalisation benefit of enforcing equivariance in a linear model is given by

$$\mathbb{E}[\Delta(f, f')] = \operatorname{Var}[\xi] \, r(n, d)(dk - (\chi_\psi|\chi_\phi)) + \mathcal{E}_\mathcal{G}(n, d)$$

where

$$r(n, d) = \begin{cases} \frac{n}{d(d-n-1)} & n < d - 1 \\ (n - d - 1)^{-1} & n > d + 1 \\ \infty & \text{otherwise} \end{cases}$$

and $\mathcal{E}_\mathcal{G}(n, d) \geq 0$ is the generalisation gap of the corresponding noiseless problem, that vanishes when $n \geq d$. The divergence at the interpolation threshold $n \in [d - 1, d + 1]$ is consistent with the double descent literature (Hastie et al., 2019).

The quantity $dk - (\chi_\psi|\chi_\phi)$ represents the significance of the group symmetry to the task. The dimension of the space of linear maps $\mathbb{R}^d \to \mathbb{R}^k$ is dk, while $(\chi_\psi|\chi_\phi)$ is the dimension of the space of equivariant linear maps. We will see later that the quantity $dk - (\chi_\psi|\chi_\phi)$ represents the dimension of the space of linear maps that vanish when averaged over $\mathcal{G}$; it is through the dimension of this space that the symmetry in the task controls the generalisation gap. Although invariance is a special case of equivariance, we find it instructive to discuss it separately. In Theorem 7 we provide a result that is analogous to Theorem 13 for invariant predictors, along with a separate proof.

In order to arrive at Theorems 7 and 13 we make use of general results about the structure of function spaces under averaging operators. In Section 4 we show how averaging operators can be used to decompose function spaces into orthogonal subspaces of symmetric (invariant/equivariant) and anti-symmetric (vanish when averaged) functions. In Section 5 we use these insights to provide new perspectives on feature averaging. Our main results are in Section 6. Finally, in Section 7 we apply our insights to derive principled methods for training invariant/equivariant neural networks and provide open questions for future work.

2. Related Work

Implementations While there has been a recent surge in interest, symmetry is not a new concept in machine learning. Recent literature is dominated by neural networks, but other methods do exist: e.g. kernels (Haasdonk et al., 2005), support vector machines (Schölkopf et al., 1996) or feature spaces such as polynomials (Schulz-Mirbach, 1994; 1992). The engineering of invariant neural networks dates back at least to Wood & Shawe-Taylor (1996), in which ideas from representation theory are applied to find weight tying schemes that result in group invariant architectures; similar themes are present in Ravanbakhsh et al. (2017). Recent work follows in this vein, borrowing ideas from fundamental physics to construct invariant/equivariant convolutional architectures (Cohen & Welling, 2016; Cohen et al., 2018). Correspondingly, a sophisticated theory of invariant/equivariant networks has arisen (Kondor & Trivedi, 2018; Cohen et al., 2019) including universal approximation results (Maron et al., 2019; Yarotsky, 2018).

Learning and Generalisation The intuition that invariant or equivariant models are more sample efficient or generalisable is widespread in the literature, but arguments are often heuristic and, to the best of our knowledge, a provably strict (non-zero) generalisation benefit has not appeared before this paper. It was noted (Abu-Mostafa, 1993) that constraining a model to be invariant cannot increase its VC dimension. An intuitive argument for reduced sample complexity is made in Mroueh et al. (2015) in the case that the input space has finite cardinality. The sample complexity of linear classifiers with invariant representations trained on a simplified image task is discussed briefly in Anselmi et al. (2014), the authors conjecture that a general result may be obtained using wavelet transforms. The framework of robustness (Xu & Mannor, 2012) is used in Sokolic et al. (2017) to obtain a generalisation bound for interpolating large-margin classifiers that are invariant to a finite set of transformations; note that the results contain an implicit margin constraint on the training data. The generalisation of models invariant or equivariant to finite permutation groups is considered in Sannai & Imaizumi (2019). Both of Lyle et al. (2019; 2020) cover the PAC Bayes approach to generalisation of invariant models, the latter also considers the relative benefits of feature averaging and data augmentation.

3. Preliminaries

We assume familiarity with the basic notions of group theory, as well as the definition of a group action and a linear representation. The reader may consult Wadsley (2012); Serre (1977, Chapters 1-4) for background. We define some key concepts and notation here and introduce more as necessary throughout the paper.

Notation and Technicalities We write $\mathcal{X}$ and $\mathcal{Y}$ for input and output spaces respectively. We assume for simplicity that $\mathcal{Y} = (\mathbb{R}^k, +)$ is a k-dimensional real vector space (with k finite) but we expect our results to apply in other settings too. Throughout the paper, $\mathcal{G}$ will represent an arbitrary compact group that has a measurable action ϕ on $\mathcal{X}$ and a representation ψ on $\mathcal{Y}$. By this we mean that $\phi : \mathcal{G} \times \mathcal{X} \to \mathcal{X}$ is a measurable map and the same for ψ. We sometimes write gx as a shorthand for $\phi(g)x$ and similarly for actions on $\mathcal{Y}$. Some notation for specific groups: C_m and S_m are, respectively, the cyclic and symmetric groups on m elements; while $O(m)$ and $SO(m)$ are the m-dimensional orthogonal and special orthogonal groups respectively. For any matrix A we write A^+ for the Moore-Penrose pseudo-inverse and $\|A\|_{\mathrm{F}} = \sqrt{\mathrm{Tr}(A^\top A)}$ for the Frobenius/Hilbert-Schmidt norm. We write $\mathbb{G}_n(\mathbb{R}^d)$ for the Grassmannian manifold of subspaces of dimension n in $\mathbb{R}^d$. The results of this paper require some mild care with technical considerations such as topology/measurability. We do not stress these in the main paper but they do appear in the proofs, some of which are deferred to the appendix.

Invariance, Equivariance and Symmetry A function $f : \mathcal{X} \to \mathcal{Y}$ is $\mathcal{G}$-invariant if $f(\phi(g)x) = f(x)\ \forall x \in \mathcal{X}\ \forall g \in \mathcal{G}$ and is $\mathcal{G}$-equivariant if $f(\phi(g)x) = \psi(g)f(x)\ \forall x \in \mathcal{X}\ \forall g \in \mathcal{G}$. A measure μ on $\mathcal{X}$ is $\mathcal{G}$-invariant if $\forall g \in \mathcal{G}$ and any μ-measurable $B \subset \mathcal{X}$ the pushforward of μ by the action ϕ equals μ, i.e. $(\phi(g)_* \mu)(B) = \mu(B)$. This means that if $X \sim \mu$ then $\phi(g)X \sim \mu\ \forall g \in \mathcal{G}$. We will sometimes use the catch-all term symmetric to describe an object that is invariant or equivariant.

4. Symmetric and Anti-Symmetric Functions

Averaging the inputs of a function over the orbit of a group is a well known method to enforce invariance, for instance see Schulz-Mirbach (1994). Approaching this from another perspective, averaging can also be used to identify invariance. That is, a function is $\mathcal{G}$-invariant if and only if it is preserved by orbit averaging with respect to $\mathcal{G}$. The same can be said for equivariant functions, using a modified average. After introducing the relevant concepts, we will use this observation and other properties of the averaging operators to decompose function spaces into mutually orthogonal symmetric (invariant/equivariant) and anti-symmetric (vanish when averaged) subspaces. This observation provides the foundation for many results later in the paper.

4.1. Setup

Haar Measure Let $\mathcal{G}$ be a compact group. The Haar measure is the unique invariant measure on $\mathcal{G}$ and we denote it by λ. By invariance we mean that for any measurable subset $A \subset \mathcal{G}$ and for any $g \in \mathcal{G}$, $\lambda(gA) = \lambda(Ag) = \lambda(A)$. We assume normalisation such that $\lambda(\mathcal{G}) = 1$, which is always possible when $\mathcal{G}$ is compact. The (normalised) Haar measure can be interpreted as the uniform distribution on $\mathcal{G}$. See Kallenberg (2006) for more details.

Orbit Averaging For any feature map $f : \mathcal{X} \to \mathcal{Y}$, we can construct a $\mathcal{G}$-invariant feature map by averaging with respect to λ. We represent this by the operator $\mathcal{O}$, where

$$(\mathcal{O}f)(x) = \int_{\mathcal{G}} f(gx)\, \mathrm{d}\lambda(g).$$

Similarly, if ψ is a representation of $\mathcal{G}$ on $\mathcal{Y}$, we can transform f into an equivariant feature map by applying $\mathcal{Q}$, where

$$(\mathcal{Q}f)(x) = \int_{\mathcal{G}} \psi(g^{-1}) f(gx)\, \mathrm{d}\lambda(g).$$

Notice that $\mathcal{O}$ is a special case of $\mathcal{Q}$ corresponding to ψ being the trivial representation. The operator $\mathcal{O}$ can be thought of as performing feature averaging with respect to $\mathcal{G}$. This interpretation is widely adopted, for instance appearing in Lyle et al. (2020).

Function Spaces We now show how to construct the relevant spaces of functions. We present this in an abstract way, but these functions can be interpreted as predictors, feature maps, feature extractors and so on. Let μ be a $\mathcal{G}$-invariant measure on $\mathcal{X}$ and let $\langle a, b \rangle$ for $a, b \in \mathcal{Y}$ be an inner product on $\mathcal{Y} = \mathbb{R}^k$ that is preserved by ψ.[1] By preserved we mean that $\langle \psi(g)a, \psi(g)b \rangle = \langle a, b \rangle, \forall g \in \mathcal{G}, \forall a, b \in \mathcal{Y}$ and any inner product can be transformed to satisfy this property using the Weyl trick $\langle a, b \rangle \mapsto \int_{\mathcal{G}} \langle \psi(g)a, \psi(g)b \rangle \, \mathrm{d}\lambda(g)$. Given two functions $f, h : \mathcal{X} \to \mathcal{Y}$, we define their inner product by

$$\langle f, h \rangle_\mu = \int_{\mathcal{X}} \langle f(x), h(x) \rangle \, \mathrm{d}\mu(x).$$

This inner product can be thought of as comparing the similarity between functions and can used to define a notion of distance with the norm $\|f\|_\mu = \sqrt{\langle f, f \rangle_\mu}$. We then define V as the space of all (measurable) functions $f : \mathcal{X} \to \mathcal{Y}$ such that $\|f\|_\mu < \infty$.[2] Formally, V is a Bochner space.

4.2. Averaging and the Structure of Function Spaces

We have seen how to define orbit averaging operators to produce invariant and equivariant functions as well as how to construct spaces of functions on which these operators can act. The reason for all of this is the following result, which shows that the averaging operators allow us to decompose any function in V into orthogonal $\mathcal{G}$-symmetric and $\mathcal{G}$-anti-symmetric parts. Recall that since $\mathcal{O}$ is just a special case of $\mathcal{Q}$, Lemma 1 applies to both operators.

Lemma 1. Let U be any subspace of V that is closed under $\mathcal{Q}$. Define the subspaces S and A of, respectively, the $\mathcal{G}$-symmetric and $\mathcal{G}$-anti-symmetric functions in U: $S = \{f \in U : f \text{ is } \mathcal{G}\text{-equivariant}\}$ and $A = \{f \in U : \mathcal{Q}f = 0\}$. Then U admits admits an orthogonal decomposition into symmetric and anti-symmetric parts

$$U = S \oplus A.$$

See Appendix B for the proof. The proof consists of establishing the following properties: (A) for any $f \in V$, $\mathcal{Q}f \in V$; (B) any $f \in V$ is $\mathcal{G}$-equivariant if and only if $\mathcal{Q}f = f$; (C) $\mathcal{Q}$ has only two eigenvalues, 1 and 0; and, (D) $\mathcal{Q}$ is self-adjoint with respect to $\langle \cdot, \cdot \rangle_\mu$. The last of these is critical and depends on the $\mathcal{G}$-invariance of μ. There are many tasks for which $\mathcal{G}$-invariance of the input distribution is a natural assumption, for instance in medical imaging (Winkels & Cohen, 2018).

Lemma 1 says that any function $u \in U$ can be written $u = s + a$, where s is $\mathcal{G}$-equivariant, $\mathcal{Q}a = 0$ and $\langle s, a \rangle_\mu = 0$. We refer to s and a as the symmetric and anti-symmetric parts of u. In general this does not imply that a is odd, that it outputs an anti-symmetric matrix or that it is negated by swapping two inputs. These are, however, special cases. If $\mathcal{G} = C_2$ acts by $x \mapsto -x$ then odd functions $f : \mathbb{R} \to \mathbb{R}$ will be anti-symmetric in the sense of this paper. If $\mathcal{G} = C_2$ acts on matrices by $M \mapsto M^\top$ then $f : M \mapsto \frac{1}{2}(M - M^\top)$ is also anti-symmetric. Finally, if $\mathcal{G} = S_n$ and $f : \mathbb{R}^n \to \mathbb{R}$ with $f(x_1, \dots, x_j, x_{j+1}, \dots, x_n) = -f(x_1, \dots, x_{j+1}, x_j, \dots, x_n)$, then f is anti-symmetric in the sense of this paper.

Although it is straightforward to demonstrate and has surely been observed before, we will see in the rest of the paper that the perspective provided by Lemma 1 is highly fruitful. Before that, we conclude this section with an example for intuition.

Example 2. Let V consist of all functions $f : \mathbb{R}^2 \to \mathbb{R}$ such that $\mathbb{E}[f(X)^2] < \infty$ where $X \sim \mathcal{N}(0, I_2)$. Let $\mathcal{G} = \mathrm{SO}(2)$ act by rotation about the origin, with respect to which the normal distribution is invariant. Using Lemma 1 we may write $V = S \oplus A$. Alternatively, consider polar coordinates (r, θ), then for any feature map $f(r, \theta)$ we have $\mathcal{O}f(r, \theta) = \frac{1}{2\pi} \int_0^{2\pi} f(r, \theta) \, \mathrm{d}\theta$. So any $\mathcal{G}$-invariant feature map (i.e. anything in S) depends only on the radial coordinate. Similarly, any h for which $\mathcal{O}h = 0$ must have $\mathcal{O}h(r, \theta) = \frac{1}{2\pi} \int_0^{2\pi} h(r, \theta) \, \mathrm{d}\theta = 0$ for any r, and A consists entirely of such functions. For example, $r^3 \cos\theta \in A$. We then recover $\langle s, h \rangle_\mu = \frac{1}{2\pi} \int_{\mathcal{X}} s(r) h(r, \theta) \mathrm{e}^{-r^2/2} r \, \mathrm{d}r \, \mathrm{d}\theta = 0$ for any $s \in S$ by integrating h over θ. Intuitively, one can think of the functions in S as varying perpendicular to the flow of $\mathcal{G}$ on $\mathcal{X} = \mathbb{R}^2$ and so are preserved by it, while the functions in A average to 0 along this flow, see Fig. 1.

5. Feature Averaging

We remarked earlier that $\mathcal{O}$ can be thought of as performing feature averaging. Before our study of the generalisation of symmetric models, we use this interpretation to derive our first consequence of Lemma 1. We show that feature averaging

[1] It is permissible for inner product itself to depend on the point x at which the feature maps are evaluated. The only requirement is that evaluating the inner product between two fixed vectors is a $\mathcal{G}$-invariant function $\langle a, b \rangle(x) = \langle a, b \rangle(gx), \forall a, b \in \mathbb{R}^k$, $g \in \mathcal{G}$ and $x \in \mathcal{X}$. We believe that this allows our results to extend to the case of the features defined as maps from a manifold to its tangent bundle.

[2] Equality is defined μ-almost-everywhere.

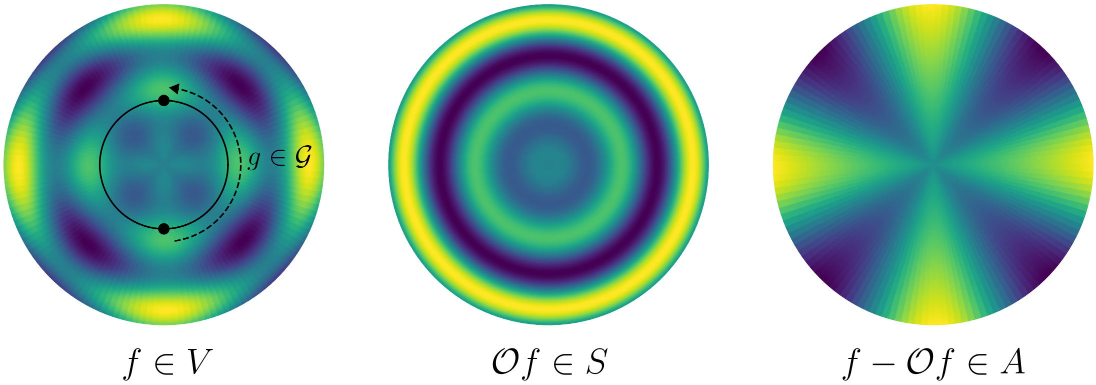

Figure 1. Example of a function decomposition. The figure shows $f(r, \theta) = r\cos(r - 2\theta)\cos(r + 2\theta)$ decomposed into its symmetric and anti-symmetric parts in $V = S \oplus A$ under the natural action of $\mathcal{G} = \mathrm{SO}(2)$ on $\mathbb{R}^2$. See Example 2. Best viewed in colour.

can be viewed as solving a least squares problem in the space of features extractors V. That is, feature averaging sends f to $\bar{f}$, where $\bar{f}$ is the *closest $\mathcal{G}$-invariant feature extractor to f.*

Proposition 3 (Feature Averaging as a Least-Squares Problem)**.** Let V be the space of all normalisable feature extractors as defined above. Define S and A as in Lemma 1. For any $f \in V$, feature averaging with $\mathcal{O}$ maps $f \mapsto \bar{f}$ where $\bar{f}$ is the (μ-a.e.) unique solution to the least-squares problem

$$\bar{f} = \operatorname*{argmin}_{s \in S} \|f - s\|_\mu^2.$$

The proof of Proposition 3 is a straightforward exercise, so we postpone it to Appendix B.2.

Example 4. Consider again the setting of Example 2. For simplicity, let $f(r, \theta) = f_{\text{rad}}(r) f_{\text{ang}}(\theta)$ be separable in polar coordinates. Notice that $\mathcal{O}f = c_f f_{\text{rad}}$ where $c_f = \frac{1}{2\pi}\int_0^{2\pi} f_{\text{ang}}(\theta)\,\mathrm{d}\theta$. Then for any $s \in S$ can calculate:

$$\begin{aligned}
\|f - s\|_\mu^2 &= \frac{1}{2\pi}\int_{\mathcal{X}} (f(r,\theta) - s(r))^2 \mathrm{e}^{-r^2/2} r\,\mathrm{d}r\,\mathrm{d}\theta \\
&= \frac{1}{2\pi}\int_{\mathcal{X}} (f(r,\theta) - c_f f_{\text{rad}}(r))^2 \mathrm{e}^{-r^2/2} r\,\mathrm{d}r\,\mathrm{d}\theta \\
&\quad + \frac{1}{2\pi}\int_{\mathcal{X}} (c_f f_{\text{rad}}(r) - s)^2 \mathrm{e}^{-r^2/2} r\,\mathrm{d}r\,\mathrm{d}\theta
\end{aligned}$$

which is minimised by $s = c_f f_{\text{rad}}$, as predicted.

5.1. Feature Averaging and Generalisation

We end our discussion of feature averaging with an analysis of its impact on generalisation. To do this we consider the reduction in the Rademacher complexity of a feature averaged class.

Rademacher Complexity Let $T = \{t_1, \dots, t_n\}$ be a collection of points of some space $\mathcal{T}$. The empirical Rademacher complexity of a set $\mathcal{F}$ of functions $f : \mathcal{T} \to \mathbb{R}$ evaluated on T is defined by

$$\widehat{\mathfrak{R}}_T(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i f(t_i) \right| \right]$$

where $\sigma_i \sim \mathrm{Unif}\{-1, 1\}$ for $i = 1, \dots, n$ are Rademacher random variables over which the expectation $\mathbb{E}_{\boldsymbol{\sigma}}$ is taken. Let ν be a distribution on $\mathcal{T}$ and take $T \sim \nu^n$, in which case the empirical Rademacher complexity $\widehat{\mathfrak{R}}_T(\mathcal{F})$ is a random quantity. The Rademacher complexity $\mathfrak{R}_n(\mathcal{F})$ is defined by taking an expectation over T: $\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}[\widehat{\mathfrak{R}}_T(\mathcal{F})]$. The Rademacher complexity appears in the study of generalisation in statistical learning, for instance see Wainwright (2019, Theorem 4.10 and Proposition 4.12).

Proposition 5. Let $\mathcal{G}$ be a compact group acting measurably on a set $\mathcal{T}$. Let $\mathcal{F}$ be a class of functions $f : \mathcal{T} \to \mathbb{R}$ and define the symmetric and anti-symmetric classes $\overline{\mathcal{F}} = \{\mathcal{O}f : f \in \mathcal{F}\}$ and $\mathcal{F}^{\perp} = \{f - \mathcal{O}f : f \in \mathcal{F}\}$. Let ν be a distribution over $\mathcal{T}$ that is $\mathcal{G}$-invariant. Then the Rademacher complexity of the feature averaged class satisfies

$$0 \leq \mathfrak{R}_n(\mathcal{F}) - \mathfrak{R}_n(\overline{\mathcal{F}}) \leq \mathfrak{R}_n(\mathcal{F}^{\perp})$$

where the expectations in the definition of $\mathfrak{R}_n$ are taken over $t_i \sim \nu$ i.i.d..

See Appendices B and B.3 for setup, proof and technicalities. Proposition 5 says that the Rademacher complexity is reduced by feature averaging, but not by more than the complexity of the anti-symmetric component of the class. This can be thought of as quantifying the benefit of enforcing invariance by averaging in terms of the extent to which the inductive bias is already present in the function class. Although the form of this result is suggestive, it does not imply a strict benefit. We provide stronger results in the following section.

6. Generalisation Benefit from Symmetric Models

In this section we apply Lemma 1 to derive a strict (i.e. non-zero) generalisation gap between models that have and have not been specified to have the invariance/equivariance that is present in the task. We start with the following general result, which equates the generalisation gap between any predictor and its closest equivariant function to the norm of the anti-symmetric component of the predictor.

Lemma 6. Let $X \sim \mu$ where μ is a $\mathcal{G}$-invariant distribution on $\mathcal{X}$. Let $Y = f^*(X) + \xi \in \mathbb{R}^k$, where ξ is a random element of $\mathbb{R}^k$ that is independent of X with zero mean and finite variance and $f^* : \mathcal{X} \to \mathbb{R}^k$ is $\mathcal{G}$-equivariant. Then, for any $f \in V$, the generalisation gap satisfies

$$\Delta(f, \mathcal{Q}f) \coloneqq \mathbb{E}[\|Y - f(X)\|_2^2] - \mathbb{E}[\|Y - \mathcal{Q}f(X)\|_2^2] = \|f^{\perp}\|_{\mu}^2.$$

Proof. By Lemma 1, we can decompose $f = \bar{f} + f^{\perp}$ into its symmetric $\bar{f} = \mathcal{Q}f$ and anti-symmetric $f^{\perp} = f - \mathcal{Q}f$ parts. Since ξ is independent of X with zero mean and finite variance,

$$\Delta(f, \mathcal{Q}f) = \mathbb{E}[\|Y - f(X)\|_2^2] - \mathbb{E}[\|Y - \bar{f}(X)\|_2^2] = \mathbb{E}[\|f^*(X) - f(X)\|_2^2] - \mathbb{E}[\|f^*(X) - \bar{f}(X)\|_2^2].$$

Using the decomposition of f,

$$\begin{aligned} \mathbb{E}[\|f^*(X) - f(X)\|_2^2] - \mathbb{E}[\|f^*(X) - \bar{f}(X)\|_2^2] &= -2\,\mathbb{E}[\langle f^*(X) - \bar{f}(X), f^{\perp}(X)\rangle] + \mathbb{E}[\|f^{\perp}(X)\|_2^2] \\ &= \|f^{\perp}\|_{\mu}^2. \end{aligned}$$

Here we relied on $\mathbb{E}[\langle f^*(X) - \bar{f}(X), f^{\perp}(X)\rangle] = \langle f^* - \bar{f}, f^{\perp}\rangle_{\mu} = 0$, which follows from f^* being $\mathcal{G}$-equivariant and hence orthogonal to $f^{\perp}$. □

Lemma 6 demonstrates the existence of barrier in the generalisation of any predictor on a problem that has a symmetry. Notice that the barrier turns out to be the measure of how well the predictor encodes the symmetry. Clearly, the only way of overcoming this is to set $f^{\perp} = 0$ (μ-a.e.), which from Lemma 1 equivalent to enforcing $\mathcal{G}$-equivariance in f (μ-a.e.). Lemma 6 therefore provides a *strict generalisation benefit for equivariant predictors*.

In a sense, $\mathcal{Q}f$ is the archetypal equivariant predictor to which f should be compared. A trivial extension to Proposition 3 shows that $\mathcal{Q}f$ is the closest equivariant predictor to f and, more importantly, if h is a $\mathcal{G}$-equivariant predictor with smaller test risk than $\mathcal{Q}f$ then $\Delta(f, h) = \Delta(f, \mathcal{Q}f) + \Delta(\mathcal{Q}f, h) \geq \Delta(f, \mathcal{Q}f)$ which cannot weaken our result.

Later in this section we will use Lemma 6 to explicitly calculate the generalisation gap for invariant/equivariant linear models. We will see that it displays a natural relationship between the number of training examples and the dimension of the space of anti-symmetric models A, which is a property of the group action. Intuitively, the model needs enough examples to learn to be orthogonal to A.

This result also has a useful theoretical implication for test-time data augmentation, which is commonly used to increase test accuracy (Simonyan & Zisserman, 2015; Szegedy et al., 2015; He et al., 2016). Test-time augmentation consists of averaging the output of a learned function f over random transformations of the same input when making predictions at test-time. When the transformations belong to a group $\mathcal{G}$ and are sampled from its Haar measure, test-time averaging can be viewed as a Monte Carlo estimate of $\mathcal{O}f$. Lemma 6 then shows that test-time averaging is beneficial for generalisation when the target function is itself $\mathcal{G}$-invariant, regardless of the learned function f.

6.1. Regression with Invariant Target

Let $\mathcal{X} = \mathbb{R}^d$ with the Euclidean inner product and $\mathcal{Y} = \mathbb{R}$. Consider linear regression with the squared-error loss $\ell(y, y') = (y - y')^2$. Let $\mathcal{G}$ be a compact group that acts on $\mathcal{X}$ via an orthogonal representation $\phi : \mathcal{G} \to O(d)$ and let $X \sim \mu$ where μ is now an arbitrary $\mathcal{G}$-invariant probability distribution on $\mathcal{X}$ with $\Sigma := \mathbb{E}[XX^\top]$ finite and positive definite.[3] We consider linear predictors $h_w : \mathcal{X} \to \mathcal{Y}$ with $h_w(x) = w^\top x$ where $w \in \mathcal{X}$. Define the space of all linear predictors $V_{\text{lin}} = \{h_w : w \in \mathcal{X}\}$ which is a subspace of V. Notice that V_{lin} is closed under $\mathcal{O}$: for any $x \in \mathcal{X}$

$$\mathcal{O}h_w(x) = \int_{\mathcal{G}} h_w(gx)\,\mathrm{d}\lambda(g) = \int_{\mathcal{G}} w^\top \phi(g) x\,\mathrm{d}\lambda(g) = \left(\int_{\mathcal{G}} \phi(g^{-1}) w\,\mathrm{d}\lambda(g)\right)^\top x = h_{\Phi_{\mathcal{G}}(w)}(x)$$

where in the last line we substituted $g \mapsto g^{-1}$ and defined the linear map $\Phi_{\mathcal{G}} : \mathbb{R}^d \to \mathbb{R}^d$ by $\Phi_{\mathcal{G}}(w) = \int_{\mathcal{G}} \phi(g) w\,\mathrm{d}\lambda(g)$.[4] We also have

$$\langle h_a, h_b\rangle_\mu = \int_{\mathcal{X}} a^\top x x^\top b\,\mathrm{d}\mu(x) = a^\top \Sigma b$$

and we denote the induced inner product on $\mathcal{X}$ by $\langle a, b\rangle_\Sigma := a^\top \Sigma b$ and the corresponding norm by $\|\cdot\|_\Sigma$. Since V_{lin} is closed under $\mathcal{O}$ we can apply Lemma 1 to decompose $V_{\text{lin}} = S \oplus A$ where the orthogonality is with respect to $\langle \cdot, \cdot\rangle_\mu$. It follows that we can write any $h_w \in V_{\text{lin}}$ as

$$h_w = \overline{h_w} + h_w^\perp$$

where we have shown that there must exist $\bar{w}, w^\perp \in \mathcal{X}$ with $\langle \bar{w}, w^\perp\rangle_\Sigma = 0$ such that $\overline{h_w} = h_{\bar{w}}$ and $h_w^\perp = h_{w^\perp}$. By choosing a basis for $\mathcal{X}$, there is a natural bijection $\mathcal{X} \to V_{\text{lin}}$ where $w \mapsto h_w$. Using this identification, we abuse notation slightly and write $\mathcal{X} = S \oplus A$ to represent the induced structure on $\mathcal{X}$.

Suppose examples are labelled by a target function $h_\theta \in V_{\text{lin}}$ that is $\mathcal{G}$-invariant. Let $X \sim \mu$ and $Y = \theta^\top X + \xi$ where ξ is independent of X, has mean 0 and finite variance. Recall the definition of the *generalisation gap* between predictors as the difference in their test errors. We study the generalisation gap $\Delta(h_w, h_{\bar{w}})$ between predictors h_w and $h_{\bar{w}}$ defined above. Lemma 6 gives $\Delta(h_w, h_{\bar{w}}) = \|h_{w^\perp}\|_\mu^2 = \|w^\perp\|_\Sigma^2$. In Theorem 7 we calculate this quantity where w is the minimum-norm least-squares estimator and $\bar{w} = \Phi_{\mathcal{G}}(w)$. To the best of our knowledge, this is the first result to specify exactly the generalisation benefit for invariant models.

Theorem 7. Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$ and let $\mathcal{G}$ be a compact group with an orthogonal representation ϕ on $\mathcal{X}$. Let $X \sim \mathcal{N}(0, \sigma_X^2 I)$ and $Y = h_\theta(X) + \xi$ where $h_\theta(x) = \theta^\top x$ is $\mathcal{G}$-invariant with $\theta \in \mathbb{R}^d$ and ξ has mean 0, variance $\sigma_\xi^2 < \infty$ and is independent of X. Let w be the least-squares estimate of θ from i.i.d. examples $\{(X_i, Y_i) : i = 1, \dots, n\}$ and let A be the orthogonal complement of the subspace of $\mathcal{G}$-invariant linear predictors (c.f. Lemma 1).

- If $n > d + 1$ then the generalisation gap is

$$\mathbb{E}[\Delta(h_w, h_{\bar{w}})] = \sigma_\xi^2 \frac{\dim A}{n - d - 1}.$$

- At the interpolation threshold $n \in [d - 1, d + 1]$, if h_w is not $\mathcal{G}$-invariant then the generalisation gap diverges to ∞.
- If $n < d - 1$ the generalisation gap is

$$\mathbb{E}[\Delta(h_w, h_{\bar{w}})] = \dim A \left(\sigma_X^2 \|\theta\|_2^2 \frac{n(d-n)}{d(d-1)(d+2)} + \sigma_\xi^2 \frac{n}{d(d-n-1)}\right).$$

In each case, the generalisation gap has a term of the form $\sigma_\xi^2 r(n, d) \dim A$ that arises due to the noise in the target distribution. In the overparameterised setting $d > n+1$ there is an additional term (the first) that represents the generalisation gap in the noiseless setting $\xi \overset{\text{a.s.}}{=} 0$. This term is the error in the least-squares estimate of θ in the noiseless problem, which of course vanishes in the fully determined case $n > d + 1$. In addition, the divergence at the so called interpolation threshold $n \approx d$ is consistent with the literature on double descent (Hastie et al., 2019).

Notice the central role of $\dim A$ in Theorem 7. This quantity is a property of the group action as it describes the codimension of the set of invariant models. The generalisation gap is then dictated by how ‘significant’ the symmetry is to the problem. Before turning to the proof, we give two examples that represent extremal cases of this ‘significance’.

[3] If Σ is only positive semi-definite then the developments are similar. We assume $\Sigma > 0$ for simplicity.

[4] Since $\mathcal{G}$ is compact it is unimodular and this change of variables is valid, e.g. see Folland (2016, Corollary 2.28) and adjacent results.

Example 8 (Permutations, $\dim A = d-1$)**.** Let S_d act on $\mathcal{X} = \mathbb{R}^d$ by permutation of the coordinates, so $(\phi(\rho)w)_i = w_{\rho(i)}$ for $\rho \in S_d$. Observe that, since the Haar measure λ is uniform on S_d, for any $i = 1, \dots, d$

$$\Phi_{S_d}(w)_i = \frac{1}{d!}\sum_{\rho \in S_d} w_{\rho(i)} = \frac{1}{d}\sum_j w_j$$

so S is the one dimensional subspace $\{t(1,\dots,1)^\top : t \in \mathbb{R}\}$. Since $\mathcal{X} = S \oplus A$ we get $\dim A = d-1$.

Example 9 (Reflection, $\dim A = 1$)**.** Let C_2 be the cyclic group of order 2 and let it act on $\mathcal{X} = \mathbb{R}^d$ by reflection in the first coordinate. A is then the subspace consisting of w such that for any $j = 2, \dots, d$

$$\Phi_{C_2}(w)_j = \frac{1}{|C_2|}\sum_{g \in C_2}(\phi(g)w)_j = w_j = 0$$

since the action fixes all coordinates apart from the first. Hence $A = \{t(1,0,\dots,0)^\top : t \in \mathbb{R}\}$.

Proof of Theorem 7. Note that X is $\mathcal{G}$-invariant for any $\mathcal{G}$ since the representation ϕ is orthogonal. We have seen above that the space of linear maps $V_{\text{lin}} = \{h_w : w \in \mathbb{R}^d\}$ is closed under $\mathcal{O}$, so by Lemma 1 we can write $V_{\text{lin}} = S \oplus A$. Let $\Phi_{\mathcal{G}}{}^{\perp} = I - \Phi_{\mathcal{G}}$, which is the orthogonal projection onto the subspace A. By isotropy of X we have

$$\Delta(h_w, h_{\bar{w}}) = \sigma_X^2 \|w^\perp\|_2^2$$

for any $w \in \mathbb{R}^d$, where $w^\perp = \Phi_{\mathcal{G}}{}^{\perp}(w)$. The proof consists of calculating this quantity in the case that w is the least-squares estimator.

Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{Y} \in \mathbb{R}^n$ correspond to row-stacked training examples drawn i.i.d. as in the statement, so $\boldsymbol{X}_{ij} = (X_i)_j$ and $\boldsymbol{Y}_i = Y_i$. Similarly, set $\boldsymbol{\xi} = \boldsymbol{X}\theta - \boldsymbol{Y}$. The least squares estimate is the minimum norm solution of $\operatorname{argmin}_{u\in\mathbb{R}^d}\|\boldsymbol{Y} - \boldsymbol{X}u\|_2^2$, i.e.

$$w = (\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{X}\theta + (\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{\xi} \tag{1}$$

where $(\cdot)^+$ denotes the Moore-Penrose pseudo-inverse. Define $P_E = (\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{X}$, which is an orthogonal projection onto E, the rank of $\boldsymbol{X}^\top\boldsymbol{X}$ (this can be seen by diagonalising).

We first calculate $\mathbb{E}[\|w^\perp\|_2^2|\boldsymbol{X}]$ where $w^\perp = \Phi_{\mathcal{G}}{}^{\perp}(w)$. The contribution from the first term of Eq. (1) is $\|\Phi_{\mathcal{G}}{}^{\perp}(P_E\theta)\|_2^2$ the cross term vanishes using $\xi \perp\!\!\!\perp X$ and $\mathbb{E}[\xi] = 0$ and the contribution from the second term of Eq. (1) is $\mathbb{E}[\|\Phi_{\mathcal{G}}{}^{\perp}((\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{\xi})\|_2^2|\boldsymbol{X}]$. Notice that $\Phi_{\mathcal{G}}{}^{\perp}$ is just projection matrix and so is idempotent, hence (briefly writing it without the parenthesis to emphasis matrix interpretation)

$$\begin{aligned}\mathbb{E}[\|\Phi_{\mathcal{G}}{}^{\perp}((\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{\xi})\|_2^2|\boldsymbol{X}] &= \mathbb{E}[\operatorname{Tr}(\boldsymbol{\xi}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^+\Phi_{\mathcal{G}}{}^{\perp}(\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{\xi})|\boldsymbol{X}]\\ &= \operatorname{Tr}(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^+\Phi_{\mathcal{G}}{}^{\perp}(\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\,\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top])\\ &= \sigma_\xi^2\operatorname{Tr}(\Phi_{\mathcal{G}}{}^{\perp}(\boldsymbol{X}^\top\boldsymbol{X})^+).\end{aligned}$$

We have obtained

$$\mathbb{E}[\|w^\perp\|_2^2|\boldsymbol{X}] = \|\Phi_{\mathcal{G}}{}^{\perp}(P_E\theta)\|_2^2 + \sigma_\xi^2\operatorname{Tr}(\Phi_{\mathcal{G}}{}^{\perp}((\boldsymbol{X}^\top\boldsymbol{X})^+))$$

and conclude by taking expectations, treating each term separately.

First Term If $n \geq d$ then $\dim E = d$ with probability 1, so the first term vanishes almost surely. We treat the $n < d$ case using Einstein notation, in which repeated indices are implicitly summed over. In components, recalling that $\Phi_{\mathcal{G}}{}^{\perp}$ is a matrix,

$$\mathbb{E}[\|\Phi_{\mathcal{G}}{}^{\perp}(P_E\theta)\|_2^2] = \Phi_{\mathcal{G}}{}^{\perp}_{fa}\Phi_{\mathcal{G}}{}^{\perp}_{fc}\,\mathbb{E}[P_E \otimes P_E]_{abce}\theta_b\theta_e$$

and applying Lemma 22 we get

$$\begin{aligned}\mathbb{E}[\|\Phi_{\mathcal{G}}{}^{\perp}(P_E\theta)\|_2^2] &= \frac{n(d-n)}{d(d-1)(d+2)}\left(\Phi_{\mathcal{G}}{}^{\perp}_{fa}\Phi_{\mathcal{G}}{}^{\perp}_{fa}\theta_b\theta_b + \Phi_{\mathcal{G}}{}^{\perp}_{fa}\Phi_{\mathcal{G}}{}^{\perp}_{fb}\theta_b\theta_a\right)\\ &\quad + \frac{n(d-n)+n(n-1)(d+2)}{d(d-1)(d+2)}\Phi_{\mathcal{G}}{}^{\perp}_{fa}\Phi_{\mathcal{G}}{}^{\perp}_{fc}\theta_a\theta_c\\ &= \|\theta\|_2^2\dim A\frac{n(d-n)}{d(d-1)(d+2)}\end{aligned}$$

where we have used that ${\Phi_{\mathcal{G}}}^{\perp}(\theta) = 0$ and $\|{\Phi_{\mathcal{G}}}^{\perp}\|_{\mathrm{F}}^2 = \dim A$.

Second Term By linearity,

$$\mathbb{E}[\operatorname{Tr}({\Phi_{\mathcal{G}}}^{\perp}((\boldsymbol{X}^\top \boldsymbol{X})^+))] = \operatorname{Tr}({\Phi_{\mathcal{G}}}^{\perp}(\mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^+])).$$

Then Lemmas 18 and 20 give $\mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^+] = \sigma_X^{-2} r(n,d) I_d$ where

$$r(n,d) = \begin{cases} \frac{n}{d(d-n-1)} & n < d-1 \\ (n-d-1)^{-1} & n > d+1 \\ \infty & \text{otherwise} \end{cases}.$$

When $n \in [d-1, d+1]$ it is well known that the expectation diverges, see Appendix A. Hence

$$\mathbb{E}[\operatorname{Tr}({\Phi_{\mathcal{G}}}^{\perp}((\boldsymbol{X}^\top \boldsymbol{X})^+))] = \sigma_X^{-2} r(n,d) \dim A.$$

□

6.2. Regression with Equivariant Target

One can apply the same construction to equivariant models. Assume the same setup, but now let $\mathcal{Y} = \mathbb{R}^k$ with the Euclidean inner product and let the space of predictors be $W_{\text{lin}} = \{f_W : \mathbb{R}^d \to \mathbb{R}^k,\ f_W(x) = W^\top x : W \in \mathbb{R}^{d\times k}\}$. We consider linear regression with the squared-error loss $\ell(y, y') = \|y - y'\|_2^2$. Let ψ be an orthogonal representation of $\mathcal{G}$ on $\mathcal{Y}$. We define the linear map, which we call the intertwining average, $\Psi_{\mathcal{G}} : \mathbb{R}^{d\times k} \to \mathbb{R}^{d\times k}$ by[5]

$$\Psi_{\mathcal{G}}(W) = \int_{\mathcal{G}} \phi(g) W \psi(g^{-1}) \, \mathrm{d}\lambda(g).$$

Similarly, define the intertwining complement as ${\Psi_{\mathcal{G}}}^{\perp} : \mathbb{R}^{d\times k} \to \mathbb{R}^{d\times k}$ by ${\Psi_{\mathcal{G}}}^{\perp}(W) = W - \Psi_{\mathcal{G}}(W)$. We establish the following results, which are in fact generalisations of the invariant case. In the proofs we will leverage the expression of $\Psi_{\mathcal{G}}$ as a 4-tensor with components ${\Psi_{\mathcal{G}}}_{abce} = \int_{\mathcal{G}} \phi(g)_{ac} \psi(g)_{be} \, \mathrm{d}\lambda(g)$ where $a, c = 1, \dots d$ and $b, e = 1, \dots, k$.[6]

Proposition 10. For any $f_W \in W_{\text{lin}}$, $\mathcal{Q} f_W = f_{\Psi_{\mathcal{G}}(W)}$ and hence W_{lin} is closed under $\mathcal{Q}$.

Proof. Let $f_W(x) = W^\top x$ with $W \in \mathbb{R}^{d\times k}$, then using orthogonality and unimodularity

$$\mathcal{Q} f_W(x) = \int_{\mathcal{G}} \psi(g^{-1}) W^\top \phi(g) x \, \mathrm{d}\lambda(g) = \left(\int_{\mathcal{G}} \phi(g) W \psi(g^{-1}) \, \mathrm{d}\lambda(g) \right)^\top x = \Psi_{\mathcal{G}}(W)^\top x.$$

□

Proposition 11. The inner product on W_{lin} satisfies, for any $f_{W_1}, f_{W_2} \in W_{\text{lin}}$,

$$\langle f_{W_1}, f_{W_2} \rangle_\mu = \operatorname{Tr}(W_1^\top \Sigma W_2)$$

where $\Sigma = \mathbb{E}[XX^\top]$ and $X \sim \mu$.

Proof.

$$\begin{aligned} \langle f_{W_1}, f_{W_2} \rangle_\mu &= \int_{\mathcal{X}} \left(W_1^\top x\right)^\top W_2^\top x \, \mathrm{d}\mu(x) \\ &= \int_{\mathcal{X}} x^\top W_1 W_2^\top x \, \mathrm{d}\mu(x) \\ &= \int_{\mathcal{X}} \operatorname{Tr}(x^\top W_1 W_2^\top x) \, \mathrm{d}\mu(x) \\ &= \operatorname{Tr}(W_1^\top \Sigma W_2) \end{aligned}$$

□

[5] The reader may have noticed that we define $\Psi_{\mathcal{G}}$ backwards, in the sense that its image contains maps that are equivariant in the direction $\psi \to \phi$. This is because of the transpose in the linear model, which is there for consistency with the $k = 1$ invariance case. This choice is arbitrary and gives no loss in generality.

[6] If necessary, the reader can see Appendix A.3 for a derivation.

Proposition 10 allows us to apply Lemma 1 to write $W_{\text{lin}} = S \oplus A$, so for any $f_W \in W_{\text{lin}}$ there exists $\overline{f_W} \in S$ and $f_W^\perp \in A$ with $\langle \overline{f_W}, f_W^\perp \rangle_\mu = 0$. The corresponding parameters $\overline{W} = \Psi_{\mathcal{G}}(W)$ and $W^\perp = \Psi_{\mathcal{G}}{}^\perp(W)$ must therefore satisfy $\operatorname{Tr}(\overline{W}^\top \Sigma W^\perp) = 0$, with Σ defined as in Proposition 11. Repeating our abuse of notation, we identify $\mathbb{R}^{d\times k} = S \oplus A$ with $S = \Psi_{\mathcal{G}}(\mathbb{R}^{d\times k})$ and A its orthogonal complement with respect to the induced inner product.

Proposition 12. Let $X \sim \mu$ and let ξ a random element of $\mathbb{R}^k$ that is independent of X with $\mathbb{E}[\xi] = 0$ and finite variance. Set $Y = f_\Theta(X) + \xi$ where f_Θ is $\mathcal{G}$-equivariant. For any $f_W \in W_{\text{lin}}$, the generalisation gap satisfies

$$\Delta(f_W, f_{\overline{W}}) \coloneqq \mathbb{E}[\|Y - f_W(X)\|_2^2] - \mathbb{E}[\|Y - f_{\overline{W}}(X)\|_2^2] = \|\Sigma^{1/2} W^\perp\|_{\text{F}}^2$$

where $\overline{W} = \Psi_{\mathcal{G}}(W)$, $W^\perp = \Psi_{\mathcal{G}}{}^\perp(W)$ and $\Sigma = \mathbb{E}[XX^\top]$.

Proof. Recall that $W = \overline{W} + W^\perp$ and that these satisfy $\operatorname{Tr}(\overline{W}\Sigma W^\perp) = 0$ from the above. Then, using Lemma 6 and Proposition 11,

$$\Delta(f_W, f_{\overline{W}}) = \|f_{W^\perp}\|_\mu^2 = \operatorname{Tr}((W^\perp)^\top \Sigma W^\perp) = \|\Sigma^{1/2} W^\perp\|_{\text{F}}^2$$

□

Having followed the same path as the previous section, we provide a characterisation of the generalisation benefit of equivariance. In the same fashion, we compare the least-squares estimate W with its equivariant version $\overline{W} = \Psi_{\mathcal{G}}(W)$. As we explained at the beginning of the section, the choice of $\overline{W} = \Psi_{\mathcal{G}}(W)$ is natural and costs us nothing.

Theorem 13. Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}^k$ and let $\mathcal{G}$ be a compact group with orthogonal representations ϕ on $\mathcal{X}$ and ψ on $\mathcal{Y}$. Let $X \sim \mathcal{N}(0, \sigma_X^2 I_d)$ and $Y = h_\Theta(X) + \xi$ where $h_\Theta(x) = \Theta^\top x$ is $\mathcal{G}$-equivariant and $\Theta \in \mathbb{R}^{d\times k}$. Assume ξ is a random element of $\mathbb{R}^k$, independent of X, with mean 0 and $\mathbb{E}[\xi\xi^\top] = \sigma_\xi^2 I_k < \infty$. Let W be the least-squares estimate of Θ from n i.i.d. examples $\{(X_i, Y_i) : i = 1, \dots, n\}$ and let $(\chi_\psi|\chi_\phi) = \int_{\mathcal{G}} \chi_\psi(g)\chi_\phi(g)\,\mathrm{d}\lambda(g)$ denote the scalar product of the characters of the representations of $\mathcal{G}$.

- If $n > d + 1$ the generalisation gap is

$$\mathbb{E}[\Delta(f_W, f_{\overline{W}})] = \sigma_\xi^2 \frac{dk - (\chi_\psi|\chi_\phi)}{n - d - 1}.$$

- At the interpolation threshold $n \in [d-1, d+1]$, if f_W is not $\mathcal{G}$-equivariant then the generalisation gap diverges to ∞.
- If $n < d - 1$ then the generalisation gap is

$$\mathbb{E}[\Delta(f_W, f_{\overline{W}})] = \sigma_X^2 \frac{n(d-n)}{d(d-1)(d+2)} \left((d+1)\|\Theta\|_{\text{F}}^2 - \operatorname{Tr}(J_{\mathcal{G}}\Theta^\top\Theta)\right) + \sigma_\xi^2 \frac{n(dk - (\chi_\psi|\chi_\phi))}{d(d-n-1)}$$

  where each term is non-negative and $J_{\mathcal{G}} \in \mathbb{R}^{k\times k}$ is given by

$$J_{\mathcal{G}} = \int_{\mathcal{G}} (\chi_\phi(g)\psi(g) + \psi(g^2))\,\mathrm{d}\lambda(g).$$

The proof of Theorem 13 is longer than for Theorem 7 but follows the same scheme, so we defer it to Appendix B.4.

Theorem 13 is a direct generalisation of Theorem 7. As we remarked in the introduction, $dk - (\chi_\psi|\chi_\phi)$ plays the role of $\dim A$ in Theorem 7 and is a measure of the significance of the symmetry to the problem. The dimension of W_{lin} is dk, while $(\chi_\psi|\chi_\phi)$ is the dimension of the space of equivariant maps. In our notation $(\chi_\psi|\chi_\phi) = \dim S$.

Just as with Theorem 7, there is an additional term (the first) in the overparameterised case $d > n + 1$ that represents the estimation in a noiseless setting $\xi \overset{\text{a.s.}}{=} 0$. Notice that if $k = 1$ and ψ is trivial we find

$$J_{\mathcal{G}} = \int_{\mathcal{G}} \chi_\phi(g)\,\mathrm{d}\lambda(g) + 1 = (\chi_\phi|1) + 1 = \dim S + 1$$

which confirms that Theorem 13 reduces exactly to Theorem 7.

Interestingly, the first term in the $d > n + 1$ case can be made independent of ψ, since the equivariance of h_Θ implies

$$\operatorname{Tr}(J_{\mathcal{G}}\Theta^\top\Theta) = \operatorname{Tr}(\Theta^\top J_\phi \Theta)$$

where

$$J_\phi = \int_{\mathcal{G}} (\chi_\phi(g)\phi(g) + \phi(g^2))\,\mathrm{d}\lambda(g).$$

Finally, we remark that Theorem 13 is possible for more general probability distributions on X. For instance, it sufficient that the distribution is absolutely continuous with respect to the Lebesgue measure, has finite variance and is $O(d)$ invariant. The final condition implies the existence of a scalar r_n such that $\mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{X})^+] = r_n I_d$ where $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ are the row-stacked training inputs as defined in the proof.

7. Neural Networks

In this section we discuss how the insights of this paper apply to neural networks and raise some open questions for future work. Let $F : \mathbb{R}^d \to \mathbb{R}^k$ be a feedforward neural network with L layers, layer widths κ_i $i = 1, \dots, L$ and weights $W^i \in \mathbb{R}^{\kappa_i \times \kappa_{i+1}}$ for $i = 1, \dots, L$ where $\kappa_1 = d$ and $\kappa_{L+1} = k$. We will assume F has the form

$$F(x) = W^L\sigma(W^{L-1}\sigma(\dots\sigma(W^1x)\dots)) \tag{2}$$

where σ is an element-wise non-linearity.

7.1. Invariant and Equivariant Networks

The standard method for engineering neural networks to be invariant/equivariant to the action of a finite group on the inputs is weight tying. This method has been around for a while (Wood & Shawe-Taylor, 1996) but has come to recent attention via Ravanbakhsh et al. (2017). We will briefly describe this approach, its connections to Theorems 7 and 13 and how the ideas of this paper can be used to find new algorithms for both enforced and learned invariance/equivariance. We leave analyses of these suggested approaches to future work.

The methods of Wood & Shawe-Taylor (1996); Ravanbakhsh et al. (2017) can be described as follows. Let $\mathcal{G}_{\text{fin}}$ be a finite group. For each $i = 2, \dots, L+1$, the user chooses a matrix representation $\psi_i : \mathcal{G}_{\text{fin}} \to \mathrm{GL}_{k_i}(\mathbb{R})$ of $\mathcal{G}_{\text{fin}}$ that acts on the inputs for each layer $i = 2, \dots, L$ and on the outputs of the network when $i = L + 1$.[7] For $i = 2, \dots, L$, these representations must be chosen such that they commute with the activation function

$$\sigma(\psi_i(g)\cdot) = \psi_i(g)\sigma(\cdot) \tag{3}$$

$\forall g \in \mathcal{G}_{\text{fin}}$.[8] One then chooses weights for the network such that at each layer and $\forall g \in \mathcal{G}_{\text{fin}}$

$$W^i\psi_i(g) = \psi_{i+1}(g)W^i. \tag{4}$$

By induction on the layers, satisfying Eqs. (3) and (4) ensures that the network is $\mathcal{G}_{\text{fin}}$-equivariant. Invariance occurs when ψ_{L+1} is the trivial representation.

The condition in Eq. (4) can be phrased as saying that that W^i belongs to the space of *intertwiners* of the representations ψ_i and ψ_{i+1}. By denoting the space of all weight matrices in layer i as $U = \mathbb{R}^{\kappa_i\times\kappa_{i+1}}$, the space of intertwiners is immediately recognisable as $S = \Psi_{\mathcal{G}}(U)$ from Lemma 1.

Typically, the practitioner will hand-engineer the structure of weight matrices to belong to the correct intertwiner space. In the following sections we will propose alternative procedures that build naturally on the ideas of this paper. Moreover, as a benefit of our framework, these new approaches extend weight-tying to any compact group that admits a finite-dimensional, real representation. We end this section with a bound on the sample complexity of invariant networks, which follows from Bartlett et al. (2019). Similar results are possible for different activation functions.

Lemma 14. Let $\mathcal{G}$ be a compact group with layer-wise representations as described. Let $F : \mathbb{R}^d \to \mathbb{R}$ be a $\mathcal{G}$-invariant neural network with ReLU activation and weights that intertwine the representations. Let $\mathcal{H}$ be the class of all functions

[7] ψ_1 is the representation on the inputs, which we consider as an aspect of the task and not a design choice.

[8] This condition is somewhat restrictive, but note that a permutation representation will commute with any element-wise non-linearity.

realisable by this network. Then

$$\mathrm{VC}(\mathcal{H}) \leq L + \frac{1}{2}\alpha(F)L(L+1) \max_{1\leq i\leq L} (\chi_i|\chi_{i+1})$$

where $\alpha(F) = \log_2\left(4\mathrm{e}\log_2\left(\sum_{i=1}^{L} 2\mathrm{e}i\kappa_i\right)\sum_{i=1}^{L} i\kappa_i\right)$.

Proof. For a ReLU network t_i independent parameters at each layer we have $\mathrm{VC}(\mathcal{H}) \leq L + \alpha(F)\sum_{i=1}^{L}(L-i+1)t_i$, which follows by application of Bartlett et al. (2019, Theorem 7). The proof of Bartlett et al. (2019, Theorem 7) depends on the representation of the network in terms of piecewise polynomials of bounded degree. Observe that since weights are tied only *within* layers (so weights in different layers can vary independently) and the activation is ReLU, there is no increase in the degree of said polynomials from weight-tying and the proof given in Bartlett et al. (2019) applies in our case. The condition Eq. (4) insists that the weight matrix W^i belongs to the intertwiner space $\mathrm{Hom}_{\mathcal{G}}(\mathbb{R}^{\kappa_{i-1}\times\kappa_i}, \mathbb{R}^{\kappa_i\times\kappa_{i+1}})$. The number of independent parameters at each layer is at most the dimension of this space. The conclusion follows by simple algebra and the relation of the dimension to the characters as given above. □

Example 15 (Permutation invariant networks)**.** Permutation invariant networks (and permutation invariant functions more generally) are studied in Wood & Shawe-Taylor (1996); Zaheer et al. (2017); Bloem-Reddy & Teh (2020) and many other works, see references therein. In particular, multiple authors have given the form of a permutation equivariant weight matrix as $W = \lambda I + \gamma\mathbf{1}\mathbf{1}^\top$ for scalars $\alpha, \beta \in \mathbb{R}$ and with $\mathbf{1} = (1, \dots, 1)^\top$. Consider an L-layer ReLU network with, for simplicity, widths $\kappa_i = d\ \forall i$. Let $\mathcal{H}$ be the class of all functions realisable by this network, then $\mathrm{VC}(\mathcal{H}) = O(L^2\log(Ld\log(Ld)))$.

7.2. Projected Gradients

As we have seen, provided that the activation function satisfies Eq. (3), specifying the weight matrices to intertwine between layer-wise representations is sufficient to ensure equivariance in a neural network. We have also seen from Section 6.2 that it is possible to project any weight matrix into an intertwiner space using $\Psi_{\mathcal{G}}$. For each layer l of the network we have a linear map $\Psi_{\mathcal{G}}{}^l$, which is a 4-tensor with components $\Psi_{\mathcal{G}}{}^l_{abce} = \int_{\mathcal{G}} \psi_{l+1}(g)_{ac}\psi_l(g)_{be}\,\mathrm{d}\lambda(g)$.

The tensors $\{\Psi_{\mathcal{G}}{}^l : l = 1, \dots, L\}$ depend only on the representations and so can be computed before training. One can therefore obtain invariant/equivariant networks by a form of projected gradient descent. Explicitly, with loss ℓ and learning rate η, the update rule for the l^{th} layer is

$$\begin{aligned}\widetilde{W}^l(t+1) &= W^l(t) - \eta\nabla_{W^l}\ell(W^1(t), \dots, W^L(t))\\ W^l(t+1) &= \Psi_{\mathcal{G}}{}^l\left(\widetilde{W}^l(t+1)\right).\end{aligned}$$

If Eq. (3) holds the network will be exactly invariant/equivariant after any iteration.

7.3. Regularisation for Equivariance

We have seen from Lemma 1 and Section 6.2 that any weight matrix can be written $W = \overline{W} + W^\perp$ where $\overline{W} = \Psi_{\mathcal{G}}(W)$ belongs to an intertwiner space (so is equivariant) and $W^\perp = \Psi_{\mathcal{G}}{}^\perp(W)$ belongs to an orthogonal space that parametrises the anti-symmetric linear maps. This suggests a method of learned invariance/equivariance by using a regularisation term of the form $\sum_{l=1}^{L}\|\Psi_{\mathcal{G}}{}^{l\perp}(W^l)\|_{\mathrm{F}}^2$. Where the 4-tensor $\Psi_{\mathcal{G}}{}^{l\perp}$ has components $\Psi_{\mathcal{G}}{}^{l\perp}_{abce} = \delta_{ac}\delta_{be} - \Psi_{\mathcal{G}}{}^l_{abce}$ and can be computed before training. If $\Psi_{\mathcal{G}}{}^{l\perp}(W^l) = 0$ for $l = 1, \dots, L$ and the activation function satisfies Eq. (3), then the resulting network will be exactly $\mathcal{G}$-invariant/equivariant. This method could also allow for learned/approximate invariance. Indeed, Proposition 16 suggests $\|\Psi_{\mathcal{G}}{}^{l\perp}(W)\|_{\mathrm{F}}^2$ as a measure of the layer-wise invariance/equivariance of the network.

Proposition 16. Let $\mathcal{G}$ be a compact group. Let $f_W : \mathbb{R}^d \to \mathbb{R}^k$ with $f_W(x) = \sigma(Wx)$ be a single neural network layer with C-Lipschitz, element-wise activation σ. Let $\phi : \mathcal{G} \to O(d)$ and $\psi : \mathcal{G} \to O(k)$ be orthogonal representations of $\mathcal{G}$ on the input and output spaces respectively and assume that ψ commutes with σ as in Eq. (3). Let $X \in \mathbb{R}^d$, $X \sim \mathcal{N}(0, I_d)$. We can consider the network as belonging to V from Section 4 with $\mu = \mathcal{N}(0, I_d)$. Write $V = S \oplus A$, where S contains the equivariant functions in V, then

$$\inf_{s\in S}\mathbb{E}[\|f_W(X) - s(X)\|_2^2] \leq 2C^2\|W^\perp\|_{\mathrm{F}}^2.$$

Proof. First note that the infimum clearly exists, since the left hand side vanishes when W intertwines ϕ and ψ. Recognise that in the notation of Section 4 we can write $\|a-b\|_\mu^2 = \mathbb{E}[\|a(X)-b(X)\|_2^2]$. By applying the proof of Proposition 3 to $\mathcal{Q}$ we get $\inf_{s\in S}\mathbb{E}[\|f_W(X)-s(X)\|^2] = \inf_{s\in S}\|f_W - s\|_\mu^2 = \|f_W - \mathcal{Q}f_W\|_\mu^2$. Then recalling the definition of $\mathcal{Q}$ we have

$$\begin{aligned}
\|\mathcal{Q}f_W(x) - f_W(x)\|_2^2 &= \left\|\int_{\mathcal{G}} \psi(g^{-1})f_W(\phi(g)x) \; - f_W(x)\right\|_2^2 \mathrm{d}\lambda(g)\\
&\le \int_{\mathcal{G}} \|\psi(g^{-1})f_W(\phi(g)x) \; - f_W(x)\|_2^2 \,\mathrm{d}\lambda(g)\\
&= \int_{\mathcal{G}} \|\psi(g^{-1})\sigma(W\phi(g)x) \; - \sigma(Wx)\|_2^2 \,\mathrm{d}\lambda(g)\\
&= \int_{\mathcal{G}} \|\sigma(\overline{W}x + \psi(g^{-1})W^\perp\phi(g)x) \; - \sigma(Wx)\|_2^2 \,\mathrm{d}\lambda(g)\\
&\le C^2\int_{\mathcal{G}} \|\overline{W}x + \psi(g^{-1})W^\perp\phi(g)x \; - Wx\|_2^2 \,\mathrm{d}\lambda(g)\\
&= C^2\int_{\mathcal{G}} \|(\psi(g^{-1})W^\perp\phi(g) - W^\perp)x\|_2^2 \,\mathrm{d}\lambda(g)
\end{aligned}$$

Then by an application of Fubini's theorem and the covariance of X we get

$$\begin{aligned}
&\mathbb{E}[\|\mathcal{Q}f_W(X) - f_W(X)\|_2^2]\\
&\le C^2\int_{\mathcal{G}} \operatorname{Tr}\left((\psi(g^{-1})W^\perp\phi(g) - W^\perp)^\top\,\mathbb{E}[XX^\top](\psi(g^{-1})W^\perp\phi(g) - W^\perp)\right)\mathrm{d}\lambda(g)\\
&= C^2\|W^\perp\|_{\mathrm{F}}^2 + C^2\int_{\mathcal{G}}\|\psi(g^{-1})W^\perp\phi(g)\|_{\mathrm{F}}^2\,\mathrm{d}\lambda(g)
\end{aligned}$$

and then the argument of the integral can be analysed as

$$\begin{aligned}
\|\psi(g^{-1})W^\perp\phi(g)\|_{\mathrm{F}}^2 &= \operatorname{Tr}\left((\psi(g^{-1})W^\perp\phi(g))^\top\psi(g^{-1})W^\perp\phi(g)\right)\\
&= \operatorname{Tr}\left(\phi(g^{-1})(W^\perp)^\top\psi(g)\psi(g^{-1})W^\perp\phi(g)\right)\\
&= \operatorname{Tr}\left((W^\perp)^\top W^\perp\right).
\end{aligned}$$

The proof is complete. □

Proposition 16 shows that the distance between the outputs of a single layer neural network and its closest equivariant function is bounded by the norm of the anti-symmetric component of the weights $W^\perp$. This quantity can be interpreted as a measure of the equivariance of the layer and regularising $\|W^\perp\|_{\mathrm{F}}$ will encourage the network to become (approximately) equivariant. It is easy to generalise Proposition 16 so that X follows any $\mathcal{G}$-invariant distribution with finite second moment.

7.4. Open Questions

Equivariant Convolutions There has been much work on engineering convolutional layers to be group equivariant, for instance Cohen & Welling (2016); Cohen et al. (2018); Kondor & Trivedi (2018); Cohen et al. (2019). The convolution is a linear operator parameterised by the kernel. This suggests that it may be possible to analyse the generalisation properties of group equivariant convolutions in the framework of Lemma 1, similar to Section 6.

Invariant/Equivariant Networks We have discussed enforcing invariance/equivariance in a neural network $F_{(W^1,\dots,W^L)}$ (with the dependence on the weights now explicit) by restricting weight matrices to intertwine between representations at each layer. We ask: is this the best way to encode symmetry? Mathematically, let $X\sim\mu$ with $\mathcal{G}$-invariant μ and embed the functions realised by the network in $V = S\oplus A$. Given an invariant/equivariant target $s\in S$, must the best approximating neural network be layer-wise invariant/equivariant? That is, are there $s\in S$ such that the following holds

$$\inf_{\mathcal{W}}\mathbb{E}[\|F_{(W^1,\dots,W^L)}(X) - s(X)\|_2^2] < \inf_{\mathcal{U}}\mathbb{E}[\|F_{(U^1,\dots,U^L)}(X) - s(X)\|_2^2],$$

where $\mathcal{W} = \{W^l \in \mathbb{R}^{\kappa_l \times \kappa_{l+1}} : l = 1, \dots, L\}$ is the set of all possible weight matrices and $\mathcal{U} = \{U^l \in {\Psi_{\mathcal{G}}}^l(\mathbb{R}^{\kappa_l \times \kappa_{l+1}}) : l = 1, \dots, L\}$ is the set of all weight matrices restricted to be intertwiners? A resolution to this might shed light on new ways of encoding symmetry in neural networks.

Acknowledgements

We would like to thank Varun Kanade and Yee Whye Teh for advice and support throughout this project. We also thank Will Sawin for providing Lemma 22, Michael Hutchinson for instructive feedback on an earlier draft and Bobby He for a helpful discussion on inverse Wishart matrices. BE receives support from the UK EPSRC CDT in Autonomous Intelligent Machines and Systems (grant reference EP/L015897/1). SZ receives support from the Aker Scholarship.

A. Useful Facts

A.1. Inverses

Lemma 17. Let $D \in \mathbb{R}^{d\times d}$ be orthogonal and let $B \in \mathbb{R}^{d\times d}$ be any symmetric matrix, then

$$(DBD^\top)^+ = DB^+D^\top.$$

Proof. Set $X = DB^+D^\top$ and $A = DBD^\top$. It suffices to check that A and X satisfy the Penrose equations, the solution of which is unique (Penrose, 1955), namely: 1. $AXA = A$, 2. $XAX = X$, 3. $(AX)^\top = AX$ and 4. $(XA)^\top = XA$. It is straightforward to check that this is the case. □

Lemma 18 ((Gupta, 1968))**.** Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ have i.i.d. $\mathcal{N}(0, 1)$ elements with $n > d + 1$. Then

$$\mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^+] = \frac{1}{n-d-1} I.$$

Remark 19. It is well known that the expectation in Lemma 18 diverges for $d \leq n \leq d+1$. To see this, first notice that since the normal distribution is $O(d)$ invariant $R\,\mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^+]R^\top = \mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^+]$ for any $R \in O(d)$. Hence $\mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^+]$ is a scalar multiple of the identity: it is symmetric so diagonalisable, hence diagonal in every basis by the invariance, then permutation matrices can be used to show the diagonals are all equal. It remains to consider the eigenvalues. The eigenvalues $\lambda_1, \dots, \lambda_d$ of $\boldsymbol{X}^\top \boldsymbol{X}$ have joint density (w.r.t. Lebesgue) that is proportional to

$$\exp\left(-\frac{1}{2}\sum_{i=1}^d \lambda_i\right) \prod_{i=1}^d \lambda_i^{(n-d-1)/2} \prod_{i<j}^d |\lambda_i - \lambda_j|$$

when $n \geq d$ and 0 otherwise, e.g. see Muirhead (2009, Corollary 3.2.19). We need to calculate the mean of $1/\lambda$ with respect to this density, which diverges unless $n \geq d+2$. Taking the mean of λ_k, there is a term from the Vandermonde product that does not contain λ_k, so the integrand in the expectation goes like $\sqrt{\lambda_k^{n-d-3}}$ as $\lambda_k \to 0$.

Lemma 20 ((Cook et al., 2011, Theorem 2.1))**.** Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ have i.i.d. $\mathcal{N}(0, 1)$ elements with $n < d - 1$. Then

$$\mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^+] = \frac{n}{d(d-n-1)} I.$$

Remark 21. The statement of Lemma 20 in Cook et al. (2011, Theorem 2.1) gives the condition $n < d - 3$, but this is not necessary for the first moment. This is easily seen by examining the proof of Cook et al. (2011, Theorem 2.1). In addition, the proof uses a transformation followed by an application of Lemma 18 with the roles of n and d switched. It follows that the expectation diverges when $d \geq n \geq d - 1$.

A.2. Projections

Lemma 22. Let $E \sim \text{Unif}\, \mathbb{G}_n(\mathbb{R}^d)$ where $0 < n < d$ and let P_E be the orthogonal projection onto E, then in components

$$\mathbb{E}[P_E \otimes P_E]_{abce} = \frac{n(d-n)}{d(d-1)(d+2)}(\delta_{ab}\delta_{ce} + \delta_{ac}\delta_{be} + \delta_{ae}\delta_{bc}) + \frac{n(n-1)}{d(d-1)}\delta_{ab}\delta_{ce}$$

Proof. We use the Einstein convention of implicitly summing over repeated indices. The distribution of E is orthogonally invariant, so $\mathbb{E}[P_E \otimes P_E]$ is isotropic. Thus, $\mathbb{E}[P_E \otimes P_E]$ must have components

$$\Gamma_{abce} \coloneqq \mathbb{E}[P_E \otimes P_E]_{abce} = \alpha\delta_{ab}\delta_{ce} + \beta\delta_{ac}\delta_{be} + \gamma\delta_{ae}\delta_{bc}$$

e.g. see Hodge (1961). Contracting indices gives

$$\begin{aligned} n^2 &= \mathbb{E}[\mathrm{Tr}(P_E)^2] = \Gamma_{aabb} = d^2\alpha + d\beta + d\gamma \\ n &= \mathbb{E}[\mathrm{Tr}(P_E^\top P_E)] = \Gamma_{abab} = d\alpha + d^2\beta + d\gamma \\ n &= \mathbb{E}[\mathrm{Tr}(P_E^2)] = \Gamma_{abba} = d\alpha + d\beta + d^2\gamma \end{aligned}$$

from which one finds

$$\begin{aligned} \beta &= \frac{n(d-n)}{d(d-1)(d+2)} \\ \alpha &= \beta + \frac{n(n-1)}{d(d-1)} \\ \gamma &= \beta. \end{aligned}$$

□

A.3. Component representation of $\Psi_{\mathcal{G}}$

Using the Einstein convention of implicitly summing over repeated indices, one can write

$$\begin{aligned} \Psi_{\mathcal{G}}(W)_{ab} &= \int_{\mathcal{G}} \phi(g)_{ac} W_{ce} \psi(g^{-1})_{eb} \, \mathrm{d}\lambda(g) \\ &= \int_{\mathcal{G}} \phi(g)_{ac} W_{ce} \psi(g)_{be} \, \mathrm{d}\lambda(g) \quad \text{representation is orthogonal} \\ &= \left(\int_{\mathcal{G}} \phi(g)_{ac} \psi(g)_{be} \, \mathrm{d}\lambda(g) \right) W_{ce} \quad \text{components are scalars} \\ &= \Psi_{\mathcal{G}\,abce} W_{ce} \end{aligned}$$

where in the last line we identify the components of the 4-tensor $\Psi_{\mathcal{G}}$.

B. Additional Proofs

In this section we give proofs of Lemma 1 and Proposition 5. Throughout this section, as in the rest of the paper, $\mathcal{G}$ will be a compact, second countable and Hausdorff topological group. There exists a unique left and right invariant Haar measure λ on $\mathcal{G}$ (Kallenberg, 2006, Theorem 2.27), which we may normalise to be a probability measure $\lambda(\mathcal{G}) = 1$. The Haar measure is Radon which means it is finite on any compact set, so it is clearly normalisable $\lambda(\mathcal{G}) < \infty$. This also immediately implies that λ is σ-finite, allowing us to use Fubini's theorem.[9]

B.1. Proof of Lemma 1

Let $\mathcal{X}$ be an input space and μ be a σ-finite, $\mathcal{G}$-invariant measure on $\mathcal{X}$. We consider vector-valued functions $f : \mathcal{X} \to \mathbb{R}^k$ for some $k \in \mathbb{N}$. Let $\langle \cdot, \cdot \rangle : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ be an inner product on $\mathbb{R}^k$ and let $\|\cdot\|$ be the induced norm. It is possible for the inner product to vary with $x \in \mathcal{X}$, making the norm a local metric on $\mathcal{X}$, as long as the inner product evaluated at any point $x \in \mathcal{X}$, i.e. $\iota_{a,b}(x) = \langle a, b \rangle(x)$, is $\mathcal{G}$-invariant as a function of x for any $a, b \in \mathbb{R}^k$. We will consider the Bochner space V of all integrable, normalisable $f : \mathcal{X} \to \mathbb{R}^k$. By integrable we mean that $\int_{\mathcal{X}} |\langle a, f(x) \rangle| \, \mathrm{d}\mu(x) < \infty \; \forall a \in \mathbb{R}^k$ (this allows us to use Jensen's inequality). To construct the Bochner space we define an inner product

$$\langle f, h \rangle_\mu = \int_{\mathcal{X}} \langle f(x), h(x) \rangle \, \mathrm{d}\mu(x)$$

[9] Weaker technical conditions are possible to achieve σ-finite λ. See Section 2.3 of Folland (2016).

with corresponding norm $\|f\|_\mu = \sqrt{\langle f, f\rangle_\mu}$ and set V to be the space of all f with $\|f\|_\mu < \infty$. Two feature maps are equal if they disagree only on sets with μ-measure 0.

Let the measurable map $\psi : \mathcal{G} \to \mathrm{GL}_k(\mathbb{R})$ be a representation of $\mathcal{G}$. We will assume that ψ is unitary with respect to $\langle \cdot, \cdot \rangle$, by which we mean that $\forall a, b \in \mathbb{R}^k$ and $\forall g \in \mathcal{G}$

$$\langle \psi(g)a, \psi(g)b\rangle = \langle a, b\rangle.$$

Notice that this implies $\langle \psi(g)a, b\rangle = \langle a, \psi(g^{-1})b\rangle$. If $\langle \cdot, \cdot \rangle$ is the Euclidean inner product, then this is the usual notion of a unitary representation. The assumption of unitarity is not stringent, since one can always apply the Weyl trick. We say that $f \in V$ is equivariant if $f(gx) = \psi(g)f(x)\ \forall g \in \mathcal{G}, x \in \mathcal{X}$. Define the operator $\mathcal{Q} : V \to V$ to have values

$$(\mathcal{Q}f)(x) = \int_{\mathcal{G}} \psi(g^{-1})f(gx)\,\mathrm{d}\lambda(g).$$

For convenience we will write this as $\mathcal{Q}f(x) = \mathbb{E}[\psi(G^{-1})f(Gx)]$ where $G \sim \lambda$ will be distributed according to the Haar measure on $\mathcal{G}$. The developments of this section apply to $\mathcal{O} : f(x) \mapsto \int_{\mathcal{G}} f(gx)\,\mathrm{d}\lambda(g)$ by letting ψ be the trivial representation.

We first check that $\mathcal{Q}$ is well-defined.

Proposition 23. Let $f \in V$, then

1. $\mathcal{Q}f$ is μ-measurable, and
2. $\mathcal{Q}f \in V$.

Proof.

1. Writing the action ϕ of $\mathcal{G}$ on $\mathcal{X}$ explicitly, the function $\psi \circ f \circ \phi : \mathcal{G} \times \mathcal{X} \to \mathcal{Y}$ with $\psi \circ f \circ \phi : (g, x) \mapsto \psi(g^{-1})f(\phi(g)x)$ is $(\lambda \otimes \mu)$-measurable, so $\mathcal{Q}f$ is μ-measurable by Kallenberg (2006, Lemma 1.26).
2. We apply in sequel Jensen's inequality (Kallenberg, 2006, Lemma 3.5), the unitarity of ψ, Fubini's theorem (Kallenberg, 2006, Theorem 1.27) and finally the invariance of μ.

$$\begin{aligned}
\|\mathcal{Q}f\|_\mu^2 &= \int_{\mathcal{X}} \|\mathbb{E}[\psi(G^{-1})f(Gx)]\|^2\,\mathrm{d}\mu(x) \\
&\leq \int_{\mathcal{X}} \mathbb{E}[\|\psi(G^{-1})f(Gx)\|^2]\,\mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} \mathbb{E}[\|f(Gx)\|^2]\,\mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} \|f(x)\|^2\,\mathrm{d}\mu(x) \\
&= \|f\|_\mu < \infty
\end{aligned}$$

□

Proposition 24. f is equivariant if and only if $\mathcal{Q}f = f$.

Proof. Suppose f is equivariant then $f(gx) = \psi(g)f(x)\ \forall g \in \mathcal{G}, \forall x \in \mathcal{X}$. Hence for any $x \in \mathcal{X}$

$$\mathcal{Q}f(x) = \mathbb{E}[\psi(G^{-1})f(Gx)] = \mathbb{E}[\psi(G^{-1})\psi(G)f(x)] = f(x).$$

Now assume that $\mathcal{Q}f = f$, so for any $x \in \mathcal{X}$ $f(x) = \mathbb{E}[\psi(G^{-1})f(Gx)]$. Take any $h \in \mathcal{G}$, then

$$\begin{aligned}
f(hx) &= \mathbb{E}[\psi(G^{-1})f(Ghx)] \\
&= \psi(h)\,\mathbb{E}[\psi((Gh)^{-1})f(Ghx)] \\
&= \psi(h)\,\mathbb{E}[\psi(G^{-1})f(Gx)] \\
&= \psi(h)f(x)
\end{aligned}$$

where in the third line we used the right invariance of the Haar measure. □

Proposition 25. $\mathcal{Q}$ has only two eigenvalues, 1 and 0.

Proof. By Proposition 24, $\mathcal{Q}^2 f = \mathcal{Q} f$. So $\mathcal{Q} f = \lambda f$ implies $\lambda^2 = \lambda$. □

Let S and A be the eigenspaces with eigenvalues 1 and 0 respectively. Any $f \in V$ can be written $f = \bar{f} + f^\perp$ where $\bar{f} = \mathcal{Q} f$ and $f^\perp = f - \mathcal{Q} f$. This implies that $V = S + A$. We conclude by showing that $\mathcal{Q}$ is self-adjoint with respect to $\langle \cdot, \cdot \rangle$. Lemma 1 follows immediately, since if $f \in S$ and $h \in A$ then

$$\langle f, h \rangle_\mu = \langle \mathcal{Q} f, h \rangle_\mu = \langle f, \mathcal{Q} h \rangle_\mu = \langle f, 0 \rangle_\mu = 0.$$

Proposition 26. $\mathcal{Q}$ is self-adjoint with respect to $\langle \cdot, \cdot \rangle_\mu$.

Proof.

$$\begin{aligned} \langle \mathcal{Q} f, h \rangle_\mu &= \int_{\mathcal{X}} \langle \mathbb{E}[\psi(G^{-1}) f(Gx)], h(x) \rangle \,\mathrm{d}\mu(x) \\ &= \int_{\mathcal{X}} \mathbb{E}[\langle f(Gx), \psi(G) h(x) \rangle] \,\mathrm{d}\mu(x) \\ &= \int_{\mathcal{X}} \langle f(x), \mathbb{E}[\psi(G) h(G^{-1} x)] \rangle \,\mathrm{d}\mu(x) \end{aligned}$$

where we have used the unitarity of ψ and the invariance of μ. We conclude with the following claim.

Claim. For any $x \in \mathcal{X}$

$$\mathbb{E}[\psi(G) h(G^{-1} x)] = (\mathcal{Q} h)(x).$$

Proof of claim. Since $\mathcal{G}$ is compact it is unimodular (Folland, 2016, Corollary 2.28). Let $A \subset \mathcal{G}$ be measurable, then by Folland (2016, Proposition 2.31) $\lambda(\{a^{-1} : a \in A\}) = \lambda(A)$. So we can just make the change of variables $g \mapsto g^{-1}$

$$\begin{aligned} \mathbb{E}[\psi(G) h(G^{-1} x)] &= \int_{\mathcal{G}} \psi(g) h(g^{-1} x) \,\mathrm{d}\lambda(g) \\ &= \int_{\mathcal{G}} \psi(g^{-1}) h(g x) \,\mathrm{d}\lambda(g^{-1}) \\ &= \int_{\mathcal{G}} \psi(g^{-1}) h(g x) \,\mathrm{d}\lambda(g) \\ &= (\mathcal{Q} h)(x). \end{aligned}$$

■

□

B.2. Proof of Proposition 3

Proposition (Proposition 3)**.** Let V be the space of all normalisable feature extractors as defined above. Define S and A as in Lemma 1. For any $f \in V$, feature averaging with $\mathcal{O}$ maps $f \mapsto \bar{f}$ where $\bar{f}$ is the (μ-a.e.) unique solution to the least-squares problem

$$\bar{f} = \operatorname*{argmin}_{s \in S} \| f - s \|_\mu^2.$$

Proof. Using Lemma 1 and Proposition 24 we can write $S = \{f \in V : \mathcal{O}f = f\}$ which in turn implies that any $f \in V$ has the decomposition $f = s + a$ where $\mathcal{O}s = s$, $\mathcal{O}a = 0$ and $\langle s, a\rangle_\mu = 0$. Hence $\overline{f} = s = \mathcal{O}f$. Now take any $h \in S$ and recall that $\langle h, a\rangle_\mu = 0$, then

$$\|f - h\|_\mu^2 = \|(s - h) + a\|_\mu^2 = \|s - h\|_\mu^2 + \|a\|_\mu^2 \geq \|a\|_\mu^2 = \|f - s\|_\mu^2.$$

Uniqueness follows by a simple convexity argument. Suppose $\exists s' \in S$ with $s' \neq s$ and $\|f - s'\|_\mu = \|f - s\|_\mu$, then since S is a vector space we have $s_{\frac{1}{2}} = \frac{1}{2}(s + s') \in S$. It follows that

$$\begin{aligned}
\|f - s_{\frac{1}{2}}\|_\mu^2 &= \|(f - s)/2 + (f - s')/2\|_\mu^2 \\
&= \frac{1}{4}\|f - s\|_\mu^2 + \frac{1}{4}\|f - s'\|_\mu^2 + \frac{1}{4}\langle f - s, f - s'\rangle_\mu \\
&\leq \frac{1}{4}\|f - s\|_\mu^2 + \frac{1}{4}\|f - s\|_\mu^2 + \frac{1}{4}\|f - s\|_\mu^2 \\
&= \frac{3}{4}\|f - s\|_\mu^2
\end{aligned}$$

a contradiction unless $\|f - s\|_\mu^2 = 0$, in which case $f = s$ μ-almost-everywhere. □

B.3. Proof of Proposition 5

Let $\mathcal{T}$ be a space on which $\mathcal{G}$ acts measurably. Let $f : \mathcal{T} \to \mathbb{R}$ be a integrable function. Recall the orbit averaging of f is the function $\mathcal{O}f : \mathcal{T} \to \mathbb{R}$ with values

$$(\mathcal{O}f)(t) = \mathbb{E}[f(Gt)]$$

where $G \sim \lambda$. For any set $\mathcal{F}$ of integrable functions $f : \mathcal{T} \to \mathbb{R}$ we define the symmetric and anti-symmetric classes as

$$\overline{\mathcal{F}} = \{\mathcal{O}f : f \in \mathcal{F}\} \quad \text{and} \quad \mathcal{F}^\perp = \{f - \mathcal{O}f : f \in \mathcal{F}\},$$

respectively. Notice that: 1. by Proposition 24, f is $\mathcal{G}$-invariant iff $\mathcal{O}f = f$, 2. that everything in the symmetric class is preserved by $\mathcal{O}$ and everything in the anti-symmetric class vanishes under $\mathcal{O}$, and 3. that $\mathcal{O}f$ is measurable whenever f is measurable by Proposition 23.

Proposition (Proposition 5)**.** Suppose ν is a $\mathcal{G}$-invariant probability distribution on $\mathcal{T}$. Let $\mathcal{T}$ be some input space and let $\mathcal{F}$ be a set of ν-integrable functions $f : \mathcal{T} \to \mathbb{R}$. Then the Rademacher complexity of the feature averaged class satisfies

$$0 \leq \mathfrak{R}_n(\mathcal{F}) - \mathfrak{R}_n(\overline{\mathcal{F}}) \leq \mathfrak{R}_n(\mathcal{F}^\perp)$$

where the expectations in the definition of $\mathfrak{R}_n$ are taken over $t_i \sim \nu$ i.i.d..

We start with the left hand side.

Claim.

$$\mathfrak{R}_n\,\overline{\mathcal{F}} \leq \mathfrak{R}_n\,\mathcal{F}.$$

Proof of claim. Let $t \in \mathcal{T}^n$. The action of $\mathcal{G}$ on $\mathcal{T}$ induces an action on $\mathcal{T}^n$ by $gt = (gt_1, \dots, gt_n)$. If $T \sim \nu^n$ then by $\mathcal{G}$-invariance of ν we have $gT \overset{\text{d}}{=} T\ \forall g \in \mathcal{G}$. Let $T_1, \dots, T_n \sim \nu$ with $T = \{T_1, \dots, T_n\}$. Then, with subscripts on expectations for clarity,

$$
\begin{aligned}
\mathfrak{R}_n \overline{\mathcal{F}} &= \frac{1}{n} \mathbb{E}_T \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\bar{f} \in \overline{\mathcal{F}}} \left| \sum_{i=1}^n \sigma_i \bar{f}(T_i) \right| \\
&= \frac{1}{n} \mathbb{E}_T \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} \left| \mathbb{E}_G \sum_{i=1}^n \sigma_i f(GT_i) \right| \\
&\leq \frac{1}{n} \mathbb{E}_T \mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_G \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i f(GT_i) \right| \\
&= \frac{1}{n} \mathbb{E}_T \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i f(T_i) \right| \qquad (\star) \\
&= \mathfrak{R}_n \mathcal{F}
\end{aligned}
$$

In deducing ($\star$) we used Fubini's theorem (Kallenberg, 2006, Theorem 1.27) and the $\mathcal{G}$-invariance of ν. Fubini's theorem applies by (Kallenberg, 2006, Lemma 1.26). □

Now for the right hand side.

Proof. For any $f \in \mathcal{F}$ we can write $f = \bar{f} + f^\perp$ where $\bar{f} \in \overline{\mathcal{F}}$ and $f^\perp \in \mathcal{F}^\perp$. Then, for any $\tau = \{t_1, \dots, t_n\}$

$$
\begin{aligned}
\widehat{\mathfrak{R}}_\tau \mathcal{F} &= \frac{1}{n} \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i f(t_i) \right| \\
&= \frac{1}{n} \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i (\bar{f}(t_i) + f^\perp(t_i)) \right| \\
&\leq \frac{1}{n} \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\bar{f} \in \overline{\mathcal{F}}} \left| \sum_{i=1}^n \sigma_i \bar{f}(t_i) \right| + \frac{1}{n} \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f^\perp \in \mathcal{F}^\perp} \left| \sum_{i=1}^n \sigma_i f^\perp(t_i) \right| \\
&= \widehat{\mathfrak{R}}_\tau \overline{\mathcal{F}} + \widehat{\mathfrak{R}}_\tau \mathcal{F}^\perp
\end{aligned}
$$

Taking an expectation over $\tau \sim \nu^n$ and combining with the claim gives

$$\mathfrak{R}_n \overline{\mathcal{F}} \leq \mathfrak{R}_n \mathcal{F} \leq \mathfrak{R}_n \overline{\mathcal{F}} + \mathfrak{R}_n \mathcal{F}^\perp$$

from which the proposition follows immediately. □

B.4. Proof of Theorem 13

Theorem (Theorem 13)**.** Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}^k$ and let $\mathcal{G}$ be a compact group with orthogonal representations ϕ on $\mathcal{X}$ and ψ on $\mathcal{Y}$. Let $X \sim \mathcal{N}(0, \sigma_X^2 I_d)$ and $Y = h_\Theta(X) + \xi$ where $h_\Theta(x) = \Theta^\top x$ is $\mathcal{G}$-equivariant and $\Theta \in \mathbb{R}^{d \times k}$. Assume ξ is a random element of $\mathbb{R}^k$, independent of X, with mean 0 and $\mathbb{E}[\xi\xi^\top] = \sigma_\xi^2 I_k < \infty$. Let W be the least-squares estimate of Θ from n i.i.d. examples $\{(X_i, Y_i) : i = 1, \dots, n\}$ and let $(\chi_\psi | \chi_\phi) = \int_{\mathcal{G}} \chi_\psi(g) \chi_\phi(g) \, \mathrm{d}\lambda(g)$ denote the scalar product of the characters of the representations of $\mathcal{G}$.

- If $n > d + 1$ the generalisation gap is

$$\mathbb{E}[\Delta(f_W, f_{\overline{W}})] = \sigma_\xi^2 \frac{dk - (\chi_\psi | \chi_\phi)}{n - d - 1}.$$

- At the interpolation threshold $n \in [d - 1, d + 1]$, if f_W is not $\mathcal{G}$-equivariant then the generalisation gap diverges to ∞.
- If $n < d - 1$ then the generalisation gap is

$$\mathbb{E}[\Delta(f_W, f_{\overline{W}})] = \sigma_X^2 \frac{n(d-n)}{d(d-1)(d+2)} \left((d+1) \|\Theta\|_{\mathrm{F}}^2 - \mathrm{Tr}(J_{\mathcal{G}} \Theta^\top \Theta) \right) + \sigma_\xi^2 \frac{n(dk - (\chi_\psi | \chi_\phi))}{d(d - n - 1)}$$

where each term is non-negative and $J_{\mathcal{G}} \in \mathbb{R}^{k \times k}$ is given by

$$J_{\mathcal{G}} = \int_{\mathcal{G}} (\chi_\phi(g)\psi(g) + \psi(g^2)) \,\mathrm{d}\lambda(g).$$

Proof. We use Einstein notation, in which repeated indices are summed over. δ_{ij} represents the Kronecker delta, which is 1 when $i = j$ and 0 otherwise.

Since the representation ϕ is orthogonal, X is $\mathcal{G}$-invariant for any $\mathcal{G}$. We have seen from Proposition 12 that

$$\mathbb{E}[\Delta(f_W, f_{\overline{W}})] = \sigma_X^2 \,\mathbb{E}[\|W^\perp\|_{\mathrm{F}}^2]$$

and we want to understand this quantity for the least-squares estimate

$$W = (\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top \boldsymbol{Y} = (\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top \boldsymbol{X} \Theta + (\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top \boldsymbol{\xi}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, $\boldsymbol{Y} \in \mathbb{R}^{n \times k}$ are the row-stacked training examples with $(\boldsymbol{X})_{ij} = (X_i)_j$, $(\boldsymbol{Y}_i)_j = (Y_i)_j$ and $\boldsymbol{\xi} = \boldsymbol{Y} - \boldsymbol{X}\Theta$. We have

$$\begin{aligned}
\mathbb{E}[\Delta(f_W, f_{\overline{W}})] &= \sigma_X^2 \,\mathbb{E}[\|\Psi_{\mathcal{G}}{}^\perp(W)\|_{\mathrm{F}}^2] \\
&= \sigma_X^2 \,\mathbb{E}[\|\Psi_{\mathcal{G}}{}^\perp((\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top \boldsymbol{X} \Theta + (\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top \boldsymbol{\xi})\|_{\mathrm{F}}^2] \\
&= \sigma_X^2 \,\mathbb{E}[\|\Psi_{\mathcal{G}}{}^\perp((\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top \boldsymbol{X} \Theta)\|_{\mathrm{F}}^2] + \sigma_X^2 \,\mathbb{E}[\|\Psi_{\mathcal{G}}{}^\perp((\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top \boldsymbol{\xi})\|_{\mathrm{F}}^2]
\end{aligned}$$

using linearity and $\mathbb{E}[\boldsymbol{\xi}] = 0$. We treat the two terms separately, starting with the second.

Second Term Now consider the second term, setting $\boldsymbol{Z} = (\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top$ we have

$$\mathbb{E}[\|\Psi_{\mathcal{G}}{}^\perp(\boldsymbol{Z}\boldsymbol{\xi})\|_{\mathrm{F}}^2] = \mathbb{E}[\mathrm{Tr}(\Psi_{\mathcal{G}}{}^\perp(\boldsymbol{Z}\boldsymbol{\xi})^\top \Psi_{\mathcal{G}}{}^\perp(\boldsymbol{Z}\boldsymbol{\xi}))].$$

One gets

$$\begin{aligned}
\mathbb{E}[\mathrm{Tr}(\Psi_{\mathcal{G}}{}^\perp(\boldsymbol{Z}\boldsymbol{\xi})^\top \Psi_{\mathcal{G}}{}^\perp(\boldsymbol{Z}\boldsymbol{\xi}))] &= \mathbb{E}[\Psi_{\mathcal{G}}{}^\perp_{abcj} \boldsymbol{Z}_{ce} \boldsymbol{\xi}_{ej} \Psi_{\mathcal{G}}{}^\perp_{abfg} \boldsymbol{Z}_{fh} \boldsymbol{\xi}_{hg}] \\
&= \sigma_\xi^2 \,\mathbb{E}[\Psi_{\mathcal{G}}{}^\perp_{abcj} \boldsymbol{Z}_{ce} \Psi_{\mathcal{G}}{}^\perp_{abfg} \boldsymbol{Z}_{fh} \delta_{eh} \delta_{jg}] \quad \text{integrating } \boldsymbol{\xi} \\
&= \sigma_\xi^2 \Psi_{\mathcal{G}}{}^\perp_{abcj} \Psi_{\mathcal{G}}{}^\perp_{abfj} \,\mathbb{E}[\boldsymbol{Z}_{ce} \boldsymbol{Z}_{fe}] \\
&= \sigma_\xi^2 \Psi_{\mathcal{G}}{}^\perp_{abcj} \Psi_{\mathcal{G}}{}^\perp_{abfj} \,\mathbb{E}[(\boldsymbol{Z}\boldsymbol{Z}^\top)_{cf}]
\end{aligned} \tag{5}$$

and then (WLOG relabelling $f \mapsto e$)

$$\begin{aligned}
\Psi_{\mathcal{G}}{}^\perp_{abcj} \Psi_{\mathcal{G}}{}^\perp_{abej} &= \left(\delta_{ac}\delta_{bj} - \int_{\mathcal{G}} \phi(g)_{ac} \psi(g)_{bj} \,\mathrm{d}\lambda(g)\right) \left(\delta_{ae}\delta_{bj} - \int_{\mathcal{G}} \phi(g)_{ae} \psi(g)_{bj} \,\mathrm{d}\lambda(g)\right) \\
&= \delta_{ac}\delta_{bj}\delta_{ae}\delta_{bj} - \delta_{ae}\delta_{bj} \int_{\mathcal{G}} \phi(g)_{ac} \psi(g)_{bj} \,\mathrm{d}\lambda(g) \\
&\quad - \delta_{ac}\delta_{bj} \int_{\mathcal{G}} \phi(g)_{ae} \psi(g)_{bj} \,\mathrm{d}\lambda(g) + \int_{\mathcal{G}} \phi(g_1)_{ac} \phi(g_2)_{ae} \psi(g_1)_{bj} \psi(g_2)_{bj} \,\mathrm{d}\lambda(g_1) \,\mathrm{d}\lambda(g_2) \\
&= k\, \delta_{ce} - \int_{\mathcal{G}} \mathrm{Tr}(\psi(g))(\phi(g)_{ec} + \phi(g)_{ce}) \,\mathrm{d}\lambda(g) \\
&\quad + \int_{\mathcal{G}} \mathrm{Tr}(\psi(g_1)^\top \psi(g_2))(\phi(g_1)^\top \phi(g_2))_{ce} \,\mathrm{d}\lambda(g_1) \,\mathrm{d}\lambda(g_2)
\end{aligned}$$

where we have used that the indices $b, j = 1, \dots, k$. Consider the final term

$$\begin{aligned}
\int_{\mathcal{G}} \mathrm{Tr}(\psi(g_1)^\top \psi(g_2))(\phi(g_1)^\top \phi(g_2))_{ce} \,\mathrm{d}\lambda(g_1) \,\mathrm{d}\lambda(g_2) &= \int_{\mathcal{G}} \mathrm{Tr}(\psi(g_1^{-1} g_2))(\phi(g_1^{-1} g_2))_{ce} \,\mathrm{d}\lambda(g_1) \,\mathrm{d}\lambda(g_2) \\
&= \int_{\mathcal{G}} \mathrm{Tr}(\psi(g))\phi(g)_{ce} \,\mathrm{d}\lambda(g)
\end{aligned}$$

where we used that the representations are orthogonal, Fubini's theorem and that the Haar measure is invariant ($\mathcal{G}$ is compact so λ is σ-finite and Fubini's theorem applies). Now we put things back together. To begin with

$$\Psi_{\mathcal{G}}{}^{\perp}_{abcj}\Psi_{\mathcal{G}}{}^{\perp}_{abej} = k\,\delta_{ce} - \int_{\mathcal{G}} \operatorname{Tr}(\psi(g))\phi(g^{-1})_{ce}\,\mathrm{d}\lambda(g)$$

and putting this into Eq. (5) with $\boldsymbol{Z}\boldsymbol{Z}^\top = (\boldsymbol{X}^\top\boldsymbol{X})^+$ gives

$$\mathbb{E}[\operatorname{Tr}(\Psi_{\mathcal{G}}{}^{\perp}(\boldsymbol{Z}\boldsymbol{\xi})^\top\Psi_{\mathcal{G}}{}^{\perp}(\boldsymbol{Z}\boldsymbol{\xi}))] = \sigma_\xi^2\left(k\,\delta_{ce} - \int_{\mathcal{G}} \operatorname{Tr}(\psi(g))\phi(g^{-1})_{ce}\,\mathrm{d}\lambda(g)\right)\mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{X})^+_{ce}]$$

where $c, e = 1, \dots, d$. Applying Lemmas 18 and 20 gives $\mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{X})^+_{ce}] = \sigma_X^{-2} r(n,d)\delta_{ce}$ where

$$r(n,d) = \begin{cases} \frac{n}{d(d-n-1)} & n < d-1 \\ (n-d-1)^{-1} & n > d+1 \\ \infty & \text{otherwise} \end{cases}.$$

When $n \in [d-1, d+1]$ it is well known that the expectation diverges, see Appendix A. We can conclude:

$$\sigma_X^2\,\mathbb{E}[\|\Psi_{\mathcal{G}}{}^{\perp}((\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{\xi})\|_{\mathrm{F}}^2] = \sigma_\xi^2 r(n,d)\left(dk - \int_{\mathcal{G}} \operatorname{Tr}(\psi(g))\operatorname{Tr}(\phi(g))\,\mathrm{d}\lambda(g)\right)$$

where we have used the orthogonality of ϕ.

First Term If $n \geq d$ then $(\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{X}\Theta \overset{\text{a.s.}}{=} \Theta$ and since $h_\Theta \in S$ the first term vanishes almost surely. This gives the case of equality in the statement. If $n < d$ we proceed as follows. Write $P_E = (\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{X}$ which is the orthogonal projection onto the rank of $\boldsymbol{X}^\top\boldsymbol{X}$. By isotropy of X, $E \sim \operatorname{Unif}\mathbb{G}_n(\mathbb{R}^d)$ with probability 1. Recall that $\Psi_{\mathcal{G}}{}^{\perp}(\Theta) = 0$, which in components reads

$$\Psi_{\mathcal{G}}{}^{\perp}_{abce}\Theta_{ce} = 0 \quad \forall a, b. \tag{6}$$

Also in components, we have

$$\mathbb{E}[\|\Psi_{\mathcal{G}}{}^{\perp}((\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{X}\Theta)\|_{\mathrm{F}}^2] = \Psi_{\mathcal{G}}{}^{\perp}_{fhai}\Psi_{\mathcal{G}}{}^{\perp}_{fhcj}\,\mathbb{E}[P_E \otimes P_E]_{abce}\Theta_{bi}\Theta_{ej}$$

and using Lemma 22 we get

$$\begin{aligned}\mathbb{E}[\|\Psi_{\mathcal{G}}{}^{\perp}((\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{X}\Theta)\|_{\mathrm{F}}^2] &= \frac{n(d-n)}{d(d-1)(d+2)}\left(\Psi_{\mathcal{G}}{}^{\perp}_{fhai}\Psi_{\mathcal{G}}{}^{\perp}_{fhaj}\Theta_{bi}\Theta_{bj} + \Psi_{\mathcal{G}}{}^{\perp}_{fhai}\Psi_{\mathcal{G}}{}^{\perp}_{fhbj}\Theta_{bi}\Theta_{aj}\right)\\ &\quad + \frac{n(d-n)+n(n-1)(d+2)}{d(d-1)(d+2)}\Psi_{\mathcal{G}}{}^{\perp}_{fhai}\Psi_{\mathcal{G}}{}^{\perp}_{fhcj}\Theta_{ai}\Theta_{cj}.\end{aligned}$$

The final term vanishes using Eq. (6), while the first term is

$$\Psi_{\mathcal{G}}{}^{\perp}_{fhai}\Psi_{\mathcal{G}}{}^{\perp}_{fhaj}\Theta_{bi}\Theta_{bj} = (\Theta^\top\Theta)_{ij}\Psi_{\mathcal{G}}{}^{\perp}_{fhai}\Psi_{\mathcal{G}}{}^{\perp}_{fhaj}$$

where

$$\begin{aligned}\Psi_{\mathcal{G}}{}^{\perp}_{fhai}\Psi_{\mathcal{G}}{}^{\perp}_{fhaj} &= \left(\delta_{fa}\delta_{hi} - \int_{\mathcal{G}} \phi(g)_{fa}\psi(g)_{hi}\,\mathrm{d}\lambda(g)\right)\left(\delta_{fa}\delta_{hj} - \int_{\mathcal{G}} \phi(g)_{fa}\psi(g)_{hj}\,\mathrm{d}\lambda(g)\right)\\ &= d\delta_{ij} - \int_{\mathcal{G}} \operatorname{Tr}(\phi(g))\psi(g)_{ij}\,\mathrm{d}\lambda(g) - \int_{\mathcal{G}} \operatorname{Tr}(\phi(g))\psi(g)_{ji}\,\mathrm{d}\lambda(g)\\ &\quad + \int_{\mathcal{G}} \phi(g_1)_{fa}\phi(g_2)_{fa}\psi(g_1)_{hi}\psi(g_2)_{hj}\,\mathrm{d}\lambda(g_1)\,\mathrm{d}\lambda(g_2)\\ &= d\delta_{ij} - \int_{\mathcal{G}} \operatorname{Tr}(\phi(g))\psi(g)_{ji}\,\mathrm{d}\lambda(g)\end{aligned}$$

using the orthogonality of the representations and invariance of the Haar measure. Therefore

$$\begin{aligned}\Psi_{\mathcal{G}^{\perp}_{fhai}}\Psi_{\mathcal{G}^{\perp}_{fhaj}}\Theta_{bi}\Theta_{bj} &= d\|\Theta\|_{\mathrm{F}}^2 - \int_{\mathcal{G}} \chi_\phi(g)\,\mathrm{Tr}\left(\psi(g^{-1})\Theta^\top\Theta\right)\mathrm{d}\lambda(g)\\ &= d\|\Theta\|_{\mathrm{F}}^2 - \int_{\mathcal{G}} \chi_\phi(g)\,\mathrm{Tr}\left(\psi(g)\Theta^\top\Theta\right)\mathrm{d}\lambda(g).\end{aligned}$$

Now for the second part

$$\begin{aligned}\Theta_{bi}\Theta_{aj}\Psi_{\mathcal{G}^{\perp}_{fhai}}\Psi_{\mathcal{G}^{\perp}_{fhbj}} &= \Theta_{bi}\Theta_{aj}\left(\delta_{fa}\delta_{hi} - \int_{\mathcal{G}}\phi(g)_{fa}\psi(g)_{hi}\,\mathrm{d}\lambda(g)\right)\left(\delta_{fb}\delta_{hj} - \int_{\mathcal{G}}\phi(g)_{fb}\psi(g)_{hj}\,\mathrm{d}\lambda(g)\right)\\ &= \Theta_{bi}\Theta_{aj}\left(\delta_{ab}\delta_{ij} - \int_{\mathcal{G}}\phi(g)_{ab}\psi(g)_{ij}\,\mathrm{d}\lambda(g)\right)\\ &= \|\Theta\|_{\mathrm{F}}^2 - \int_{\mathcal{G}}\mathrm{Tr}(\Theta^\top\phi(g)\Theta\psi(g))\,\mathrm{d}\lambda(g)\\ &= \|\Theta\|_{\mathrm{F}}^2 - \int_{\mathcal{G}}\mathrm{Tr}(\psi(g^2)\Theta^\top\Theta)\,\mathrm{d}\lambda(g).\end{aligned}$$

Putting these together gives

$$\mathbb{E}[\|\Psi_{\mathcal{G}^\perp}((\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{X}\Theta)\|_{\mathrm{F}}^2] = \frac{n(d-n)}{d(d-1)(d+2)}\left((d+1)\|\Theta\|_{\mathrm{F}}^2 - \mathrm{Tr}(J_{\mathcal{G}}\Theta^\top\Theta)\right)$$

where $J_{\mathcal{G}} \in \mathbb{R}^{k\times k}$ is the matrix-valued function of $\mathcal{G}$, ψ and ϕ

$$J_{\mathcal{G}} = \int_{\mathcal{G}}(\chi_\phi(g)\psi(g) + \psi(g^2))\,\mathrm{d}\lambda(g).$$

□

References

Abu-Mostafa, Y. S. Hints and the VC dimension. *Neural Computation*, 5(2):278–288, 1993.

Anselmi, F., Leibo, J. Z., Rosasco, L., Mutch, J., Tacchetti, A., and Poggio, T. Unsupervised learning of invariant representations in hierarchical architectures, 2014.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res.*, 20:63–1, 2019.

Bloem-Reddy, B. and Teh, Y. W. Probabilistic symmetries and invariant neural networks. *Journal of Machine Learning Research*, 21(90):1–61, 2020. URL http://jmlr.org/papers/v21/19-322.html.

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

Cohen, T. S., Geiger, M., Köhler, J., and Welling, M. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018.

Cohen, T. S., Geiger, M., and Weiler, M. A general theory of equivariant cnns on homogeneous spaces. In *Advances in Neural Information Processing Systems*, pp. 9145–9156, 2019.

Cook, R. D., Forzani, L., et al. On the mean and variance of the generalized inverse of a singular wishart matrix. *Electronic Journal of Statistics*, 5:146–158, 2011.

Folland, G. B. *A course in abstract harmonic analysis*, volume 29. CRC press, 2016.

Gupta, S. D. Some aspects of discrimination function coefficients. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 387–400, 1968.

Haasdonk, B., Vossen, A., and Burkhardt, H. Invariance in kernel methods by haar integration kernels. In *SCIA 2005, Scandinavian Conference on Image Analysis*, pp. 841–851. Springer-Verlag, 2005.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Hodge, P. G. On isotropic cartesian tensors. *The American Mathematical Monthly*, 68(8):793–795, 1961. ISSN 00029890.

Kallenberg, O. *Foundations of modern probability*. Springer Science & Business Media, 2006.

Kondor, R. and Trivedi, S. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, pp. 2747–2755, 2018.

Lyle, C., Kwiatkowksa, M., and Gal, Y. An analysis of the effect of invariance on generalization in neural networks. In *International conference on machine learning Workshop on Understanding and Improving Generalization in Deep Learning*, 2019.

Lyle, C., van der Wilk, M., Kwiatkowska, M., Gal, Y., and Bloem-Reddy, B. On the benefits of invariance in neural networks, 2020.

Maron, H., Fetaya, E., Segol, N., and Lipman, Y. On the universality of invariant networks. In *International Conference on Machine Learning*, pp. 4363–4371, 2019.

Mroueh, Y., Voinea, S., and Poggio, T. A. Learning with group invariant features: A kernel perspective. In *Advances in Neural Information Processing Systems*, pp. 1558–1566, 2015.

Muirhead, R. J. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009.

Penrose, R. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pp. 406–413. Cambridge University Press, 1955.

Pfau, D., Spencer, J. S., Matthews, A. G., and Foulkes, W. M. C. Ab initio solution of the many-electron schrödinger equation with deep neural networks. *Physical Review Research*, 2(3):033429, 2020.

Ravanbakhsh, S., Schneider, J., and Poczos, B. Equivariance through parameter-sharing. In *International Conference on Machine Learning*, pp. 2892–2901. PMLR, 2017.

Sannai, A. and Imaizumi, M. Improved generalization bound of group invariant / equivariant deep networks via quotient feature space, 2019.

Schölkopf, B., Burges, C., and Vapnik, V. Incorporating invariances in support vector learning machines. pp. 47–52. Springer, 1996.

Schulz-Mirbach, H. On the existence of complete invariant feature spaces in pattern recognition. In *International Conference On Pattern Recognition*, pp. 178–178. Citeseer, 1992.

Schulz-Mirbach, H. Constructing invariant features by averaging techniques. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 - Conference C: Signal Processing (Cat. No.94CH3440-5)*, volume 2, pp. 387–390 vol.2, 1994.

Serre, J.-P. *Linear representations of finite groups*. Graduate texts in mathematics ; 42. Springer-Verlag, New York, 1977. ISBN 9780387901909.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Sokolic, J., Giryes, R., Sapiro, G., and Rodrigues, M. Generalization error of invariant classifiers. In *Artificial Intelligence and Statistics*, pp. 1094–1103, 2017.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015. doi: 10.1109/CVPR.2015.7298594.

Wadsley, S. Lecture notes on representation theory, October 2012. URL https://www.dpmms.cam.ac.uk/~sjw47/RepThLecturesMich2012.pdf.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Winkels, M. and Cohen, T. S. 3d g-cnns for pulmonary nodule detection. *arXiv preprint arXiv:1804.04656*, 2018.

Wood, J. and Shawe-Taylor, J. Representation theory and invariant neural networks. *Discrete applied mathematics*, 69(1-2): 33–60, 1996.

Xu, H. and Mannor, S. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

Yarotsky, D. Universal approximations of invariant maps by neural networks, 2018.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. In *Advances in neural information processing systems*, pp. 3391–3401, 2017.