Symmetry and Generalisation in Machine Learning

Bryn Elesedy

St Anne's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2023

In memory of my dad, Captain Khalid El-Esedy, who would have enjoyed this being submitted at the end of Ramadan. Eid Mubarak, Baba.

And in gratitude to my mum, Linda. Thank you, Mam, for your unconditional(!) love and support.

Contents

Abstract

This work is about understanding the impact of invariance and equivariance on generalisation in supervised learning. We use the perspective afforded by an averaging operator to show that for any predictor that is not equivariant, there is an equivariant predictor with strictly lower test risk on all regression problems where the equivariance is correctly specified. This constitutes a rigorous proof that symmetry, in the form of invariance or equivariance, is a useful inductive bias.

We apply these ideas to equivariance and invariance in random design least squares and kernel ridge regression respectively. This allows us to specify the reduction in expected test risk in more concrete settings and express it in terms of properties of the group, the model and the data.

Along the way, we give examples and additional results to demonstrate the utility of the averaging operator approach in analysing equivariant predictors. In addition, we adopt an alternative perspective and formalise the common intuition that learning with invariant models reduces to a problem in terms of orbit representatives. The formalism extends naturally to a similar intuition for equivariant models. We conclude by connecting the two perspectives and giving some ideas for future work.

Acknowledgements

I would like to thank my advisors Varun Kanade and Yee Whye Teh for their support throughout this process. Their guidance has been invaluable. Their diverse expertise has broadened my perspective on the field and their judicious nudges have made a world of difference. As their student I had both of Berlin's concepts of liberty: freedom from any pressure to publish or produce results as well as support to explore whatever ideas I found interesting. Whatever topic I brought to our meetings, I would always get their honest and engaged feedback.

I have benefitted from the rich intellectual environment cultivated in Yee Whye's group at OxCSML, which has exposed me to areas outside of my research and provided inspiration for new questions. Indeed, the work in this thesis is an attempt to answer questions that arose during the regular reading groups. A special mention goes to Sheheryar Zaidi, Bobby He and Michael Hutchinson, who hold a humbling array of talents. It has been a pleasure to bring down the average of our year group.

There are many others to thank for my time at Oxford, and probably more still that I can't bring to mind at this moment. Martin Lesourd, John Mittermeier and Joel Hart for providing a counterbalance at home. Panagiotis Tigas for a rubber duck code review that saved me from despair. Tim Rudner and Adam Goliński for advice during the early stages of my PhD. Wendy Poole, superhero of the Autonomous Intelligent Machines and Systems CDT, for the million helpful things she has done. Sheheryar Zaidi for his contributions to [52] and enlightening conversations. Shahine Bouabid and Jake Fawkes for notifying me of an error in [51] and their subsequent attempts to fix it. Benjamin Bloem-Reddy for helpful discussions on the topic of this thesis and related areas, including one that lead to Section 7.1.

Few people I'm connected to through Oxford know this, but the route to this point hasn't been straightforward. In that respect, I intend the submission of this thesis to be the end of the beginning. I stopped going to school as a teenager. I left at around 16 with few qualifications and almost took a different trajectory. I ended up

going back to education and since then I've just been putting one foot in front of the other as best I can. Encouragement from others has been a precious resource and I have countless people to thank.

As it happens, I can't remember the first guy's name, but he left me sufficiently humiliated after months of sitting unemployed at my mum's that I found a job and later enrolled in a sixth form college. When I got to college, despite what I thought I was capable of, I knew nothing. Early on I was advised to drop A-Level maths, the reason being that if I worked really hard I might get a C grade, but it was a long shot.

I was thinking of doing so, but then Mike Knowles stepped in with some critical guidance: do I want to give up when I receive a challenge, or actually try? This was the pivotal moment. I kept going with the maths A-Level and it turned into my favourite subject. I was well supported by my teacher Mr. Jones, who left me to explore the subject at my own pace. Around this time Juan Bercial and Peter Giblin took time to show me exciting ideas in mathematics, while Rob Lewis and Glenn Skelhorn nurtured my interest in philosophy and provided useful feedback.

I ended up at Cambridge because I attended a summer school for under-represented students run by Geoff Parks. I learned about the summer school after bumping into my economics teacher one evening in the college corridors (another name I can't remember, sorry!). She recommended that I apply, I shrugged and she submitted on my behalf what would have had to have been quite a convincing application. At the summer school I was encouraged to apply to Cambridge. I did, and I was lucky enough to get in.

At Cambridge I was a student of Stephen Siklos, who invested a lot of time and energy in me. He took most of my supervision sessions one on one and, for whatever reason, was insistent that there was something to be made of me academically. I suspect Stephen's recommendation letter was key to my acceptance to Oxford. From what I can tell, he went out of his way to help many other students as well.

Along the way, I have received lessons, help and encouragement from others including Neil Smith, Carola-Bibiane Schönlieb, Emanuel Malek, Tom Mädler, Jean-Gabriel Prince, Filippo Altissimo, Haydn Davies, Alex Bakker, Mike Osborne, Neil Lawrence, Federico Vaggi, David Duvenaud, Mihaela Rosca and Marcus Hutter.

There are countless people for whom I'm grateful in my personal life, but I won't go into details. It doesn't feel like the forum. I have great friends and during difficult times the fun has kept me going. A loving and supporting family with a great sense of humour. Welcoming (de facto) in-laws who feed me and treat me as one of their own. And, saving the best until last, a long-suffering other half, Issy, who has been carrying my heart for more than a decade. She listens to me, supports me, and even laughs at my jokes. I couldn't ask for more.

This DPhil was supported by funding from the UK EPSRC CDT in Autonomous Intelligent Machines and Systems (grant reference EP/L015897/1).

Chapter 1

Introduction

1.1 Motivation

We will study how invariance and equivariance affect generalisation in supervised learning. Let $\mathcal{G}$ be a group acting on sets $\mathcal{X}$ and $\mathcal{Y}$, then $f : \mathcal{X} \to \mathcal{Y}$ is invariant if $f(gx) = f(x)$ and equivariant if $f(gx) = gf(x)$, each holding for all $g \in \mathcal{G}$ and $x \in \mathcal{X}$. Invariance is the special case of equivariance where the action of $\mathcal{G}$ on $\mathcal{Y}$ is trivial. In this work, the word *symmetry* specifically refers to some form of invariance or equivariance.

Supervised learning is the science of extrapolation from labelled data. The basic task is as follows: given a sequence of input-output pairs $(x_1, y_1), \dots, (x_n, y_n)$ generated by some unknown, possibly stochastic procedure, predict unseen values (x, y) generated by the same procedure. Fundamentally, it is a problem of inductive inference.

More formally, let X and Y be random elements of $\mathcal{X}$ and $\mathcal{Y}$ respectively whose distributions are unknown. We call the sequence $(x_1, y_1), \dots, (x_n, y_n)$ the *observations* and assume that they are, respectively, instances of the random variables $(X_1, Y_1), \dots, (X_n, Y_n)$. The tuple $((X_1, Y_1), \dots, (X_n, Y_n))$ is called the *training sample* and individual elements of the training sample are *training examples* or just *examples*. We assume the training examples are distributed independently and identically to (X, Y). The word *data* may refer to either fixed observations or the training sample depending on the context.

We formulate supervised learning as the problem of finding, given the observations, a minimiser over $f : \mathcal{X} \to \mathcal{Y}$ of

$$R[f] = \mathbb{E}[\ell(f(X), Y)]$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is the *loss function* which measures the quality of predictions. We will refer to R as the *risk function* and to $R[f]$ as the *risk of f*; when the distinction isn't needed we may refer to each of these as the *risk*. We call the candidate functions f *predictors* or *models*. An *algorithm* is a map from training samples to predictors.

Without access to the distribution of (X, Y) it is not possible to minimise the risk directly and, in any case, finding a minimiser of R could be computational infeasible. In practice, an approximate solution is acceptable. In the same vein, theoretical study is often specialised to certain relationships between X and Y or the minimisation constrained to certain classes of predictors.

From a mathematical perspective, the problem of finding a predictor with small risk is critically underdetermined. Judicious choice of a predictor requires extrapolating from the finite number of observations to the general relationship between X and Y. The term *inductive bias* refers loosely to a mode of extrapolation. For instance, linear predictors share an inductive bias because they extrapolate from the training sample in a conceptually similar fashion.

Symmetry is also a form of inductive bias. Along an orbit of $\mathcal{G}$, invariant models are constant while the values of equivariant models are related by the group action. Symmetry has emerged as a popular method of incorporating domain knowledge into models [34, 94, 35, 186, 57] and these models have applications in many areas.

For instance, where symmetry is known to be a fundamental property of the system such as quantum chemistry [128], or where arbitrary experimenter choices or data representation give undue privilege to a specific coordinate system such as medical imaging or protein folding [188, 84]. The purpose of this thesis is to understand, insofar as it improves generalisation, whether symmetry is a good idea.

We will mostly consider the relative performance of predictors or algorithms and the risk provides a comparator. In particular, the statement f *generalises better than* f' means that $R[f] \leq R[f']$ and generalising *strictly* better means a strict inequality in the same direction. We make the same comparison for algorithms, for which we define the risk to be the risk of the returned predictor viewed as a function of the training sample. In this case the risk is stochastic and the comparison will be in expectation. The risk of a predictor or algorithm, the extent to which it *generalises*, is the only measure of quality we consider. Although important, we do not consider other aspects such as interpretability or computational efficiency.

The theoretical study of generalisation can be considered as a narrow form of mathematical epistemology, in that it offers a quantitative formulation and analysis of the problem of induction. More practically, it has the potential to offer performance guarantees that improve the reliability of machine intelligence. In addition, theoretical analyses can provide conceptual schemes and intuitions that lead to new methods.

Closely related to generalisation is learning. An algorithm with range $\mathcal{F}$ is said to *learn* the class of functions $\mathcal{F}$ if $\exists m : (0,1)^2 \to \mathbb{N}$ such that $\forall \epsilon, \delta \in (0,1)$ if $n \geq m(\epsilon, \delta)$ then for all distributions of (X, Y), with probability at least $1 - \delta$ over the training sample, the output $\hat{f}$ of the algorithm satisfies

$$R[\hat{f}] \leq \inf_{f \in \mathcal{F}} R[f] + \epsilon.$$

The pointwise minimum over all m that satisfy the above is called the *sample complexity* of the algorithm.

The above model of learning is based on PAC learning, which was originally proposed by Valiant [170]. The variation we give is due to Haussler [76] and is known as agnostic PAC learning, because it is agnostic as to the relationship between X and Y [88]. See [155, 118] for further information and bibliographic remarks.

If the sample complexity of learning is established then one has non-asymptotic control of the risk of the algorithm. In the case of binary classification, the sample complexity of learning is governed by a combinatorial measure of complexity of $\mathcal{F}$ called the VC dimension, originally from Vapnik and Chervonenkis [173]. The same idea holds for learning in other settings such as regression, just with different complexity measures, for instance the Rademacher complexity [92] or covering numbers [41]. See [11] for more.

In the case of symmetry, it seems natural to try to derive sample complexity upper bounds that reduce when $\mathcal{F}$ has the right symmetry. All prior works on the generalisation of invariant and equivariant models take some form of this approach (see Section 2.1 for a discussion) and we also give some results of this flavour in Proposition 3.11 and Theorem 6.15.[1] However, the main goal of this thesis is to provide theory that aides the practitioner in deciding whether to incorporate symmetry (either engineered or learned) into their model. For this purpose the aforementioned results are suggestive, but insufficient.

We would like to understand the *generalisation gap*

$$\Delta(f, f') = R[f] - R[f']$$

between predictors f and f'. Specifically, we are interested in the case where f' is equivariant and f is not. If $\Delta(f, f') > 0$ then f' is preferable to f. When f and f' are the outputs of algorithms the generalisation gap is a function of the training

[1] As far as we are aware, the only other work that doesn't take this approach is by Mei, Misiakiewicz, and Montanari [116], which appeared on arXiv only four days after the first work of this thesis. The topic is similar to Section 5.2, but the approach is quite different. A comparison is given in Section 7.3.

sample so is stochastic. In this setting, the complexity based results we mention above provide tail estimates for $R[f]$ and $R[f']$. However, unless these estimates turn out to be exact it is difficult even to tell the sign of $\mathbb{E}[\Delta(f, f')]$. Ultimately, the techniques of statistical learning theory are too general so we must take a different approach. We give an outline in the next section.

1.2 Overview

The technical setting and assumptions are described in Section 1.4.1. Below we give an informal overview of our main results without reference to these conditions. Throughout the work there are many examples and additional results to illustrate the utility of the approach. We make use of some standard facts which we give in Appendix A. We review related literature in Chapter 2.

Chapter 3 contains some general observations. Let $\mathcal{G}$ be a group and let $f : \mathcal{X} \to \mathcal{Y}$ be a function between two sets on which $\mathcal{G}$ acts. Define the operator

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

which takes any function and makes it equivariant. By exploring the properties of $\mathcal{Q}$, we establish in Lemma 3.1 that any function can be written as

$$f = \bar{f} + f^{\perp}$$

where $\bar{f}$ is equivariant, $\mathcal{Q}f^{\perp} = 0$ and, crucially, the two terms are L_2-orthogonal as functions.

This turns out to be a useful observation. In particular, it means that on any regression task with squared-error loss and an equivariant target the generalisation gap $\Delta(f, \bar{f})$ is exactly the squared L_2-norm of $f^{\perp}$. If f is not equivariant then this is strictly positive. In other words, for any predictor f that is not equivariant there is an equivariant predictor f' with strictly lower risk on all regression problems with squared-error loss and an equivariant target. Another view on this is that it gives a lower bound for not using an equivariant predictor. All of this applies equally to invariance.

These insights are applied in Chapter 4, where the main result quantifies the expected generalisation benefit of equivariance in a linear model. Theorem 4.7 concerns the generalisation gap $\Delta(f, f')$ in the case that $f : \mathbb{R}^d \to \mathbb{R}^k$ is the minimum-norm least squares predictor and $f' = \mathcal{Q}f$ is its equivariant version.

Let $\mathcal{G}$ 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, $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\xi^2] = \sigma^2 < \infty$. Let

$$(\chi_\psi|\chi_\phi) = \int_{\mathcal{G}} \operatorname{Tr}(\psi(g)) \operatorname{Tr}(\phi(g)) \,\mathrm{d}g$$

denote the scalar product of the characters of the representations (which are real). We show in Theorem 4.7 that the generalisation benefit of enforcing equivariance in a linear model is given by

$$\mathbb{E}[\Delta(f, f')] = \mathcal{E}(n, d) + \sigma^2 r(n, d)(dk - (\chi_\psi|\chi_\phi))$$

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 $\mathcal{E}(n, d) \geq 0$ is the generalisation gap of the corresponding noiseless problem, specified exactly in Theorem 4.7, which vanishes when $n \geq d$. The divergence at the interpolation threshold $n \in [d - 1, d + 1]$ is consistent with the literature on double descent [75].

It's worth emphasising that if f' is any other predictor such that $\mathbb{E}[R[f']] \leq \mathbb{E}[R[\mathcal{Q}f]]$ then the above result holds as a lower bound on $\mathbb{E}[\Delta(f, f')]$. In particular, this means that the above result applies to the case where f' is generated by transforming the features before training such that the least squares estimate is automatically equivariant, provided that the resulting predictor generalises at least as well as $\mathcal{Q}f$ in expectation over the training sample.

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 under $\mathcal{Q}$. 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 4.1 we provide a result that is analogous to Theorem 4.7 for invariant predictors, along with a separate proof.

In Chapter 5 we adapt and extend these results to kernel methods. We define the operator

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

which takes any function and makes it invariant. Since $\mathcal{O}$ is a special case of $\mathcal{Q}$, we have the decomposition $f = \bar{f} + f^\perp$ with $\bar{f}$ invariant, $\mathcal{O}f^\perp = 0$ and the terms being L_2-orthogonal. In Theorem 5.2 we study the generalisation gap $\Delta(f, \mathcal{O}f)$ for kernel ridge regression on an invariant target. The comments made above about the use of $\mathcal{Q}f$ as a comparator apply here to $\mathcal{O}f$.

Let X be invariant in distribution, so $X \overset{\mathrm{d}}{=} gX$ for all $g \in \mathcal{G}$, and set $Y = f^*(X) + \xi$ with f^* invariant and $\mathbb{E}[\xi] = 0$, $\mathbb{E}[\xi^2] = \sigma^2 < \infty$. Let f be the solution to kernel ridge regression with kernel k and regularisation parameter $\rho > 0$ on n i.i.d. training examples $((X_i, Y_i) : i = 1, \dots, n)$ each distributed identically to and independently from (X, Y). In Theorem 5.2, we find that

$$\mathbb{E}[\Delta(f, \mathcal{O}f)] \geq \mathcal{E}_k(n, \rho) + \sigma^2 \frac{\|(\mathrm{id} - \mathcal{O})k\|^2_{L_2(\mu \otimes \mu)}}{(\sqrt{n}M_k + \rho/\sqrt{n})^2}$$

where for any $j : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$\|j\|_{L_2(\mu\otimes\mu)}^2 = \int_{\mathcal{X}} j(x,y)^2 \,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y),$$

$M_k = \sup_x k(x,x) < \infty$ and $\mathcal{E}_k(n,\rho) \geq 0$ is the generalisation gap for the corresponding noiseless problem. By considering the linear kernel we study how this result relates to the result for invariance in Chapter 4.

Further, we show that under mild additional conditions on $\mathcal{X}$, $\mathcal{Y}$ and ρ, $\mathcal{E}_k(n,\rho) \to 0$ as $n \to \infty$ provided the kernel k satisfies the identity

$$\int_{\mathcal{G}} k(gx,y)\,\mathrm{d}g = \int_{\mathcal{G}} k(x,gy)\,\mathrm{d}g$$

for all $x, y \in \mathcal{X}$. Assuming this condition holds, we derive an independent result about the structure of the RKHS $\mathcal{H}$. In particular, Theorem 5.14 says that the above condition on the kernel implies the orthogonal decomposition $\mathcal{H} = \mathcal{O}\mathcal{H} \oplus (\mathrm{id} - \mathcal{O})\mathcal{H}$ where the orthogonality is now with respect to the inner product on $\mathcal{H}$.

In Chapter 6 we take a different approach, making use of the observation that an invariant function can be specified by its values on one representative from each orbit of $\mathcal{X}$ under $\mathcal{G}$. We show rigorously how learning a class of invariant predictors is equivalent to learning in a reduced problem in terms of orbit representatives and extend this framework to provide a new intuition for learning equivariant predictors. In addition, we show how to use these equivalences to derive a sample complexity bound for learning invariant/equivariant classes with empirical risk minimisation.

We conclude in Chapter 7 with a relation of the orbit averaging viewpoint on invariance in Chapter 3 to the orbit representative viewpoint in Chapter 6, applications to neural networks, connections to other works and, finally, some sugggestions for future work.

1.3 Authorship

The work in this thesis is based on [52, 51, 50]. The majority of content has been revised or generalised, sometimes substantially, and some new results are given. Chapters 3 and 4 and Section 7.2 are based on [52], Chapter 5 is based on [51] and Chapter 6 is based on [50].

All original contributions of the thesis were discovered and proved by the author unless explicitly stated otherwise. Where possible, Zaidi's contributions to [52] are cited explicitly. Otherwise, his contributions to [52] were through discussions, proof reading and presentation. In addition, Zaidi suggested the connection to test time augmentation described in Section 3.3.2.

A list of other work completed during this DPhil is given in Appendix B.

1.4 Preliminaries

We assume familiarity with the basic notions of group theory including the definitions of a group action and a linear representation. The reader may consult [179, 154, Chapters 1-4] for background. We run through some results that we use often in Section 1.4.3. We provide some background definitions and technical material in Section 1.4.4.

1.4.1 Setup, Assumptions and Technicalities

In this section we outline our setup and technical conditions which, unless stated otherwise, are assumed throughout. Additional definitions and assumptions are given as needed. Technical conditions are chosen so as to vary the least between results; the reader interested in more general settings is encouraged to inspect the proofs. That said, the conditions we impose are quite general.

1.4.1.1 Spaces

There will be a background probability space $(\Omega, \mathcal{S}_\Omega, \mathbb{P})$ that is assumed to be rich enough to support our analysis. The input and output spaces will be $\mathcal{X}$ and $\mathcal{Y}$ respectively, where $\mathcal{X}$ is a non-empty Polish space and $\mathcal{Y}$ is $\mathbb{R}^k$ with an inner product $\langle \cdot, \cdot \rangle : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$, induced norm $\|\cdot\|$ and corresponding topology. Sometimes $k = 1$ or this inner product is the Euclidean one, but this will be specified. All σ-algebras will be Borel. In particular this makes $(\mathcal{X}, \mathcal{S}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{S}_\mathcal{Y})$ standard Borel spaces, but we won't make direct use of this level of detail. Unless stated otherwise, X and Y will be random elements of $\mathcal{X}$ and $\mathcal{Y}$ respectively.

1.4.1.2 Group and Action

Let $\mathcal{G}$ be a measurable, second countable, Hausdorff and compact topological group.[2] Let the Haar measure on $\mathcal{G}$ be λ, normalised so that $\lambda(\mathcal{G}) = 1$. This is the unique invariant probability measure on $\mathcal{G}$. We assume that $\mathcal{G}$ has a measurable action ϕ on $\mathcal{X}$ and measurable representation ψ on $\mathcal{Y}$. By measurable we mean that $\phi : \mathcal{G} \times \mathcal{X} \to \mathcal{X}$ is a measurable map and the same for ψ. We will assume that ψ is unitary with respect to $\langle \cdot, \cdot \rangle$, by which we mean that $\langle \psi(g)a, \psi(g)b \rangle = \langle a, b \rangle \; \forall a, b \in \mathbb{R}^k$ and $\forall g \in \mathcal{G}$. 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 an orthogonal representation (one for which the $\psi(g)$ is always an orthogonal matrix). Any inner product can be transformed to be such that ψ is unitary using the Weyl trick $\langle a, b \rangle \mapsto \int_\mathcal{G} \langle \psi(g)a, \psi(g)b \rangle \, \mathrm{d}\lambda(g)$. The *character* of a representation ψ is defined by $\chi_\psi(g) = \mathrm{Tr}(\psi(g))$. The inner product of characters is defined by

$$(\chi_1 | \chi_2) = \int_\mathcal{G} \chi_1(g)\chi_2(g) \, \mathrm{d}\lambda(g).$$

This definition typically appears with a complex conjugate, but it's not needed because all the representations we encounter are real. The inner product of the

[2] The set of compact groups covers the majority of symmetries in machine learning, including all finite groups (such as permutations or reflections), many continuous groups such as rotations or translations on a bounded domain (e.g., an image) and combinations thereof.

characters of two finite-dimensional real representations of a compact group is equal to the dimension of the space of linear maps that are equivariant with respect to these representations, e.g., see [4, Theorem 3.34].

1.4.1.3 Invariance, Equivariance and Symmetry

A function $f : \mathcal{X} \to \mathcal{Y}$ is $\mathcal{G}$*-invariant* if $f(\phi(g)x) = f(x)$ and is $\mathcal{G}$*-equivariant* if $f(\phi(g)x) = \psi(g)f(x)$, each holding for all $x \in \mathcal{X}$ and for all $g \in \mathcal{G}$. Invariance is the special case of equivariance where ψ is the *trivial representation*, i.e., $\psi(g)$ is the identity for all $g \in \mathcal{G}$. A measure μ on $\mathcal{X}$ is $\mathcal{G}$*-invariant* if for all $g \in \mathcal{G}$ 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 $\phi(g)X \sim \mu$ for all $g \in \mathcal{G}$. We say that X is $\mathcal{G}$*-invariant in distribution* or just $\mathcal{G}$*-invariant* if $X \overset{\mathrm{d}}{=} gX$ for all $g \in \mathcal{G}$.

We will often drop the $\mathcal{G}$- when the group is clear from the context and just say invariant/equivariant. We will often also drop ψ and ϕ, so the quantity $g^{-1}f(gx)$ should be interpreted as $\psi(g^{-1})f(\phi(g)x)$, and so on. We sometimes use the catch-all term $\mathcal{G}$*-symmetric* to describe an object that is invariant or equivariant.

1.4.1.4 Function Space

Throughout, μ will be a $\mathcal{G}$-invariant probability measure on $(\mathcal{X}, \mathcal{S}_\mathcal{X})$. We consider $L_2(\mathcal{X}, \mathcal{Y}, \mu)$, which we define to be the Hilbert space of equivalence classes of functions $f : \mathcal{X} \to \mathcal{Y}$ such that $\|f\|_\mu < \infty$ where the norm is induced by the following inner product

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

Equality in $L_2(\mathcal{X}, \mathcal{Y}, \mu)$ is defined μ-almost-everywhere. This space is general enough to cover pretty much any predictor used in machine learning. In the case $\mathcal{Y} = \mathbb{R}$ we consider the usual L_2-space which we write $L_2(\mu)$, where the inner product on $\mathcal{Y}$ is just multiplication.

1.4.1.5 Averaging Operator

The averaging operator $\mathcal{Q}$ will be central to much of this work. It has values

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

where λ is the normalised Haar measure on $\mathcal{G}$ with $\lambda(\mathcal{G}) = 1$. The operator $\mathcal{Q}$ can be used to transform any function into an equivariant function.[3] Due to the compactness of $\mathcal{G}$, we can change variables $g \mapsto g^{-1}$ above and view $\mathcal{Q}$ as averaging the action of $\mathcal{G}$ on $L_2(\mathcal{X}, \mathcal{Y}, \mu)$ defined by $f \mapsto g \circ f \circ g^{-1}$. Our developments will apply equally to the operator

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

which is the special case of $\mathcal{Q}$ in which ψ is the trivial representation. In this case the action of $\mathcal{G}$ on $L_2(\mathcal{X}, \mathcal{Y}, \mu)$ defined above is akin to the regular representation. The operator $\mathcal{O}$ enforces invariance, a special case of equivariance. In other works both $\mathcal{O}$ and $\mathcal{Q}$ are often referred to as Reynolds operators.

1.4.2 Additional Notation

The sets $\mathbb{R}$, $\mathbb{Z}$, $\mathbb{N}$ and $\mathbb{R}_+$ denote the reals, integers, naturals and non-negative reals respectively. We use $\circ$ for function composition $(f \circ h)(x) = f(h(x))$ and we write $\mathsf{e} = \exp(1)$. For functions $f, h : \mathbb{N} \to \mathbb{R}_+$, $f = \omega(h)$ means $\exists C > 0 \; \exists m \in \mathbb{N}$ such that $\forall n > m \; f(n) > Ch(n)$ and $f = O(h)$ means $f(n) \leq Ch(n)$ under the same quantifiers.

We write $X \sim \mu$ to mean that X has distribution μ. For probability spaces $(S, \mathcal{S}, \mu)$ and $(T, \mathcal{T}, \nu)$ the product measure is denoted by $\mu \otimes \nu$, see Theorem 1.2. For random variables A and B independence is written $A \perp\!\!\!\perp B$, equality in distribution is written $A \overset{\mathrm{d}}{=} B$ and almost sure equality is written $A \overset{\text{a.s.}}{=} B$. We write i.i.d. to

[3] The operator $\mathcal{Q}$ bears similarity to the *twirl operator* in quantum computing, see [117, Section 2.1.1] and [134, Section VII.B].

stand for independent and identically distributed.

We will use the Einstein summation convention, in which repeated indices are implicitly summed over. The Kronecker tensor is written as δ_{ij}, which is 1 when $i = j$ and 0 otherwise. We write $\mathbb{G}_n(\mathbb{R}^d)$ for the Grassmannian manifold of subspaces of dimension n in $\mathbb{R}^d$.

We write id for the identity operator and any linear operator A we write its adjoint as A^*. We use $\|\cdot\|_2$ for the Euclidean norm of vectors, $\|\cdot\|_1$ for the sum of magnitudes of components and $\|\cdot\|_\infty$ for the component-wise max magnitude. On a function $f : \mathcal{X} \to \mathbb{R}^d$, $\|f\|_\infty = \sup_{x\in\mathcal{X}} \|f(x)\|_\infty$.

We write I for the identity matrix, sometimes with a subscript to indicate the dimension of the space on which it acts. We write $\mathbb{R}^{m\times n}$ for the set of all real matrices with m rows and n columns. For any matrix $A \in \mathbb{R}^{n\times n}$ we define $\|A\|_2 = \sup_{x\in\mathbb{R}^n} \frac{\|Ax\|_2}{\|x\|_2}$, which is the operator norm induced by the Euclidean norm. In general, the operator norm of any operator between normed spaces will be written with $\|\cdot\|_{\text{op}}$. For any symmetric matrix A, we denote by $\gamma_{\max}(A)$ and $\gamma_{\min}(A)$ the largest and smallest eigenvalues of A respectively. For any matrix A we write A^+ for the Moore-Penrose pseudo-inverse and $\|A\|_{\text{F}} = \sqrt{\text{Tr}(A^\top A)}$ for the Frobenius norm.

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. The group of $m \times m$ invertible real matrices is written GL_m.

1.4.3 Commonly Used Results

We will make use of the results in this section throughout the work.

The inclusion $L_2(\mu) \subset L_1(\mu)$

Theorem 1.1 ([174, Theorem 2])**.** Let $(\mathcal{X}, \mathcal{S}_\mathcal{X}, \mu)$ be a measure space with $\mu \geq 0$. Let $A_\infty = \{A \in \mathcal{S}_\mathcal{X} : \mu(A) < \infty\}$. Then $\sup_{A\in A_\infty} \mu(A) < \infty$ is equivalent to

$L_p(\mu) \subset L_q(\mu)$ for all $p > q$.

In our case μ is a probability measure, so $L_2(\mu) \subset L_1(\mu)$. We will use this fact frequently, in particular when applying Fubini's theorem.

Fubini's theorem

Theorem 1.2 (Fubini's theorem [85, Theorem 1.27])**.** For any σ-finite measure spaces $(S, \mathcal{S}, \nu_S)$ and $(T, \mathcal{T}, \nu_T)$ there exists a unique product measure $\nu = \nu_S \otimes \nu_T$ on $(S \times T, \mathcal{S} \otimes \mathcal{T})$ such that

$$\nu(A \times B) = \nu_S(A)\nu_T(B) \quad \forall A \in \mathcal{S},\ B \in \mathcal{T}.$$

Moreover, for any measurable $f : S \times T \to \mathbb{R}_+$

$$\int_{S\times T} f(s,t)\,\mathrm{d}\nu(s,t) = \int_S \mathrm{d}\nu_S(s) \int_T f(s,t)\,\mathrm{d}\nu_T(t) = \int_T \mathrm{d}\nu_T(t) \int_S f(s,t)\,\mathrm{d}\nu_S(s).$$

The above remains valid for any ν-integrable $f : S \times T \to \mathbb{R}$.

Both μ and λ are probability measures so are σ-finite. When the integrand can be negative we must verify integrability. In doing this, we will might apply Fubini's theorem to a non-negative function, but most of the time we will use Theorem 1.1.

Change of variables $g \mapsto g^{-1}$

Theorem 1.3 ([58, Corollary 2.28 and Proposition 2.31])**.** Let $\mathcal{G}$ be a measurable compact group with Haar measure λ, then $\mathrm{d}\lambda(g) = \mathrm{d}\lambda(g^{-1})$.

1.4.4 Background Theory

We list some background definitions and results. For more details see [143, 85].

1.4.4.1 Topology

A topological space is *separable* if it contains a countable dense subset.

A metric space is *complete* if every Cauchy sequence in the space converges to a limit in the space.

A *Polish space* is a separable topological space that admits a metric with respect to which it is complete.

Let (T, τ) be a topological space, then $B \subset \tau$ is a *base* for τ if any element of τ is the union of elements of B.

A topological space is *second countable* if its topology has a countable base.

A topological space is *Hausdorff* if all pairs of distinct points have disjoint open neighbourhoods.

A subset of a topological space is *compact* if every open cover has a finite sub-cover and it is *locally compact* if every point has a compact neighbourhood. Clearly, any compact set is locally compact.

A *topological group* is a group $\mathcal{G}$ equipped with topology such that the group operations are continuous. For instance $g \mapsto g^{-1}$ is continuous and $(g, h) \mapsto gh$ is continuous when $\mathcal{G}^2$ has the product topology.

1.4.4.2 Measure Theory

Let $f : A \times B \to C$ be a function. Then a *B-section* of f is a function $f_b : A \to C$ for some $b \in B$ with $f_b(a) = f(a, b)$. An *A-section* is defined similarly.

Let $(S, \mathcal{S}, \nu)$ be a measure space. We say ν is *σ-finite* if S can be written as the disjoint union of a countable family of elements of $\mathcal{S}$, each of which has finite measure. All probability measures are σ-finite.

The *Borel σ-algebra* on a topological space is the σ-algebra generated by the topology. A *Borel measure* is a measure defined on a Borel σ-algebra.

A *measurable group* is a measurable space $(\mathcal{G}, \mathcal{S}_{\mathcal{G}})$ where $\mathcal{G}$ is a topological group and $\mathcal{S}_{\mathcal{G}}$ is the Borel σ-algebra. In this case the group operations are measurable when $\mathcal{G}$ is locally compact, second countable and Hausdorff [85, p. 39].

A measure λ on $(\mathcal{G}, \mathcal{S}_{\mathcal{G}})$ is *left-invariant* if $\lambda(gA) = \lambda(A)$ for all $A \in \mathcal{S}_{\mathcal{G}}$ and for all $g \in \mathcal{G}$. The definition of *right-invariant* is the same but with $\lambda(Ag) = \lambda(A)$. The measure λ is *invariant* if it is both left-invariant and right-invariant.

A *Radon* measure is one that is finite on any compact measurable set.

Theorem 1.4 (Haar measure [85, Theorem 2.27])**.** On any locally compact, second countable and Hausdorff measurable group $\mathcal{G}$ there exists, uniquely up to normalisation, a left-invariant Radon measure λ with $\lambda \neq 0$. If $\mathcal{G}$ is compact then λ is also right-invariant.

In our setting $\mathcal{G}$ will be compact so we can normalise λ to be the unique invariant probability measure.

Chapter 2

Related Literature

In this work we are concerned with predictors that are equivariant or invariant functions. In particular we are interested in understanding their generalisation theory. We discuss works in this area first in Section 2.1, then move on to an outline of various symmetric models and their applications in Section 2.2. We end this chapter with Section 2.3, in which we point to some other notions of symmetry in machine learning.

We note that group symmetry plays a role in the nearby field of statistics. We do not discuss this, but instead refer the reader to [49, 22, 21] and references therein. Finally, in writing this literature review we found results that are either related to or are special cases of the results in this thesis. We give a comparison in Section 7.3.

2.1 Symmetry and Generalisation

The first general theoretical justification for invariance of which we are aware is from Fyfe [64] and Abu-Mostafa [3], which, roughly, states that enforcing invariance cannot increase the VC dimension of a model. Prior to that, the specific case of neural networks was addressed by John Shawe-Taylor, who calculated sample complexity bounds for neural networks whose weights can be partitioned into equivalence classes, which is directly applicable to invariant/equivariant networks [161, 159].

A heuristic argument for reduced sample complexity for invariant models is made by Mroueh, Voinea, and Poggio [120] 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 by Anselmi, Leibo, Rosasco, Mutch, Tacchetti, and Poggio [10], the authors conjecture that a general result may be obtained using wavelet transforms. A similar argument for sample complexity reduction due to translation invariance is made by Anselmi, Rosasco, and Poggio [7] and Anselmi, Leibo, Rosasco, Mutch, Tacchetti, and Poggio [9] where it is argued that the sample complexity depends on the input space, which is, in effect, significantly smaller from the perspective of a translation invariant model. This idea is prevalent in the literature and is captured rigorously in Section 6.4.

Indeed, there are many sample complexity results that are similar in spirit to those in Section 6.4, where we consider coverings on the hypothesis class and input space that reduce for invariant/equivariant hypotheses. This result, along with the rest of Chapter 6, is relatively recent work [50]. Earlier, Sokolic, Giryes, Sapiro, and Rodrigues [165] built on the work of Xu and Mannor [192] by considering classifiers that are invariant to a finite set of transformations. Their results apply to the case where the metric on the input space is the Euclidean norm contain an implicit margin constraint on the training data. Sannai, Imaizumi, and Kawano [145] prove a generalisation bound for invariant/equivariant neural networks in the case of finite groups in terms of covering numbers of the quotient space $\mathcal{X}/\mathcal{G}$. Zhu, An, and Huang [198] use a pseudo-metric that exploits the group action to express the covering number of the hypothesis class on the training set in terms of a covering number of the hypothesis class on the orbit of the training set under the group action. The work of Shao, Montasser, and Blum [156] was published concurrently with what is presented in Chapter 6 and, at least in its learning-theoretic perspective, is mostly closely related. Their main results include upper and lower sample complexity bounds for learning with invariant hypotheses on an invariant distribution in terms of an adaptation of the VC dimension that incorporates the group action

[156, Theorem 4].

At a high level, all but three works on the generalisation of invariant/equivariant models of which we are aware construct a generic generalisation bound that can be shown to reduce under the assumption that the class of predictors satisfies a symmetry. Bietti, Venturi, and Bruna [20] study the restricted setting of a discrete grid for the input space and subsets of the symmetric group for symmetries, and use spherical harmonics to derive a sample complexity upper bound for kernel ridge regression that reduces by a factor of the size of the group for invariant kernels. Notably, their setup is a special case of our condition Eq. (5.3.1) with a finite group and an inner product kernel [20, Eq. 8]. Lyle, Kwiatkowksa, and Gal [109] study the effect of symmetry on the marginal likelihood and use this for model selection. Behboodi, Cesa, and Cohen [16] consider equivariant networks in the Fourier domain to derive norm based PAC Bayes generalisation bounds. Wang, Zhu, Park, Platt, and Walters [184] study lower bounds for the generalisation error of equivariant predictors when the symmetry is misspecified to varying degrees. Tahmasebi and Jegelka [169] derive upper and lower sample complexity bounds for invariant kernel regression on manifolds.

However, all of the aforementioned works are worst-case, in the sense that they give sample complexity bounds that apply to an entire class of predictors simultaneously. Of course, this does not guarantee that the generalisation of any given predictor (at a fixed number of examples) would be improved if it were somehow replaced with an invariant/equivariant one. This issue was not resolved until [52], one of the works contributing to this thesis, wherein it was shown in an abstract setting that for any predictor that is not invariant/equivariant there is an invariant/equivariant predictor with strictly lower risk on all regression tasks with an invariant/equivariant target. Explicit calculations of this generalisation improvement were given for linear models [52, Theorem 7 and Theorem 13] and subsequently for kernel ridge regression in a separate work which also forms part of this thesis [51]. Concurrently, Mei, Misiakiewicz, and Montanari [116] analyse the asymptotic generalisation benefit of invariance in random feature models and kernel methods, providing the only other strict generalisation benefit for invariance of which we are aware. A comparison of their results to the relevant results in this thesis is given in Section 7.3.

2.2 Invariant and Equivariant Models

While there has been a recent surge in interest, symmetry and invariance have a long history in machine learning, for instance appearing in the classical work of Fukushima [63]. See [190] and references therein. The vast majority of modern implementations concern multi-layer perceptrons (MLPs) and variations of convolutional neural networks (CNNs) and our focus will reflect this. However, implementation of invariance/equivariance in other models has been explored too, as we outline below.

2.2.1 The Symmetric Zoo

Invariance and equivariance have been implemented in various models in machine learning and related areas. To name a few: steerable filters for image processing [60], invariance in signal processing [81] and scattering transforms [111, 28], kernels [70, 71, 196, 136], support vector machines [149, 29, 42, 183] and feature spaces of invariant polynomials [151, 152, 72].

In addition, recent works include equivariant Gaussian processes and neural processes [80], capsule networks [104], self attention and transformers [82, 78, 61, 62], autoencoders [193, 189], normalising flows [140, 91, 148, 25] and graph networks [113, 147, 130], whose approximation capacity has been studied by Azizian and Lelarge [14].

Methods and results also exist that can be applied to a range of models. Of course, averaging any predictor over the group with $\mathcal{O}$ or $\mathcal{Q}$ produces invariance or equivariance respectively but is often computationally infeasible. Puny et al. [130] find a computationally tractable alternative by averaging over a selected subgroup. Taking a viewpoint similar to Chapter 6, Aslan, Platt, and Sheard [13] propose to learn invariant/equivariant models by projecting the data onto a space of orbit representatives. Villar, Hogg, Storey-Fisher, Yao, and Blum-Smith [176] show how functions equivariant to physical symmetries such as rigid body transformations and the Lorentz group can be expressed solely in terms of invariant scalar functions. In a similar vein, Blum-Smith and Villar [23] show how to parameterise polynomial or smooth equivariant functions in terms of invariant ones. Finally, Anselmi, Rosasco, and Poggio [7] discuss the foundations of learning representations that are both invariant and *selective*, meaning that the representations for two inputs are equal only if the inputs belong to the same orbit.

2.2.2 Equivariant MLPs and CNNs

The design of invariant neural networks was first considered by Shawe-Taylor and Wood, in which ideas from representation theory are applied to find weight tying schemes in MLPs that result in group invariant architectures [157, 160, 158, 191]. Similarly, Ravanbakhsh, Schneider, and Póczos [138] observe that the equivariance in a linear map $\mathbb{R}^d \to \mathbb{R}^k$ implies symmetries in the corresponding weight matrix $M \in \mathbb{R}^{k\times d}$ and use this to derive weight tying schemes for equivariant neural networks. Finzi, Welling, and Wilson [56] reduce layer-wise equivariance of the network to a finite dimensional linear system that the weights must satisfy and use this to engineer equivariant MLPs, their method works even for infinite Lie groups (provided they have finite dimension). Simard, Victorri, LeCun, and Denker [162] take a different route, regularising the directional derivatives of the learned model to encourage invariance to local transformations.

For CNNs, weight tying for equivariance to 90° rotations was proposed by Dieleman, De Fauw, and Kavukcuoglu [45]. A weight-tying approach for arbitrary symmetry groups was developed for CNNs by Gens and Domingos [65] using kernel interpolation. Marcos, Volpi, Komodakis, and Tuia [112] apply convolutional filters at a range of orientations to produce rotation equivariant networks. To achieve rotation invariance, Kim, Jung, Kim, and Lee [89] map images to polar co-ordinates before performing the convolution. In addition, an equivariant version of the subsampling/upsampling operations in CNNs was proposed by Xu, Kim, Rainforth, and Teh [193].

It is well known that CNNs are equivariant to (local) translations [103] and this is widely believed to be a key factor in their generalisation performance. The G-CNN was introduced by Cohen and Welling [34] who, by viewing the standard convolutional layer as the convolution operator on feature maps over the translation group, generalised the convolutional layer to be equivariant to other groups that preserve the two dimensional lattice $\mathbb{Z}^2$.[1] Following this were many works borrowing ideas from harmonic analysis, representation theory and mathematical physics to generalise the basic approach of the G-CNN to various groups and input spaces [36, 37, 54, 94, 35, 186, 57, 187]. Esteves [53] gives an exposition of some of these models and the underlying mathematical concepts.

Following the introduction of the G-CNN, Kondor and Trivedi [95] show that any equivariant neural network layer on a homogeneous space (one for which $\forall x, y \in \mathcal{X}$ $\exists g \in \mathcal{G}$ such that $x = gy$) can be expressed in terms of a group convolution. This was elegantly generalised to non-homogeneous spaces by Portilheiro [129, Section 4.3]. Lang and Weiler [98] give a general parameterisation for the filters in these equivariant convolutions that holds for any compact group. Getting even more general, Bloem-Reddy and Teh [22] characterise the structure of invariant/equivariant probability distributions using an extension of noise outsourcing (also known as transfer) [85, Theorem 6.10] to connect functional and probabilistic symmetries.

There are results that establish, across a variety of settings, that any continuous invariant/equivariant function can be approximated to arbitrary accuracy with an invariant/equivariant neural network [194, 137, 115]. In addition, Lawrence [99] study efficient approximation of smooth functions using neural network architectures

[1] The standard convolutional layer is actually interpreted as the group *correlation* by Cohen and Welling [34], but we follow the authors in blurring this distinction.

that are invariant to subgroups of the orthogonal group. While Qin, He, and Tao [132] compare the computational complexity of two layer ReLU networks with similar equivariant networks.

There are also works specialised to the theory of learning equivariant neural networks. Bietti and Mairal [19] study the invariance, stability and sample complexity of convolutional neural networks with smooth homogeneous activation functions by embedding them in a certain reproducing kernel Hilbert space. Petersen and Sepliarskaia [127] estimate the VC dimension of G-CNNs and prove the existence of two layer G-CNNs that are invariant to an infinite group, yet have infinite VC dimension. Lawrence, Georgiev, Dienes, and Kiani [100] analyse the implicit bias of linear G-CNNs when trained by gradient descent, while Chen and Zhu [32] study the implicit bias of linear equivariant steerable CNNs trained by gradient flow.

A sub-field has emerged for learning functions on sets. We discuss permutation symmetry in more detail below but briefly mention a few related works now. Maron, Litany, Chechik, and Fetaya [114] study the learning of functions on sets whose elements themselves satisfy symmetries. Wang, Albooyeh, and Ravanbakhsh [185] construct architectures for hierarchical symmetries, such as permutation equivariant maps of sets (that themselves are invariant to permutation of their elements) and apply it to semantic segmentation of point cloud data. A transformer for use on set valued data was developed by Lee, Lee, Kim, Kosiorek, Choi, and Teh [101].

2.2.2.1 Permutation Equivariance

Particular attention has been paid to neural networks $f : \mathbb{R}^n \to \mathbb{R}^m$ that are equivariant to the action of the permutation group S_n on their inputs. The representation corresponding to $\pi \in \mathsf{S}_n$ on the inputs is the natural action $f(x_1, \dots, x_n) \mapsto f(x_{\pi^{-1}(1)}, \dots, x_{\pi^{-1}(n)})$. The most common representations on the outputs are the trivial representation, the natural representation when $m = n$, or $\operatorname{sign}(\pi)$ where here sign denotes the sign of the permutation π, which is the parity of the product of the parities of the transpositions in its decomposition thereof. These three possibilities fall under the names of *permutation invariant*, *permutation equivariant* and *fermionic* networks respectively. Fermionic networks are also known as anti-symmetric networks, which is a special case of the usage of the term $\mathcal{G}$-anti-symmetric in this work. In particular, it is straightforward to see that if f is a fermionic network, then one must have $\mathcal{O}f = 0$ if $n > 1$.[2]

Zaheer, Kottur, Ravanbakhsh, Poczos, Salakhutdinov, and Smola [195] propose the use of permutation invariant neural networks for tasks with inputs that are sets and show that any continuous permutation invariant function $f : [0,1]^n \to \mathbb{R}$ must have the form

$$f(x) = f_1\left(\sum_{i=1}^{n} f_2(x_i)\right) \tag{2.2.1}$$

for some continuous functions f_1 and f_2 with f_2 independent of f. The importance of continuity and the dimension of any intermediate latent spaces used in the computation of f (of particular relevance to neural networks) is considered by Wagstaff, Fuchs, Engelcke, Posner, and Osborne [180] and Wagstaff, Fuchs, Engelcke, Osborne, and Posner [181]. In a sense, these results say that the only permutation invariant function is summation.

There has been considerable work on the theory of permutation equivariant networks. Abrahamsen and Lin [2] and Hutter [83] consider approximating and representing fermionic functions in terms of sums of determinants respectively. Han, Li, Lin, Lu, Zhang, and Zhang [73] study approximating permutation invariant and fermionic functions. While universal approximation of permutation equivariant functions was proved by Sannai, Takai, and Cordonnier [146] and Segol and Lipman [153]. Finally, Nachum and Yehudayoff [122] study the ability of a generic two layer neural network to learn a certain class of permutation invariant functions.

One application of permutation invariant networks is point cloud modelling. Point cloud networks are permutation invariant neural networks that are designed to pre-

[2] We have $|\mathsf{S}_n| = n!$ which is even when $n > 1$. The transposition that switches the first and second elements of a list is a bijection, so must map half of the elements of S_n to the other half. Split the sum in $\mathcal{O}f$ into these two halves and the two terms cancel.

serve the shape of point clouds [131]. Point cloud networks may also be invariant or equivariant to rigid body transformations (translation and rotation) and have applications in computer vision and robotics [68]. Their approximation properties have been studied by Dym and Maron [48].

2.2.3 A Remark on the Representation of Invariant Functions

The work from Bloem-Reddy and Teh [22] which we referred to earlier offers a probabilistic generalisation of Eq. (2.2.1) in an analogous expression for random variables. Permutations are generalised to any compact group and the sum is replaced with a *maximal invariant*, which is a function M on $\mathcal{X}$ that takes a distinct constant value on each orbit of $\mathcal{X}$ under $\mathcal{G}$. That is, M is invariant and $M(x) = M(y)$ implies $gx = y$ for some $g \in \mathcal{G}$. As it happens, using the below theorem, if there exists a function φ such that $\mathcal{O}\varphi$ is a maximal invariant then a deterministic generalisation of Eq. (2.2.1) to any compact group is immediate.

Theorem 2.1 ([102, Theorem 6.2.1])**.** Let M be a maximal invariant on $\mathcal{X}$ with respect to $\mathcal{G}$. For all functions f on $\mathcal{X}$, f is $\mathcal{G}$-invariant if and only if it can be written $f(x) = h(M(x))$ for some function h.

Proof. Clearly $h(M(x))$ is invariant and if $M(x) = M(y)$ then maximality implies $x = gy$ for some $g \in \mathcal{G}$ so $f(x) = f(y)$. □

2.2.4 Applications

There are many applications of symmetric models. Most often these are in domains where the problem is known to satisfy a symmetry a priori, where experimenter choices or data representation are arbitrary and not of intrinsic significance (e.g., choice of reference frame for measurement in classical dynamics or an ordered representation of a set), or simply where symmetry is believed to be a useful inductive bias. Below we give a few examples.

Neural networks invariant to Galilean transformations have been used for predicting the Reynolds stress anisotropy tensor in fluid dynamics [106]. Rotation invariant neural networks have been applied to galaxy morphology prediction by Dieleman, Willett, and Dambre [46] and for learning energy potentials of molecular systems by Anderson, Hy, and Kondor [6].

In particle physics, Bogatskiy, Anderson, Offermann, Roussi, Miller, and Kondor [24] apply Lorentz group equivariant neural networks to tagging top quark decays and proton-proton collisions. For quantum mechanical systems, invariant neural networks have been used to model ground state wave functions and dynamics [107, 108] and fermionic networks have been applied to quantum chemistry [128].

In the life sciences, equivariant attention was applied to protein structure prediction by Jumper et al. [84] and permutation equivariant networks applied to protein-drug binding by Hartford, Graham, Leyton-Brown, and Ravanbakhsh [74]. While, in medicine, rotation invariant networks have been applied to image analysis [17] and G-CNNs to pulmonary nodule detection [188].

Rahme, Jelassi, Bruna, and Weinberg [135] and Qin, He, Shi, Huang, and Tao [133] propose and study the application of permutation equivariant networks to auction design. Finally, Duan, Ma, and Deng [47] study the benefit of permutation equivariance when predicting game theoretic equilibria.

2.2.5 Learning Symmetries from Data

For all of the works we mentioned previously, the symmetry of the task must be known in advance of designing the model. In some circumstances this may not be possible, in which case learning the symmetry from data is a natural approach. In addition, by using this approach the model could uncover symmetries in the system unknown to the user or even refute hypothesised symmetries which it would otherwise be constrained to represent if hard-coded. Eliminating a stage of the design process using data is an exciting prospect. We mention some examples of work in this area below.

Anselmi, Evangelopoulos, Rosasco, and Poggio [8] outline principles for learning symmetries of data and learning equivariant representations without explicit knowledge of the symmetry group. In particular, they propose to learn symmetries from data using a regularisation scheme. In addition, Desai, Nachman, and Thaler [44] use a generative adversarial network to discover symmetries in data and Christie and Aston [33] develop statistical tests for equivariance.

Benton, Finzi, Izmailov, and Wilson [18] use data augmentation and learn a parameterised group of transformations in conjunction with the model parameters to learn partial invariances from data (e.g., invariance to a connected subset of SO_2). van der Wilk, Bauer, John, and Hensman [171] learn parameterised invariances in Gaussian processes through their effect on the marginal likelihood. In an adjacent area, Cubuk, Zoph, Mané, Vasudevan, and Le [40] propose the automatic learning of data augmentation policies.

Dehmamy, Walters, Liu, Wang, and Yu [43] learn Lie group symmetries in a G-CNN by parametrising the filters in terms of the basis of the Lie algebra. Learning partial equivariances in G-CNNs is studied by Romero and Lohit [141]. In addition, Zhou, Knowles, and Finn [197] propose a meta-learning approach to learn weight sharing patterns that produce equivariant neural networks. Finally, addressing the theoretical limitations of this general approach, Portilheiro [129] shows, by considering aliasing between symmetries, that under certain conditions it is impossible to simultaneously learn group symmetries and functions equivariant with respect to them using an equivariant ansatz.

2.3 Other Forms of Symmetry in Machine Learning

General ideas of symmetry appear in many places in machine learning, but not necessarily in the form of invariance/equivariance of the predictor. These areas are outside of the topic of this thesis, but we mention a few examples.

Geometric Deep Learning The study of invariant/equivariant models can be placed inside in the broader sub-field of *geometric deep learning*, that uses ideas from geometry, algebra and graph theory to enforce inductive biases [26, 27, 66].

Conservation Laws Neural networks have been used to parameterise Hamiltonians and Lagrangians for models of physical systems, automatically enforcing conservation laws [67, 39]. For each conserved quantity there is a symmetry and vice versa by Noether's theorem.

Passive Symmetries Villar, Hogg, Yao, Kevrekidis, and Schölkopf [177] highlight the potential importance of recognising symmetries that arise in tasks due to arbitrary experimenter choices, for instance when collecting or processing data. This is particularly important when the fundamental properties of a problem are independent of the co-ordinate description used in computation. In a similar vein, Villar, Yao, Hogg, Blum-Smith, and Dumitrascu [175] study the importance of *units equivariance*, meaning equivariance of the model to the choice of units used to represent the data.

Algorithmic Symmetry Abbe and Boix-Adserà [1] exploit *algorithmic symmetries*, symmetries of the training algorithm when viewed as a mapping from the training sample to a predictor, to prove limitations on what can be learned by neural networks trained by noisy gradient descent. This is a generalisation of the orthogonal equivariance of gradient descent which was exploited in the famous work of Ng [123] to argue for the LASSO over the ridge for regularisation in logistic regression. The orthogonal equivariance property was also applied in Li, Zhang, and Arora [105] to demonstrate a separation in sample complexity between convolutional and fully connected networks.

Weight Space Symmetries in Neural Networks It is well known that the description in terms of the parameters of the function computed by standard neural network architectures is degenerate. That is, the map from ordered lists of weights to functions is not injective. As far as we are aware, this fact is yet to make a significant impact on the mainstream theory of neural networks. In any case, list a few works in this area. Chen, Lu, and Hecht-Nielsen [30] analyse the group theoretical properties of a class of output-preserving weight transformations. Rüger and Ossen [144] derive a metric on the weight space of MLPs that removes the permutation symmetries. Simsek et al. [164] study the effect of weight-space permutation symmetries on the loss landscape. Kunin, Sagastuy-Brena, Ganguli, Yamins, and Tanaka [97] examine the effect of weight space symmetries on optimisation.

Data Augmentation Data augmentation is a procedure in which the covariates are replaced by their image under a random transformation. If the transformations are sampled uniformly from a compact group then in expectation (over the transformations only) data augmentation replaces the training loss function ℓ with $\mathcal{O}\ell$, where for the purposes of the averaging it is viewed as a function of the covariates. Data augmentation has been shown to help learn transformation invariances from data [55]. Moreover, Kashyap, Subramanyam, et al. [87] show that robustness to input transformations, a form of approximate invariance (although not always with respect to a group), is a strong predictor of generalisation performance in deep learning. From a theoretical perspective, Lyle, van der Wilk, Kwiatkowska, Gal, and Bloem-Reddy [110] derive a PAC Bayes bound and use it to study the relative benefits of feature averaging and data augmentation. Chen, Dobriban, and Lee [31] study the statistical properties of estimators using data augmentation when the transformations form a group. Shao, Montasser, and Blum [156] use the framework of PAC learning and some adaptations of the VC dimension to quantify the effect of data augmentation on the sample complexity of learning with empirical risk minimisation.

Chapter 3

General Theory I: Averaging Operators

Summary

We explore symmetry from a functional perspective, finding in Lemma 3.1 that any function f can be written

$$f = \bar{f} + f^{\perp}$$

where $\bar{f}$ is equivariant, $f^{\perp}$ represents the $\mathcal{G}$-anti-symmetric (non-equivariant) part of f and, most importantly, the two terms are orthogonal as functions. As a warm up, we use this result to give some new perspectives on feature averaging. We then apply it to generalisation in Lemma 3.12, deriving a strict generalisation benefit for equivariant predictors.

3.1 Averaging and the Structure of $L_2(\mathcal{X}, \mathcal{Y}, \mu)$

The following result shows that any function in $L_2(\mathcal{X}, \mathcal{Y}, \mu)$ can be decomposed orthogonally into two terms that we call its $\mathcal{G}$-symmetric and $\mathcal{G}$-anti-symmetric parts. Recall that since $\mathcal{O}$ is just a special case of $\mathcal{Q}$, Lemma 3.1 applies to both operators.

Lemma 3.1. Let U be any subspace of $L_2(\mathcal{X}, \mathcal{Y}, \mu)$ that is closed under $\mathcal{Q}$, meaning $\mathcal{Q}U \subset U$. Define the subspaces S and A of U consisting of the $\mathcal{G}$-symmetric and $\mathcal{G}$-anti-symmetric functions respectively, $S = \{f \in U : f \text{ is } \mathcal{G}\text{-equivariant}\}$ and $A = \{f \in U : \mathcal{Q}f = 0\}$. Then U admits the orthogonal decomposition $U = S \oplus A$.

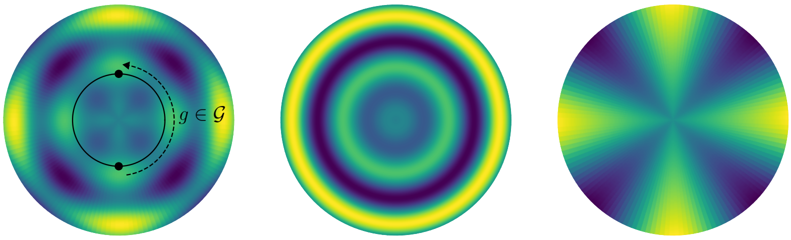
Although discovered independently, Lemma 3.1 is not a fundamentally new idea. The same theme appears in Reisert and Burkhardt [139] and likely many other works. A proof is given in Section 3.1.1 which establishes that $\mathcal{Q}$ is a self-adjoint orthogonal projection on $L_2(\mathcal{X}, \mathcal{Y}, \mu)$, from which the conclusion follows.

Lemma 3.1 says that any function $u \in U$ can be written $u = s + a$, where s is equivariant, $\mathcal{Q}a = 0$ and $\langle s, a\rangle_\mu = 0$. We refer to s and a as the $\mathcal{G}$-symmetric and $\mathcal{G}$-anti-symmetric parts of u respectively. In general this does not imply that a is an odd function, that it outputs an anti-symmetric matrix or that its values are negated by swapping arguments. These are, however, special cases. If $\mathcal{G} = \mathsf{C}_2$ acts by $x \mapsto -x$ then odd functions $f : \mathbb{R} \to \mathbb{R}$ will be C_2-anti-symmetric. If $\mathcal{G} = \mathsf{C}_2$ acts on matrices by $M \mapsto M^\top$ then $f : M \mapsto \frac{1}{2}(M - M^\top)$ is also C_2-anti-symmetric, but with respect to a different action. Finally, if $\mathcal{G} = \mathsf{S}_n$ and $f : \mathbb{R}^n \to \mathbb{R}$ with $f(x_1, \ldots, x_j, x_{j+1}, \ldots, x_n) = -f(x_1, \ldots, x_{j+1}, x_j, \ldots, x_n)$ then f is S_n-anti-symmetric.

Example 3.2. Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \mathbb{R}$, $\mu = \mathcal{N}(0, I_2)$ and set $V = L_2(\mu)$. Let $\mathcal{G} = \mathsf{SO}_2$ act by rotation about the origin, with respect to which the normal distribution is invariant. Using Lemma 3.1 we may write $V = S \oplus A$. Alternatively, consider polar coordinates (r, θ), then for all $f \in V$ we have $\mathcal{O}f(r, \theta) = \frac{1}{2\pi}\int_0^{2\pi} f(r, \theta')\,\mathrm{d}\theta'$.[1] So, naturally, any invariant f 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 all r, and A consists entirely of such functions. For example, $r^3 \cos\theta \in A$. We then recover $\langle s, h\rangle_\mu = \frac{1}{2\pi}\int_{\mathcal{X}} s(r)h(r, \theta)\mathsf{e}^{-r^2/2} r\,\mathrm{d}r\,\mathrm{d}\theta = 0$ for all $s \in S$ by integrating h over θ.

[1] This is to be interpreted informally as, strictly, evaluation is not defined. The existence of $\mathcal{O}f$ is guaranteed by Proposition 3.4 so the integral is finite for almost all r and the rest are harmless.

Intuitively, one can think of the functions in S as only 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. 3.1.



$f \in V$ $\mathcal{O}f \in S$ $f - \mathcal{O}f \in A$

Figure 3.1: Example of a function decomposition. The figure shows $f(r, \theta) = r\cos(r - 2\theta)\cos(r + 2\theta)$ decomposed into its $\mathcal{G}$-symmetric and $\mathcal{G}$-anti-symmetric parts in $V = S \oplus A$ under the natural action of $\mathcal{G} = \mathsf{SO}_2$ on $\mathbb{R}^2$. See Example 3.2. Image credit: Sheheryar Zaidi. Best viewed in colour.

Remark 3.3. Villar, Yao, Hogg, Blum-Smith, and Dumitrascu [175] discuss generalising Lemma 3.1 to the non-compact group of scalings $x \mapsto cx$ and demonstrate the impossibility of doing so for real (as opposed to complex) c.

3.1.1 Proof of Lemma 3.1

In this section we derive the following result.

Lemma (Lemma 3.1)**.** Let U be any subspace of $L_2(\mathcal{X}, \mathcal{Y}, \mu)$ that is closed under $\mathcal{Q}$, meaning $\mathcal{Q}U \subset U$. Define the subspaces S and A of U consisting of the $\mathcal{G}$-symmetric and $\mathcal{G}$-anti-symmetric functions respectively, $S = \{f \in U : f \text{ is } \mathcal{G}\text{-equivariant}\}$ and $A = \{f \in U : \mathcal{Q}f = 0\}$. Then U admits the orthogonal decomposition $U = S \oplus A$.

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

Proposition 3.4. Let $f \in L_2(\mathcal{X}, \mathcal{Y}, \mu)$, then

1. $\mathcal{Q}f$ is μ-measurable, and
2. $\mathcal{Q}f \in L_2(\mathcal{X}, \mathcal{Y}, \mu)$ with $\|\mathcal{Q}f\|_\mu \leq \|f\|_\mu$.

Proof. The mapping $m : (g, x) \mapsto g^{-1}f(gx) \in \mathcal{Y}$ is $(\lambda \otimes \mu)$-measurable by considering the following composition and applying Lemma A.1.1

$$(g, x) \mapsto (g, gx) \mapsto (g, f(gx)) \mapsto g^{-1} f(gx).$$

The lemma requires that each co-ordinate of each tuple is generated by a measurable mapping of the previous tuple, which is true either by assumption or trivially for $(g, x) \mapsto g$.

Lemma A.1.1 also allows us to verify the μ-measurability of $\mathcal{Q}f$ component-wise. By Corollary A.1.4, it's sufficient to verify that each component of m is $(\lambda \otimes \mu)$-integrable. Consider, using Fubini's theorem, the unitarity of ψ and invariance of μ

$$\begin{aligned}
\int_{\mathcal{G}} \int_{\mathcal{X}} \|m(g, x)\|^2 \, \mathrm{d}\mu(x) \, \mathrm{d}\lambda(g) &= \int_{\mathcal{G}} \int_{\mathcal{X}} \|g^{-1} f(gx)\|^2 \, \mathrm{d}\mu(x) \, \mathrm{d}\lambda(g) \\
&= \int_{\mathcal{G}} \int_{\mathcal{X}} \|f(gx)\|^2 \, \mathrm{d}\mu(x) \, \mathrm{d}\lambda(g) \\
&= \int_{\mathcal{X}} \|f(x)\|^2 \, \mathrm{d}\mu(x) \\
&= \|f\|_{\mu}^2 < \infty
\end{aligned}$$

so that each component of m is in $L_2(\lambda \otimes \mu)$. Theorem 1.1 gives $L_2(\lambda \otimes \mu) \subset L_1(\lambda \otimes \mu)$ because $\lambda \otimes \mu$ is bounded.

So far we have established the first assertion in the statement, we now verify that $\mathcal{Q}f \in L_2(\mathcal{X}, \mathcal{Y}, \mu)$. In Eq. (a) we use Fubini's theorem which we will justify at the end of the proof, in Eq. (b) we use unitarity and in Eq. (c) we use the invariance of

μ and λ

$$\begin{aligned}
\|\mathcal{Q}f\|_\mu^2 &= \int_\mathcal{X} \left\langle \int_\mathcal{G} g_1^{-1} f(g_1 x)\,\mathrm{d}\lambda(g_1), \int_\mathcal{G} g_2^{-1} f(g_2 x)\,\mathrm{d}\lambda(g_2) \right\rangle \mathrm{d}\mu(x) \\
&= \int_\mathcal{X} \int_\mathcal{G} \big\langle g_1^{-1} f(g_1 x), g_2^{-1} f(g_2 x) \big\rangle \,\mathrm{d}\lambda(g_1)\,\mathrm{d}\lambda(g_2)\,\mathrm{d}\mu(x) \\
&= \int_\mathcal{G} \int_\mathcal{X} \big\langle g_1^{-1} f(g_1 x), g_2^{-1} f(g_2 x) \big\rangle \,\mathrm{d}\mu(x)\,\mathrm{d}\lambda(g_1)\,\mathrm{d}\lambda(g_2) && \text{(a)} \\
&= \int_\mathcal{G} \int_\mathcal{X} \big\langle f(g_1 x), (g_2 g_1^{-1})^{-1} f(g_2 x) \big\rangle \,\mathrm{d}\mu(x)\,\mathrm{d}\lambda(g_1)\,\mathrm{d}\lambda(g_2) && \text{(b)} \\
&= \int_\mathcal{G} \int_\mathcal{X} \big\langle f(x), g^{-1} f(g x) \big\rangle \,\mathrm{d}\mu(x)\,\mathrm{d}\lambda(g) && \text{(c)} \\
&= \int_\mathcal{G} \langle f, g^{-1} \circ f \circ g \rangle_\mu \,\mathrm{d}\lambda(g) \\
&\le \|f\|_\mu \int_\mathcal{G} \|g^{-1} \circ f \circ g\|_\mu \,\mathrm{d}\lambda(g).
\end{aligned}$$

It remains to show that $\|g^{-1} \circ f \circ g\|_\mu = \|f\|_\mu$. For all $g \in \mathcal{G}$,

$$\begin{aligned}
\|g^{-1} \circ f \circ g\|_\mu^2 &= \int_\mathcal{X} \langle g^{-1} f(gx), g^{-1} f(gx) \rangle \,\mathrm{d}\mu(x) \\
&= \int_\mathcal{X} \langle f(gx), f(gx) \rangle \,\mathrm{d}\mu(x) \\
&= \int_\mathcal{X} \langle f(x), f(x) \rangle \,\mathrm{d}\mu(x) \\
&= \|f\|_\mu^2.
\end{aligned}$$

We now justify Eq. (a). To apply Fubini's theorem we need each $\mathcal{G}$-section of the integrand to be $(\lambda \otimes \mu)$-integrable, for which Proposition 3.5 is sufficient. □

Proposition 3.5. For all $h, f \in L_2(\mathcal{X}, \mathcal{Y}, \mu)$ the function $\langle h(x), g^{-1} f(gx) \rangle$ is $(\lambda \otimes \mu)$-integrable.

Proof. The function is measurable by expanding the inner product in a basis. Then by Cauchy-Schwarz and $a^2 + b^2 \geq ab$

$$\begin{aligned}
&\int_{\mathcal{G}} \int_{\mathcal{X}} |\langle h(x), g^{-1} f(gx) \rangle| \, \mathrm{d}\mu(x) \, \mathrm{d}\lambda(g) \\
&\leq \int_{\mathcal{G}} \int_{\mathcal{X}} \|h(x)\| \|g^{-1} f(gx)\| \, \mathrm{d}\mu(x) \, \mathrm{d}\lambda(g) \\
&\leq \int_{\mathcal{G}} \int_{\mathcal{X}} \|h(x)\|^2 \, \mathrm{d}\mu(x) \, \mathrm{d}\lambda(g) + \int_{\mathcal{G}} \int_{\mathcal{X}} \|g^{-1} f(gx)\|^2 \, \mathrm{d}\mu(x) \, \mathrm{d}\lambda(g) \\
&= \|h\|_\mu^2 + \|f\|_\mu^2 \\
&< \infty
\end{aligned}$$

where in the last line we use unitarity and the invariance of μ. □

Now for the important observation that $\mathcal{Q}$ identifies equivariance, as well as enforcing it. If there are any equivariant functions in $L_2(\mathcal{X}, \mathcal{Y}, \mu)$ then in combination with Proposition 3.4 this shows that $\mathcal{Q}$ has unit operator norm.

Proposition 3.6. *f is equivariant if and only if $\mathcal{Q}f = f$.*

Proof. Recall that f is equivariant if $f(gx) = gf(x)$ for all $g \in \mathcal{G}$ and all $x \in \mathcal{X}$. Suppose f is equivariant then for all $x \in \mathcal{X}$

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

Now assume that $\mathcal{Q}f = f$, so $f(x) = \int_{\mathcal{G}} g^{-1} f(gx) \, \mathrm{d}\lambda(g)$ for all $x \in \mathcal{X}$. Take any $\tilde{g} \in \mathcal{G}$, then

$$\begin{aligned}
f(\tilde{g}x) &= \int_{\mathcal{G}} g^{-1} f(g\tilde{g}x) \, \mathrm{d}\lambda(g) = \tilde{g} \int_{\mathcal{G}} (g\tilde{g})^{-1} f(g\tilde{g}x) \, \mathrm{d}\lambda(g) = \tilde{g} \int_{\mathcal{G}} g^{-1} f(gx) \, \mathrm{d}\lambda(g) \\
&= \tilde{g} f(x)
\end{aligned}$$

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

By Proposition 3.6, $\mathcal{Q}^2 f = \mathcal{Q}f$ so $\mathcal{Q}$ is a projection. The operator $\mathcal{Q}$ is also linear. Let U be a subspace of $L_2(\mathcal{X}, \mathcal{Y}, \mu)$ such that $\mathcal{Q}U \subset U$. Set $S = \mathcal{Q}U$ and $A =$

$(\mathrm{id} - \mathcal{Q})U$. $S \subset U$, $A \subset U$ and $S \cap A$ is trivial, so $S \oplus A \subset U$. Any $f \in U$ can be written uniquely as $f = \bar{f} + f^{\perp}$ where $\bar{f} = \mathcal{Q}f$ and $f^{\perp} = f - \mathcal{Q}f$, so $U \subset S \oplus A$. Hence $U = S \oplus A$. Proposition 3.6 gives $S = \{f \in U : f \text{ is } \mathcal{G}\text{-equivariant}\}$ and $A = \{f \in U : \mathcal{Q}f = 0\}$ is easily established: $\mathcal{Q}A = \{0\}$ by linearity and idempotence, while any f with $\mathcal{Q}f = 0$ has $f = (\mathrm{id} - \mathcal{Q})f$. Next we show that $\mathcal{Q}$ is self-adjoint with respect to $\langle \cdot, \cdot \rangle_{\mu}$ which shows that $\mathcal{Q}$ is an orthogonal projection. The orthogonality in Lemma 3.1 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 3.7. $\mathcal{Q}$ is self-adjoint with respect to $\langle \cdot, \cdot \rangle_{\mu}$.

Proof. At various points we apply Fubini's theorem, the unitarity of ψ, the invariance of μ and use the change of variables $g \mapsto g^{-1}$. The application of Fubini's theorem is valid by Proposition 3.5. The inner product on $\mathcal{Y}$ commutes with integration (e.g., by expanding in a basis). Let $f, h \in L_2(\mathcal{X}, \mathcal{Y}, \mu)$, then

$$\begin{aligned}
\langle \mathcal{Q}f, h \rangle_{\mu} &= \int_{\mathcal{X}} \left\langle \int_{\mathcal{G}} g^{-1} f(gx) \, \mathrm{d}\lambda(g), h(x) \right\rangle \mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} \int_{\mathcal{G}} \langle g^{-1} f(gx), h(x) \rangle \, \mathrm{d}\lambda(g) \, \mathrm{d}\mu(x) \\
&= \int_{\mathcal{G}} \int_{\mathcal{X}} \langle g^{-1} f(gx), h(x) \rangle \, \mathrm{d}\mu(x) \, \mathrm{d}\lambda(g) \\
&= \int_{\mathcal{G}} \int_{\mathcal{X}} \langle f(x), g h(g^{-1}x) \rangle \, \mathrm{d}\mu(x) \, \mathrm{d}\lambda(g) \\
&= \int_{\mathcal{X}} \left\langle f(x), \int_{\mathcal{G}} g h(g^{-1}x) \, \mathrm{d}\lambda(g) \right\rangle \mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} \left\langle f(x), \int_{\mathcal{G}} g^{-1} h(gx) \, \mathrm{d}\lambda(g) \right\rangle \mathrm{d}\mu(x) \\
&= \langle f, \mathcal{Q}h \rangle_{\mu}.
\end{aligned}$$

□

3.1.2 The invariance of μ

The invariance of μ is used throughout the proof of Lemma 3.1. There are many tasks for which invariance of the input distribution is a natural assumption, for instance in medical imaging [188], but the reader may wonder whether it is necessary for our results. In the following example from Sheheryar Zaidi, invariance is equivalent to the orthogonality in the decomposition $L_2(\mathcal{X}, \mathcal{Y}, \mu) = S \oplus A$. This means invariance is necessary for Lemma 3.1 to hold in general. An alternative phrasing is that the projection $\mathcal{Q}$ is only orthogonal in the below setting if μ is invariant. Note that in the example below ψ is the trivial representation.

Example 3.8 (Sheheryar Zaidi [52])**.** Let C_2 act on $\mathcal{X} = \{-1, 1\} \times \{1\}$ by multiplication on the first coordinate. Let V be the vector space of functions $f_{(t_1,t_2)} : \mathcal{X} \to \mathbb{R}$ with $f_{(t_1,t_2)}(x_1, x_2) = t_1x_1 + t_2x_2$ for all $t_1, t_2 \in \mathbb{R}$. We note that any distribution μ on $\mathcal{X}$ can be described by its probability mass function p, which is defined by $p(-1, 1)$ and $p(1, 1)$. Moreover, μ is C_2-invariant if and only if $p(-1, 1) = p(1, 1)$. Next, observe that the $\mathcal{G}$-symmetric functions S and $\mathcal{G}$-anti-symmetric functions A are precisely those for which $t_1 = 0$ and $t_2 = 0$ respectively. The inner product induced by μ is given by $\langle f_{(a_1,a_2)}, f_{(b_1,b_2)} \rangle_\mu = (a_1+a_2)(b_1+b_2)p(1, 1)+(a_2-a_1)(b_2-b_1)p(-1, 1)$. With this, we see that the inner product $\langle f_{(0,t_2)}, f_{(t_1,0)} \rangle_\mu = t_1t_2p(1, 1) - t_1t_2p(-1, 1)$ is zero for all $f_{(0,t_2)} \in S$ and $f_{(t_1,0)} \in A$ if and only if $p(-1, 1) = p(1, 1)$. That is, if and only if μ is invariant.

3.2 Warm Up

In this section we consider some basic consequences of Lemma 3.1 and of our setup more generally. Recall the special case of $\mathcal{Q}$ where ψ is the trivial representation, corresponding to invariance rather than equivariance,

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

3.2.1 Feature Averaging as a Least Squares Problem

When f is a feature extractor, e.g., the final layer representations of a neural network, $\mathcal{O}$ can be thought of as performing feature averaging. Using Lemma 3.1, feature averaging can be viewed as solving a least squares problem in $L_2(\mathcal{X}, \mathcal{Y}, \mu)$. That is, feature averaging sends f to $\bar{f}$ where $\bar{f}$ is the closest invariant feature extractor to f. This just a well known fact about orthogonal projections written in different terms, but we give a proof for completeness. The same result holds for $\mathcal{Q}$.

Proposition 3.9. Define S and A as in Lemma 3.1. For all $f \in L_2(\mathcal{X}, \mathcal{Y}, \mu)$, feature averaging with $\mathcal{O}$ maps $f \mapsto \bar{f}$ where $\bar{f}$ is the unique solution to the least squares problem $\bar{f} = \operatorname{argmin}_{s \in S} \|f - s\|_\mu^2$.

Proof. By Lemma 3.1 we can write $f = \bar{f} + f^\perp$ where $\bar{f} = \mathcal{O}f \in S$, $f^\perp \in A$ and the two terms are orthogonal. Take any $h \in S$, then using the orthogonality

$$\|f - h\|_\mu^2 = \|(\bar{f} - h) + f^\perp\|_\mu^2 = \|\bar{f} - h\|_\mu^2 + \|f^\perp\|_\mu^2 \geq \|f^\perp\|_\mu^2 = \|f - \bar{f}\|_\mu^2.$$

Set $s = \bar{f}$ and 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, using Cauchy-Schwarz, that

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

which a contradiction unless $\|f - s\|_\mu^2 = 0$, in which case $f = s$ in $L_2(\mathcal{X}, \mathcal{Y}, \mu)$. □

Example 3.10. Consider again the setting of Example 3.2. 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 $s \in S$ we can calculate:

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

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

3.2.2 Averaging and the Rademacher Complexity

Let $T = (x_1, \dots, x_n)$ be a collection of points from $\mathcal{X}$. The *empirical Rademacher complexity* of a set $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$ evaluated on T is defined by

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

where the expectation is over the random variables $\varsigma_i \sim \operatorname{Unif}\{-1,1\}$ which follow the Rademacher distribution. If the T is random then the empirical Rademacher complexity $\widehat{\mathfrak{R}}_T(\mathcal{F})$ is a random quantity and the *Rademacher complexity* $\mathfrak{R}_n(\mathcal{F})$ is defined by $\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}[\widehat{\mathfrak{R}}_T(\mathcal{F})]$. The Rademacher complexity appears in the study of generalisation in statistical learning, for instance see [182, Theorem 4.10 and Proposition 4.12] for upper and lower bounds respectively. A reduction in the Rademacher complexity means that fewer examples are required to achieve a specified worst-case risk.

Let $\mathcal{F} \subset L_2(\mu)$ and consider its $\mathcal{G}$-symmetric and $\mathcal{G}$-anti-symmetric projections $\overline{\mathcal{F}} = \mathcal{O}\mathcal{F}$ and $\mathcal{F}_\perp = (\mathrm{id} - \mathcal{O})\mathcal{F}$ respectively. We assume that both versions of the Rademacher complexity are well-defined on $\mathcal{F}$, $\overline{\mathcal{F}}$ and $\mathcal{F}_\perp$.

Proposition 3.11. The Rademacher complexity of the feature averaged class satisfies

$$0 \le \mathfrak{R}_n(\mathcal{F}) - \mathfrak{R}_n(\overline{\mathcal{F}}) \le \mathfrak{R}_n(\mathcal{F}_\perp)$$

whenever the terms are finite and where the data are distributed $T \sim \otimes^n \mu$.

Proof. We will show that $\mathfrak{R}_n(\overline{\mathcal{F}}) \leq \mathfrak{R}_n(\mathcal{F}) \leq \mathfrak{R}_n(\overline{\mathcal{F}}) + \mathfrak{R}_n(\mathcal{F}_\perp)$, from which the proposition follows immediately. We start by establishing $\mathfrak{R}_n(\overline{\mathcal{F}}) \leq \mathfrak{R}_n(\mathcal{F})$. The action of $\mathcal{G}$ on $\mathcal{X}$ induces an action on $\mathcal{X}^n$ by $g(x_1, \dots, x_n) = (gx_1, \dots, gx_n)$ under which $\otimes^n \mu$ is invariant. Let $X_1, \dots, X_n \sim \mu$, then note that

$$\begin{aligned}
\mathfrak{R}_n(\overline{\mathcal{F}}) &= \frac{1}{n} \mathbb{E}\left[\sup_{\bar{f} \in \overline{\mathcal{F}}} \left| \sum_{i=1}^n \varsigma_i \bar{f}(X_i) \right|\right] \\
&= \frac{1}{n} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varsigma_i \int_{\mathcal{G}} f(gX_i)\, \mathrm{d}\lambda(g) \right|\right] \\
&\leq \frac{1}{n} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \int_{\mathcal{G}} \left| \sum_{i=1}^n \varsigma_i f(gX_i) \right| \mathrm{d}\lambda(g)\right].
\end{aligned}$$

We obtain $\mathfrak{R}_n(\overline{\mathcal{F}}) \leq \mathfrak{R}_n(\mathcal{F})$ by the fact that the last line above is dominated by Eq. (b) below

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{F}) &= \frac{1}{n} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varsigma_i f(X_i) \right|\right] \\
&= \frac{1}{n} \int_{\mathcal{G}} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varsigma_i f(gX_i) \right|\right] \mathrm{d}\lambda(g) \qquad \text{(a)} \\
&= \frac{1}{n} \mathbb{E}\left[\int_{\mathcal{G}} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varsigma_i f(gX_i) \right| \mathrm{d}\lambda(g)\right]. \qquad \text{(b)}
\end{aligned}$$

In Eq. (a) we used the invariance of μ. In Eq. (b) we use Fubini's theorem which also guarantees that the expression is well-defined (although possibly infinite).

We now prove that $\mathfrak{R}_n(\mathcal{F}) \leq \mathfrak{R}_n(\overline{\mathcal{F}}) + \mathfrak{R}_n(\mathcal{F}_\perp)$. For any $f \in \mathcal{F}$ we can write $f = \bar{f} + f^\perp$ where $\bar{f} \in \overline{\mathcal{F}}$ and $f^\perp \in \mathcal{F}_\perp$ by Lemma 3.1. The result follows from taking expectations over $T \sim \otimes^n \mu$ of the below. For any $T = (x_1, \dots, x_n)$

$$\begin{aligned}\widehat{\mathfrak{R}}_T(\mathcal{F}) &= \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varsigma_i f(x_i)\right|\right] \\ &= \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n (\varsigma_i \bar{f}(x_i) + \varsigma_i f^\perp(x_i))\right|\right] \\ &\leq \frac{1}{n}\mathbb{E}\left[\sup_{\bar{f}\in\overline{\mathcal{F}}}\left|\sum_{i=1}^n \varsigma_i \bar{f}(x_i)\right|\right] + \frac{1}{n}\mathbb{E}\left[\sup_{f^\perp\in\mathcal{F}_\perp}\left|\sum_{i=1}^n \varsigma_i f^\perp(x_i)\right|\right] \\ &= \widehat{\mathfrak{R}}_T(\overline{\mathcal{F}}) + \widehat{\mathfrak{R}}_T(\mathcal{F}_\perp)\end{aligned}$$

□

Proposition 3.11 says that the Rademacher complexity is reduced by orbit averaging, but not by more than the complexity of the $\mathcal{G}$-anti-symmetric component of the class. This quantifies the improvement in worst-case generalisation from enforcing invariance by averaging in terms of the extent to which the inductive bias is already present in the function class. We provide stronger results in later sections and chapters, studying individual predictors and the average case for algorithms.

3.3 Implications for Generalisation

3.3.1 The Generalisation Gap in Regression Problems

In this section we apply Lemma 3.1 to derive a strict (i.e., non-zero) generalisation gap between predictors that have and have not been specified to have the symmetry that is present in the task.

Given some pair of random elements (X, Y) with values on $\mathcal{X} \times \mathcal{Y}$ defining a supervised learning task (with inputs in $\mathcal{X}$ and outputs in $\mathcal{Y}$), we define the *risk* of a predictor f as

$$R[f] = \mathbb{E}[\|f(X) - Y\|^2].$$

This is the same definition of the risk, just specialised to the loss function $\ell(y, y') = \|y - y'\|^2$, where $\|\cdot\|$ is the inner product norm on $\mathcal{Y}$. A consequence of the following result is that, on a task with equivariant structure, the difference in risk between a predictor f and any equivariant predictor f' such that $R[f'] \leq R[\mathcal{Q}f]$ is at least the norm of the $\mathcal{G}$-anti-symmetric component of of f. This shows a barrier to generalisation if f is not equivariant. It also shows that the lower bound on the risk can be (approximately) eliminated by making f (approximately) equivariant. In short: for any non-equivariant predictor, there is an equivariant predictor that performs better. The projection $\mathcal{Q}f$ is the archetypal equivariant predictor to which f should be compared. From Lemma 3.1, $\mathcal{Q}f$ can be thought of the equivariant part of f and in addition $\mathcal{Q}f$ is the closest equivariant predictor to f.

Lemma 3.12. Let $X \sim \mu$ and let Y have finite second moment. Assume either of the following models for Y

1. $Y \overset{\mathrm{d}}{=} f^\star(X) + \xi$ where $f^\star \in L_2(\mathcal{X}, \mathcal{Y}, \mu)$ is equivariant, $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\|\xi\|^2] < \infty$, or
2. $Y \overset{\mathrm{d}}{=} \tilde{f}(X, \eta)$ where $\tilde{f} : \mathcal{X} \times [0,1] \to \mathbb{R}$ is equivariant in its first argument, $\eta \sim \mathrm{Unif}[0,1]$ and $\eta \perp\!\!\!\perp X$.

Let $f \in L_2(\mathcal{X}, \mathcal{Y}, \mu)$ and let $f' \in L_2(\mathcal{X}, \mathcal{Y}, \mu)$ be any predictor such that $R[f'] \leq R[\bar{f}]$ where $\bar{f} = \mathcal{Q}f$. Let $f^\perp = f - \bar{f}$ be the $\mathcal{G}$-anti-symmetric component of f, then

$$R[f] - R[f'] \geq \|f^\perp\|_\mu^2$$

with equality when $R[f'] = R[\bar{f}]$.

Proof. Assume the statement in the case of equality. Consider $f' \in L_2(\mathcal{X}, \mathcal{Y}, \mu)$ such that $R[f'] \leq R[\bar{f}]$, then $R[f] - R[f'] = R[f] - R[\bar{f}] + R[\bar{f}] - R[f'] \geq \|f^\perp\|_\mu^2$. We now address the case of equality for each model. Suppose $Y \overset{\mathrm{d}}{=} f^\star(X) + \xi$. Using Lemma 3.1 we write $f = \bar{f} + f^\perp$ where the terms are orthogonal and the first is equivariant. Then using $\mathbb{E}[\xi]=0$ and the orthogonality

$$\begin{aligned}
R[f] &= \mathbb{E}[\|f(X)-Y\|^2] \\
&= \mathbb{E}[\|\bar{f}(X)-f^\star(X)+f^\perp(X)\|^2]+\mathbb{E}[\|\xi\|^2] \\
&= \|\bar{f}-f^\star+f^\perp\|_\mu^2+\mathbb{E}[\|\xi\|^2] \\
&= \|\bar{f}-f^\star\|_\mu^2+\|f^\perp\|_\mu^2+\mathbb{E}[\|\xi\|^2] \\
&= R[\bar{f}]+\|f^\perp\|_\mu^2.
\end{aligned}$$

Now consider the second model for Y. Let us write $f_t^\star(x)=\tilde{f}(x,t)$. Since Y has finite second moment

$$\infty > \mathbb{E}[\|Y\|^2] = \int_{[0,1]} \|f_t^\star\|_\mu^2 \, \mathrm{d}t$$

by Fubini's theorem so $f_t^\star \in L_2(\mathcal{X},\mathcal{Y},\mu)$ for almost all $t\in[0,1]$. As before, we can use Lemma 3.1 to get the orthogonal decomposition $f=\bar{f}+f^\perp$. Additionally, the equivariance of $f_t^\star$ means $\langle f_t^\star, f^\perp\rangle_\mu=0$. Hence

$$R[f]=\mathbb{E}[\|f(X)-Y\|^2]=\mathbb{E}[\|f-f_\eta^\star\|_\mu^2]=\mathbb{E}[\|\bar{f}-f_\eta^\star\|_\mu^2]+\|f^\perp\|_\mu^2=R[\bar{f}]+\|f^\perp\|_\mu^2.$$

□

The first model for Y in Lemma 3.12 covers the standard regression setup with an equivariant target function. The second model, inspired by [22], gives conditional equivariance in distribution: $Y|gX \overset{\mathrm{d}}{=} gY|X$ for all $g\in\mathcal{G}$ and is aimed at stochastic equivariant functions. The first is a special case of the second when ξ is invariant in distribution. Additionally, Bloem-Reddy and Teh [22] show that conditional invariance in distribution, i.e., $Y|gX \overset{\mathrm{d}}{=} Y|X$ for all $g\in\mathcal{G}$, is equivalent to $Y \overset{\text{a.s.}}{=} \tilde{f}(X,\eta)$ for some $\tilde{f}$ that's invariant in its first argument. It's straightforward to derive a version of Lemma 3.12 for the case that $f^\star$ or $\tilde{f}$ are approximately equivariant.

In later chapters we will use Lemma 3.12 to calculate explicitly the generalisation benefit of invariance/equivariance in random design least squares regression and kernel ridge regression. We will see that $\|f^\perp\|_\mu^2$ displays a natural relationship between the number of training examples and the dimension of the space of $\mathcal{G}$-anti-symmetric predictors A, which is a property of the group action. Intuitively, the algorithm needs enough examples to learn to be orthogonal to A.

Remark 3.13. The same idea used in Lemma 3.12 can be used to give a lower bound on the excess risk in the misspecified case where the predictor f is equivariant but the target $f^\star$ is not. For instance, under the first model in Lemma 3.12

$$R[f] - R[f^\star] = \|f - \mathcal{Q}f^\star\|_\mu^2 + \|(\mathrm{id} - \mathcal{Q})f^\star\|_\mu^2 \geq \|(\mathrm{id} - \mathcal{Q})f^\star\|_\mu^2.$$

3.3.2 Detour: Test-Time Augmentation

Test-time augmentation consists of averaging the output of a learned function f over random transformations of its input and can be used to increase test accuracy [163, 168, 77]. When the transformations belong to a group $\mathcal{G}$ and are sampled from its Haar measure, test-time augmentation can be written as

$$\widehat{\mathcal{O}}_n f(x) = \frac{1}{n}\sum_{i=1}^n f(G_i x)$$

where $G_1, \ldots, G_n \sim \lambda$ are independent and identically distributed. We can view $\widehat{\mathcal{O}}_n f$ as an unbiased Monte-Carlo estimate of $\mathcal{O}f$. For any bounded function f and any x, $\widehat{\mathcal{O}}f(x) \to \mathcal{O}f(x)$ λ-almost-surely by the strong law of large numbers [85, Theorem 4.23]. By considering this limit, Lemma 3.12 hints at an explanation for the generalisation improvement from test-time augmentation.

However, more work is needed to connect Lemma 3.12 to [163, 168, 77]. In particular, Simonyan and Zisserman [163], Szegedy et al. [168], and He, Zhang, Ren, and Sun [77] average softmax outputs rather than predictions and use classification accuracy to calculate the test error.

Chapter 4

The Linear Model

Summary

In this chapter we apply Lemmas 3.1 and 3.12 to linear models in a random design setting. Although invariance is a special case of equivariance, we find it instructive to discuss separately the fully general result of least squares regression with an equivariant target and the special case of an invariant, scalar valued target. These results are given in Theorems 4.1 and 4.7 respectively. The work in this chapter provides the first proofs of a strict generalisation benefit of invariance/equivariance in each case.

Notation

In this chapter we will write the actions ϕ and ψ explicitly. In particular we write

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

and analogously for $\mathcal{O}$.

4.1 Regression with an Invariant Target

Let $\mathcal{X} = \mathbb{R}^d$ with the Euclidean inner product and $\mathcal{Y} = \mathbb{R}$ with multiplication. Consider linear regression with the squared-error loss $\ell(y, y') = (y - y')^2$. Let $\mathcal{G}$ act on $\mathcal{X}$ via an orthogonal representation $\phi : \mathcal{G} \to \mathsf{O}_d$ and let $X \sim \mu$ be such that $\Sigma \coloneqq \mathbb{E}[XX^\top]$ is finite and positive definite.[1] 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 $L_2(\mu)$. Notice that V_{lin} is closed under $\mathcal{O}$: for all $x \in \mathcal{X}$

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

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

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

We denote the induced inner product on $\mathcal{X}$ by $\langle a, b \rangle_\Sigma \coloneqq a^\top \Sigma b$ and the corresponding norm by $\|\cdot\|_\Sigma$. Since V_{lin} is closed under $\mathcal{O}$ we can apply Lemma 3.1 to decompose $V_{\text{lin}} = S \oplus A$ with the orthogonality with respect to $\langle \cdot, \cdot \rangle_\Sigma$. 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 $h_w^\perp = h_{w^\perp}$. There is an isomorphism $\mathcal{X} \to V_{\text{lin}}$ where $w \mapsto h_w$. Using this identification, we abuse notation slightly and write $\mathcal{X} = S \oplus A$ to represent the

[1] If Σ is only positive semi-definite then the developments are similar.

induced structure on $\mathcal{X}$.

Recall the definition of the *risk* of a predictor f

$$R[f] = \mathbb{E}[\|f(X) - Y\|^2].$$

The risk is implicitly conditional on any randomness in f, e.g., occurring from its dependence on training data. We will refer to the difference in risk between two predictors as the *generalisation gap*. So the generalisation gap between f and f' is $R[f]-R[f']$. If this quantity is positive, then we have strictly better test performance from f'. This gives a method of comparing predictors.

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. We calculate the difference in risk between h_w and its invariant version $h_{\bar{w}}$. Lemma 3.12 gives

$$R[h_w] - R[h_{\bar{w}}] = \|h_{w^\perp}\|_\mu^2 = \|w^\perp\|_\Sigma^2 \tag{4.1.1}$$

In Theorem 4.1 we calculate this quantity exactly. 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 the generalisation benefit of invariant models.

Theorem 4.1. 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 where 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. training examples $((X_i, Y_i) : i = 1, \dots, n)$ distributed independently of and identically to (X, Y) and let A be the orthogonal complement of the subspace of $\mathcal{G}$-invariant linear predictors (as in Lemma 3.1).

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

$$\mathbb{E}\left[R[h_w] - R[h_{\bar{w}}]\right] = \sigma_\xi^2 \frac{\dim A}{n-d-1}.$$

- 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

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

The expectations are over the training sample. All of the above hold with $\geq$ when $h_{\bar{w}}$ is replaced by any predictor h' such that $R[h'] \leq R[h_{\bar{w}}]$ (see Lemma 3.12).

Proof. Note that X is $\mathcal{G}$-invariant for all $\mathcal{G}$ 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 3.1 we can write $V_{\text{lin}} = S \oplus A$. Let $\Phi^A = I - \Phi$, which is the orthogonal projection onto the subspace A. By isotropy of X and Eq. (4.1.1) we have

$$R[h_w] - R[h_{\bar{w}}] = \sigma_X^2 \|w^\perp\|_2^2$$

for all $w \in \mathbb{R}^d$, where $w^\perp = \Phi^A(w)$. 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.,

$$w = (\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top \boldsymbol{X}\theta + (\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top \boldsymbol{\xi} \tag{4.1.2}$$

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 $\mathbb{E}[\|w^\perp\|_2^2 \mid \boldsymbol{X}]$ where $w^\perp = \Phi^A(w)$. The contribution from the first term of Eq. (4.1.2) is

$$\|\Phi^A(P_E\theta)\|_2^2,$$

the cross term vanishes using $\xi \perp\!\!\!\perp X$ and $\mathbb{E}[\xi] = 0$, and the contribution from the second term of Eq. (4.1.2) is

$$\mathbb{E}[\|\Phi^A((\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{\xi})\|_2^2 \mid \boldsymbol{X}].$$

Φ^A is an orthogonal projection, so as a matrix is symmetric and idempotent. Hence, briefly writing Φ^A without the parenthesis to emphasise the matrix interpretation,

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

We have obtained

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

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 because $\Phi^A(\theta) = 0$. We treat the $n < d$ case using Einstein notation, in which repeated indices are implicitly summed over. In components, recalling that Φ^A is a matrix,

$$\mathbb{E}[\|\Phi^A(P_E\theta)\|_2^2] = \Phi^A_{fa}\Phi^A_{fc}\,\mathbb{E}[P_E \otimes P_E]_{abce}\theta_b\theta_e$$

and applying Lemma A.3.6 we get

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

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

Second Term By linearity,

$$\mathbb{E}[\mathrm{Tr}(\Phi^A((\boldsymbol{X}^\top\boldsymbol{X})^+))] = \mathrm{Tr}(\Phi^A(\mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{X})^+])).$$

Then Lemmas A.3.2 and A.3.4 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.3.1. Hence

$$\mathbb{E}[\mathrm{Tr}(\Phi^A((\boldsymbol{X}^\top\boldsymbol{X})^+))] = \sigma_X^{-2}r(n,d)\dim A.$$

□

In each case in Theorem 4.1, 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 [75].

Notice the central role of $\dim A$ in Theorem 4.1. This quantity is a property of the group action as it describes the codimension of the space of invariant models. The generalisation gap is then dictated by how significant the symmetry is to the problem. We give two examples representing (non-trivial) extremal cases.

Example 4.2 (Permutations, $\dim A = d-1$)**.** The matrix Φ is invariant under the action of any $g' \in \mathcal{G}$

$$\phi(g')\Phi = \phi(g') \int_{\mathcal{G}} \phi(g) \, \mathrm{d}\lambda(g) = \int_{\mathcal{G}} \phi(g'g) \, \mathrm{d}\lambda(g) = \int_{\mathcal{G}} \phi(g) \, \mathrm{d}\lambda(g) = \Phi.$$

In the case of S_d with its representation as permutations matrices this implies $\Phi_{\mathsf{S}_d} = c\mathbf{1}\mathbf{1}^\top$ for some c that may depend on other quantities. This implies $\dim S = 1$ so $\dim A = d-1$ and the invariance $\Phi_{\mathsf{S}_d}\mathbf{1} = \mathbf{1}$ gives $c = 1/d$.

Example 4.3 (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 all $j = 1, \dots, d$

$$\Phi_{\mathsf{C}_2}(w)_j = \frac{1}{|\mathsf{C}_2|} \sum_{g \in \mathsf{C}_2} (\phi(g)w)_j = 0.$$

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

4.2 Regression with an 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 $\|y - y'\|_2^2$. Let $X \sim \mu$ be such that $\Sigma := \mathbb{E}[XX^\top]$ is finite and positive definite. Let ψ be an orthogonal representation of $\mathcal{G}$ on $\mathcal{Y}$. We define the linear map, which we call the *intertwining average*, $\Psi : \mathbb{R}^{d\times k} \to \mathbb{R}^{d\times k}$ by[2]

$$\Psi(W) = \int_{\mathcal{G}} \phi(g) W \psi(g^{-1}) \, \mathrm{d}\lambda(g). \tag{4.2.1}$$

Similarly, define the *intertwining complement* as $\Psi^A : \mathbb{R}^{d\times k} \to \mathbb{R}^{d\times k}$ by $\Psi^A(W) = W - \Psi(W)$. We establish the following results, which are generalisations of the invariant case. In the proofs we will leverage the expression of Ψ as a 4-tensor with components

$$\Psi_{abce} = \int_{\mathcal{G}} \phi(g)_{ac} \psi(g)_{be} \, \mathrm{d}\lambda(g)$$

where $a, c = 1, \dots d$ and $b, e = 1, \dots, k$. This expression follows from the orthogonality of ψ by

$$\Psi(W)_{ab} = \int_{\mathcal{G}} \phi(g)_{ac} W_{ce} \psi(g^{-1})_{eb} \, \mathrm{d}\lambda(g) = \int_{\mathcal{G}} \phi(g)_{ac} W_{ce} \psi(g)_{be} \, \mathrm{d}\lambda(g) = \Psi_{abce} W_{ce}.$$

Proposition 4.4. $\forall f_W \in W_{\text{lin}}$: $\mathcal{Q} f_W = f_{\Psi(W)}$ so W_{lin} is closed under $\mathcal{Q}$.

Proof. Let $f_W(x) = W^\top x$ with $W \in \mathbb{R}^{d\times k}$. By orthogonality and substituting $g \mapsto g^{-1}$

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

□

Proposition 4.5. For all $f_{W_1}, f_{W_2} \in W_{\text{lin}}$, $\langle f_{W_1}, f_{W_2} \rangle_\mu = \mathrm{Tr}(W_1^\top \Sigma W_2)$.

[2] The reader may have noticed that we define Ψ 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.

Proof.

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

□

Proposition 4.4 allows us to apply Lemma 3.1 to write $W_{\text{lin}} = S \oplus A$, so for all $f_W \in W_{\text{lin}}$ 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 $\overline{W} = \Psi(W)$ and $W^\perp = \Psi^A(W)$ must therefore satisfy $\mathrm{Tr}(\overline{W}^\top \Sigma W^\perp) = 0$. Repeating our abuse of notation, we identify $\mathbb{R}^{d\times k} = S \oplus A$ with $S = \Psi(\mathbb{R}^{d\times k})$ and A its orthogonal complement with respect to the induced inner product.

Proposition 4.6. Let $X \sim \mu$ and let ξ be a random vector in $\mathbb{R}^k$ that is independent of X with $\mathbb{E}[\xi] = 0$ and finite variance. Set $Y = h_\Theta(X) + \xi$ where h_Θ is $\mathcal{G}$-equivariant. For all $f_W \in W_{\text{lin}}$, the generalisation gap satisfies

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

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

Proof. Recall that $W = \overline{W} + W^\perp$ and that these satisfy $\mathrm{Tr}(\overline{W}\Sigma W^\perp) = 0$ from the above. Then, using Lemma 3.12 and Proposition 4.5,

$$R[f_W] - R[f_{\overline{W}}] = \|f_{W^\perp}\|_\mu^2 = \mathrm{Tr}((W^\perp)^\top \Sigma W^\perp) = \|\Sigma^{1/2} W^\perp\|_{\mathrm{F}}^2.$$

□

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 $\overline{W} = \Psi(W)$. The choice of $\overline{W} = \Psi(W)$ as a comparator is natural. Indeed, following directly from Lemma 3.12 it costs us nothing in terms of the strength of the following result.

Theorem 4.7. 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. training examples $((X_i, Y_i) : i = 1, \dots, n)$ distributed independently of and identically to (X, Y) 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

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

- 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

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

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

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

All of the above hold with $\geq$ when $f_{\overline{W}}$ is replaced by any predictor f' such that $R[f'] \leq R[f_{\overline{W}}]$ (see Lemma 3.12).

Proof. We use Einstein notation, in which repeated indices are summed over. Since the representation ϕ is orthogonal, X is $\mathcal{G}$-invariant for all $\mathcal{G}$. We have seen from Proposition 4.6 that

$$\mathbb{E}\left[R[f_W] - R[f_{\overline{W}}]\right] = \sigma_X^2\, \mathbb{E}[\|W^\perp\|_{\mathrm{F}}^2]$$

and we want to calculate 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

$$\begin{aligned}
&\mathbb{E}\left[R[f_W] - R[f_{\overline{W}}]\right] \\
&= \sigma_X^2\, \mathbb{E}[\|\Psi^A(W)\|_{\mathrm{F}}^2] \\
&= \sigma_X^2\, \mathbb{E}[\|\Psi^A((\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top \boldsymbol{X}\Theta + (\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top \boldsymbol{\xi})\|_{\mathrm{F}}^2] \\
&= \sigma_X^2\, \mathbb{E}[\|\Psi^A((\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top \boldsymbol{X}\Theta)\|_{\mathrm{F}}^2] + \sigma_X^2\, \mathbb{E}[\|\Psi^A((\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top \boldsymbol{\xi})\|_{\mathrm{F}}^2]
\end{aligned}$$

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

Second Term Setting $\boldsymbol{Z} = (\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top$ we have

$$\mathbb{E}[\|\Psi^A(\boldsymbol{Z}\boldsymbol{\xi})\|_{\mathrm{F}}^2] = \mathbb{E}[\mathrm{Tr}(\Psi^A(\boldsymbol{Z}\boldsymbol{\xi})^\top \Psi^A(\boldsymbol{Z}\boldsymbol{\xi}))].$$

One gets

$$\begin{aligned}
\mathbb{E}[\mathrm{Tr}(\Psi^A(\boldsymbol{Z\xi})^\top \Psi^A(\boldsymbol{Z\xi}))] &= \mathbb{E}[\Psi^A_{abcj}\boldsymbol{Z}_{ce}\boldsymbol{\xi}_{ej}\Psi^A_{abfg}\boldsymbol{Z}_{fh}\boldsymbol{\xi}_{hg}] \\
&= \sigma_\xi^2\,\mathbb{E}[\Psi^A_{abcj}\boldsymbol{Z}_{ce}\Psi^A_{abfg}\boldsymbol{Z}_{fh}\delta_{eh}\delta_{jg}] \\
&= \sigma_\xi^2\Psi^A_{abcj}\Psi^A_{abfj}\,\mathbb{E}[\boldsymbol{Z}_{ce}\boldsymbol{Z}_{fe}] \\
&= \sigma_\xi^2\Psi^A_{abcj}\Psi^A_{abfj}\,\mathbb{E}[(\boldsymbol{Z}\boldsymbol{Z}^\top)_{cf}] \qquad (4.2.2)
\end{aligned}$$

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

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

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. Now we put things back together. To begin with

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

and putting this into Eq. (4.2.2) with $\boldsymbol{Z}\boldsymbol{Z}^\top = (\boldsymbol{X}^\top \boldsymbol{X})^+$ gives

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

where $c, e = 1, \ldots, d$. Applying Lemmas A.3.2 and A.3.4 gives $\mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^+_{ce}] = \sigma_X^{-2} r(n,d)\delta_{ce}$ where

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

When $n \in [d-1, d+1]$ it is well known that the expectation diverges, see Appendix A.3.1. Using the orthogonality of ϕ we arrive at

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

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.[3] Recall that $\Psi^A(\Theta) = 0$, which in components reads

$$\Psi^A_{abce} \Theta_{ce} = 0 \qquad \forall a, b. \tag{4.2.3}$$

Also in components, we have

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

[3] Regarding Remark 4.8, this is where absolute continuity with respect to the Lebesgue measure is required.

and using Lemma A.3.6 we get

$$\begin{aligned}
&\mathbb{E}[\|\Psi^A((\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{X}\Theta)\|_{\mathrm{F}}^2] \\
&= \frac{n(d-n)}{d(d-1)(d+2)}\Psi^A_{fhai}\Psi^A_{fhaj}\Theta_{bi}\Theta_{bj} \\
&\quad+ \frac{n(d-n)}{d(d-1)(d+2)}\Psi^A_{fhai}\Psi^A_{fhbj}\Theta_{bi}\Theta_{aj} \\
&\quad+ \frac{n(d-n)+n(n-1)(d+2)}{d(d-1)(d+2)}\Psi^A_{fhai}\Psi^A_{fhcj}\Theta_{ai}\Theta_{cj}.
\end{aligned} \tag{4.2.4}$$

The third term vanishes by Eq. (4.2.3). Consider the remaining two terms separately. Start with the first term of Eq. (4.2.4), in which

$$\Psi^A_{fhai}\Psi^A_{fhaj}\Theta_{bi}\Theta_{bj} = (\Theta^\top\Theta)_{ij}\Psi^A_{fhai}\Psi^A_{fhaj}$$

where

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

using the orthogonality of the representations and invariance of the Haar measure (the calculation is the same as for the first term). Therefore

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

Now for the second term of Eq. (4.2.4)

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

Putting these together gives

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

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

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

□

Theorem 4.7 is a direct generalisation of Theorem 4.1. As we remarked in the introduction, $dk-(\chi_\psi|\chi_\phi)$ plays the role of $\dim A$ in Theorem 4.1 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 4.1, there is an additional term (the first) in the overparameterised case $d>n+1$ that represents the estimation error in the 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 4.7 reduces exactly to Theorem 4.1.

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

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

where

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

Remark 4.8. Versions of Theorem 4.7 are possible for other probability distributions on X. For instance, a similar result would hold for any isotropic distribution that is absolutely continuous with respect to the Lebesgue measure and has finite variance. The isotropy implies the existence of a scalar r (which depends on n and the distribution of X) such that $\mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{X})^+] = rI_d$ where $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ are the row-stacked training inputs as defined in the proof.

Chapter 5

Kernel Methods

Summary

In this chapter we apply and extend the results of Chapter 3 to reproducing kernel Hilbert spaces, focussing on invariance rather than equivariance. We go a step further than the previous chapter on linear models, applying Lemma 3.1 to establish Theorem 5.2, which gives a strict generalisation benefit for invariance in kernel ridge regression when the symmetry is also present in the target. As with Chapter 4, we study the random design setting. This result is specialised to the linear kernel and the orthogonal group, showing a fundamental connection to Theorem 4.1. The other main developments of this chapter are more abstract. Theorem 5.2 gives a lower bound for the generalisation of kernel ridge regression with an invariant target. By studying the result, we uncover a condition on the kernel under which this lower bound vanishes in the $n \to \infty$ limit. Assuming that this condition holds, we derive an independent result in Theorem 5.14 that provides a decomposition of a reproducing kernel Hilbert space into orthogonal $\mathcal{G}$-symmetric and $\mathcal{G}$-anti-symmetric parts, analogous to Lemma 3.1.

5.1 Background, Assumptions and Preliminaries

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive definite* if, for all n and all distinct $x_1, \ldots, x_n \in \mathcal{X}$ the matrix with components $K_{ij} = k(x_i, x_j)$ is positive semi-definite. For $\boldsymbol{X} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ we write $f(\boldsymbol{X}) \in \mathbb{R}^n$ as the vector with elements $f(\boldsymbol{X})_i = f(x_i)$. For any Borel measure ν we write $\operatorname{supp} \nu$ for the support of ν, which is the smallest closed subset of $\mathcal{X}$ whose complement has measure 0. We will say that a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is invariant if $k(x, gx') = k(x, x')$ for all $g \in \mathcal{G}$ and all $x, x' \in \mathcal{X}$. For any measurable $j : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ we define

$$\|j\|^2_{L_2(\mu \otimes \mu)} = \int_{\mathcal{X}} j(x, y)^2 \, \mathrm{d}\mu(x) \, \mathrm{d}\mu(y).$$

5.1.1 The Basics of Reproducing Kernel Hilbert Spaces

A *Hilbert space* is an inner product space that is complete with respect to the norm topology induced by the inner product. We will only consider spaces over $\mathbb{R}$. A *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$ is a Hilbert space of real functions $f : \mathcal{X} \to \mathbb{R}$ on which the evaluation functional $\delta_x : \mathcal{H} \to \mathbb{R}$ with $\delta_x[f] = f(x)$ is continuous for all $x \in \mathcal{X}$ or, equivalently, is a bounded operator. The Riesz representation theorem tells us that there is a unique function $k_x \in \mathcal{H}$ such that $\delta_x[f] = \langle f, k_x \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is the inner product on $\mathcal{H}$. We will refer to the function k_x as the *representer (of evaluation at x)*. We identify the function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with $k(x, y) = \langle k_x, k_y \rangle_{\mathcal{H}}$ as the *(reproducing) kernel* of $\mathcal{H}$. Using the inner product representation, one can see that k is positive definite and symmetric. Conversely, the Moore-Aronszajn theorem states that for any positive definite and symmetric function k, there is a unique RKHS with reproducing kernel k [12]. In addition, any Hilbert space of functions admitting a reproducing kernel is an RKHS. Finally, another characterisation of $\mathcal{H}$ is as the completion (with respect to the norm topology) of the space of linear combinations $f_c(x) = \sum_{i=1}^n c_i k(x, x_i)$ for $c_1, \ldots, c_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in \mathcal{X}$ with the inner product between f_c and $f_{\tilde{c}}$ being $\sum_{i,j=1}^n c_i \tilde{c}_j k(x_i, x_j)$. For (many) more details, see [166, Chapter 4].

5.1.2 Assumptions

In this chapter we make the additional assumption that $\operatorname{supp}\mu = \mathcal{X}$. The output space will be $\mathcal{Y} = \mathbb{R}$. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a measurable kernel with RKHS $\mathcal{H}$ such that $k(\cdot, x) : \mathcal{X} \to \mathbb{R}$ is continuous for all $x \in \mathcal{X}$. Assume that $\sup_{x\in\mathcal{X}} k(x,x) = M_k < \infty$ and note that this implies that k is bounded since

$$k(x,x') = \langle k_x, k_{x'}\rangle_{\mathcal{H}} \leq \|k_x\|_{\mathcal{H}}\|k_{x'}\|_{\mathcal{H}} = \sqrt{k(x,x)}\sqrt{k(x',x')} \leq M_k.$$

Every $f \in \mathcal{H}$ is μ-measurable, bounded and continuous by Lemmas A.4.1 and A.4.2. Recall that $\mathcal{X}$ is assumed to be a Polish space, so it is separable and $\mathcal{H}$ is separable using Lemma A.4.3.

5.1.3 Relating $\mathcal{H}$ to $L_2(\mu)$

Define the integral operator $S_k : L_2(\mu) \to \mathcal{H}$

$$S_k f(x) = \int_{\mathcal{X}} k(x,x') f(x') \, \mathrm{d}\mu(x').$$

S_k assigns a function in $\mathcal{H}$ to every element of $L_2(\mu)$. On the other hand, every $f \in \mathcal{H}$ is measurable and bounded so has $\|f\|_\mu < \infty$ and belongs to some element of $L_2(\mu)$. We write $\iota : \mathcal{H} \to L_2(\mu)$ for the inclusion map that sends f to the element of $L_2(\mu)$ that contains f. By Theorem A.4.4, S_k is well-defined and ι is its adjoint. Both ι and S_k are bounded with operator norms bounded by M_k. This is immediate from Theorem A.4.4 or follows directly from

$$\|\iota f\|_\mu^2 = \int_{\mathcal{X}} f(x)^2 \, \mathrm{d}\mu(x) = \int_{\mathcal{X}} \langle k_x, f\rangle_{\mathcal{H}}^2 \, \mathrm{d}\mu(x) \leq \int_{\mathcal{X}} \|k_x\|_{\mathcal{H}}^2 \|f\|_{\mathcal{H}}^2 \, \mathrm{d}\mu(x) \leq M_k \|f\|_{\mathcal{H}}^2$$

for all $f \in \mathcal{H}$ and $\|S_k\|_{\mathrm{op}} = \|\iota\|_{\mathrm{op}}$ by Theorem A.2.4. Define $T_k : L_2(\mu) \to L_2(\mu)$ by $T_k = \iota \circ S_k$ and it is easy to check that T_k is self-adjoint, $\|T_k\|_{\mathrm{op}} \leq M_k$ and that $\langle T_k f, f\rangle \geq 0$ for all $f \in L_2(\mu)$.

We will make use of the following fact, which is both standard and easy to see.

Lemma 5.1. The image of $L_2(\mu)$ under S_k is dense in $\mathcal{H}$ and ι is injective, so any element of $L_2(\mu)$ contains at most one $f \in \mathcal{H}$.

Proof. Theorem A.4.4 says that $S_k(L_2(\mu))$ is dense in $\mathcal{H}$ if and only if ι is injective. Injectivity of ι is equivalent to the statement that for all $f, f' \in \mathcal{H}$ the set $A(f, f') = \{x \in \mathcal{X} : f(x) \neq f'(x)\}$ has $A \neq \varnothing \implies \mu(A) > 0$. Continuity implies that for all $f, f' \in \mathcal{H}$, either $f = f'$ pointwise or $A(f, f')$ contains an open set. By the support of μ this implies $\mu(A) > 0$: if there was a non-empty open set B with $\mu(B) = 0$ then $\mathcal{X} \setminus B$ is a closed proper subset of $\mathcal{X}$, contradicting $\operatorname{supp} \mu = \mathcal{X}$. Thus, ι is injective. □

Note that we do not assume that $\mathcal{X}$ is compact. This allows for application to common settings such as $\mathcal{X} = \mathbb{R}^n$ but prevents the application of Mercer's theorem. We work around this, but for generalisations of Mercer's theorem see [167] and references therein.

5.1.4 Orbit Averaging

Recall the definition of the averaging operator for invariance $\mathcal{O} : L_2(\mu) \to L_2(\mu)$ as

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

Note that we have not yet defined $\mathcal{O}$ as operator on $\mathcal{H}$. In Section 5.3 we give an additional condition under which $\mathcal{O} : \mathcal{H} \to \mathcal{H}$ is well defined and, as a consequence, an analogy of Lemma 3.1 for RKHSs. For all $x \in \mathcal{X}$ we define $\overline{k_x} = \mathcal{O}\iota k_x$ and $k_x^\perp = \iota k_x - \overline{k_x}$. We also define $\overline{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k^\perp : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$\overline{k}(x, y) = \int_{\mathcal{G}} k(x, gy) \, \mathrm{d}\lambda(g)$$

and $k^\perp(x, y) \coloneqq k(x, y) - \overline{k}(x, y)$. Note that the averaging is only over one argument. The function $k_x[y] : \mathcal{G} \to \mathbb{R}$ with $k_x[y](g) = k(x, gy)$ is λ-measurable by Lemma A.1.2 and a simple composition argument as in Proposition 3.4; it is also bounded so $\overline{k}$

exists, is finite and itself is μ-measurable by Lemma A.1.2. For all $x \in \mathcal{X}$ we have $\overline{k}(x, y) = \overline{k_x}(y)$ μ-almost-surely in y. On the other hand, the functions $\overline{k}(x, \cdot)$ are not necessarily in $\mathcal{H}$.

5.2 Generalisation

In this section we apply the theory developed in Chapter 3 to study the impact of invariance on kernel ridge regression with an invariant target. We analyse the generalisation benefit of invariance, finding a strict benefit when the symmetry is present in the target.

5.2.1 Kernel Ridge Regression

Given observations $((x_i, y_i) : i = 1, \dots, n)$ where $x_i \in \mathcal{X}$ and $y_i \in \mathbb{R}$, *kernel ridge regression* (KRR) returns a predictor that solves the optimisation problem

$$\operatorname*{argmin}_{f \in \mathcal{H}} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \rho \|f\|_{\mathcal{H}}^2 \tag{5.2.1}$$

and $\rho > 0$ is the regularisation parameter. In general ρ is a function of n, but we leave this dependence implicit to save on notation. KRR can be thought of as performing ridge regression in a possibly infinite dimensional feature space $\mathcal{H}$. The solution to this problem is of the form $f(x) = \sum_{i=1}^n \alpha_i k_{x_i}(x)$ where $\alpha \in \mathbb{R}^n$ solves

$$\operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \|\boldsymbol{Y} - K\alpha\|_2^2 + \rho \alpha^\top K \alpha, \tag{5.2.2}$$

$\boldsymbol{Y} \in \mathbb{R}^n$ is the standard row-stacking of the training outputs with $\boldsymbol{Y}_i = y_i$ and K is the kernel Gram matrix with $K_{ij} = k(x_i, x_j)$. The form of the solution is an immediate consequence of the representer theorem [90, 150], or can be seen directly by expanding f in terms of its components in and orthogonal to the span of $\{k_{x_1}, \dots, k_{x_n}\}$ and substituting into Eq. (5.2.1). We consider solutions of the form

$\alpha = (K + \rho I)^{-1}\boldsymbol{Y}$ which results in the predictor

$$f(x) = k_{\boldsymbol{X}}(x)^\top (K + \rho I)^{-1}\boldsymbol{Y} \tag{5.2.3}$$

where $k_{\boldsymbol{X}} : \mathcal{X} \to \mathbb{R}^n$ has values $k_{\boldsymbol{X}}(x)_i = k_{x_i}(x) = k(x_i, x)$. It's easy to see that this solution is unique. If K is a positive definite matrix then of course α is the unique solution to Eq. (5.2.2). On the other hand suppose K is degenerate and let β be the component of α in the null space of K, then $0 = \beta^\top K \beta = \|\sum_{i=1}^n \beta_i k_{x_i}\|_{\mathcal{H}}^2$ so f is independent of β.

In calculating the generalisation benefit of invariance, we will use as a comparator its averaged version

$$\bar{f}(x) = \mathcal{O}\iota f(x) = \overline{k_{\boldsymbol{X}}}(x)^\top (K + \rho I)^{-1}\boldsymbol{Y}$$

which is $\mathcal{G}$-invariant. Once again, in the proof we will compare the risks of f and $\bar{f}$, recalling the definition of the risk of f, for any random variables (X, Y), as

$$R[f] = \mathbb{E}[\|f(X) - Y\|_2^2]$$

with the expectation conditional on f if it is random. Just as with Lemma 3.12, the choice of $\bar{f}$ as a comparator costs us nothing and this choice is also natural: $\bar{f}$ is the closest invariant predictor to f in $L_2(\mu)$ (see Proposition 3.9).

5.2.2 Generalisation Benefit of Invariance

In this section we give a characterisation of the generalisation benefit of invariance in kernel ridge regression. Theorem 5.2 shows that invariance improves generalisation in kernel ridge regression when the conditional mean of the target distribution is invariant. In a sense, Theorem 5.2 is a generalisation of Theorem 4.1 and we will return to this comparison later.

Theorem 5.2. Let the training sample be $((X_i, Y_i) : i = 1, \dots, n)$ i.i.d. with $X_i \sim \mu$

and $Y_i = f^\star(X_i) + \xi_i$ where $f^\star \in L_2(\mu)$ is $\mathcal{G}$-invariant with $\mathbb{E}[\xi_i \mid X_1, \dots, X_n] = 0$ and $\mathbb{E}[\xi_i \xi_j \mid X_1, \dots, X_n] = \sigma^2 \delta_{ij} < \infty$ for $i, j = 1, \dots, n$. Similarly, let $X \sim \mu$ and $Y = f^\star(X) + \xi$ where $\mathbb{E}[\xi \mid X] = 0$ and $\mathbb{E}[\xi^2] = \sigma^2 < \infty$. Let f be the solution to the KRR problem Eq. (5.2.1) given in Eq. (5.2.3) and let $\bar{f} = \mathcal{O}\iota f$ be its averaged version. Let $f' \in L_2(\mu)$ be any predictor with risk not larger than $\bar{f}$, i.e., $R[f'] \leq R[\bar{f}]$. Then

$$\mathbb{E}\left[R[f] - R[f']\right] \geq \mathbb{E}[\|(\mathrm{id} - \mathcal{O})\iota\Lambda_{n,\rho} f^\star\|_\mu^2] + \sigma^2 \frac{\|k^\perp\|^2_{L_2(\mu\otimes\mu)}}{(\sqrt{n}M_k + \rho/\sqrt{n})^2}$$

where $\Lambda_{n,\rho} f^\star$ solves the corresponding noiseless problem

$$\underset{f\in\mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^n (f(X_i) - f^\star(X_i))^2 + \rho\|f\|^2_{\mathcal{H}}. \tag{5.2.4}$$

Proof. It is sufficient to set $f' = \bar{f}$, since if $R[f'] \leq R[\bar{f}]$ we have

$$R[f] - R[f'] = R[f] - R[\bar{f}] + R[\bar{f}] - R[f'] \geq R[f] - R[\bar{f}].$$

Let $J^\perp$ be the Gram matrix with components $J^\perp_{ij} = \langle k^\perp_{X_i}, k^\perp_{X_j}\rangle_\mu$ and let $u \in \mathbb{R}^n$ have components $u_i = f^\star(X_i)$. We can use Lemma 3.12 to get

$$R[f] - R[\bar{f}] = \mathbb{E}[(k^\perp_{\boldsymbol{X}}(X)^\top (K + \rho I)^{-1}\boldsymbol{Y})^2 \mid \boldsymbol{X}, \boldsymbol{Y}]$$

where $X \sim \mu$ and $k^\perp_{\boldsymbol{X}} : \mathcal{X} \to \mathbb{R}^n$ with $k^\perp_{\boldsymbol{X}}(x)_i = k^\perp_{X_i}(x)$. Let $\boldsymbol{\xi} \in \mathbb{R}^n$ have components $\boldsymbol{\xi}_i = \xi_i$ then one finds

$$\begin{aligned}
&\mathbb{E}[R[f] - R[\bar{f}] \mid \boldsymbol{X}] \\
&= \mathbb{E}[(k^\perp_{\boldsymbol{X}}(X)^\top (K + \rho I)^{-1} u)^2 \mid \boldsymbol{X}] + \mathbb{E}[(k^\perp_{\boldsymbol{X}}(X)^\top (K + \rho I)^{-1}\boldsymbol{\xi})^2 \mid \boldsymbol{X}] \\
&= \mathbb{E}[(k^\perp_{\boldsymbol{X}}(X)^\top (K + \rho I)^{-1} u)^2 \mid \boldsymbol{X}] + \sigma^2 \operatorname{Tr}\left(J^\perp (K + \rho I)^{-2}\right)
\end{aligned} \tag{5.2.5}$$

which follows from the assumptions on $\xi_1, \dots, \xi_n$ (by looking at components or via the trace trick). Applying the representer theorem, we find that for all $x \in \mathcal{X}$

$$\Lambda_{n,\rho} f^\star(x) = k_{\boldsymbol{X}}(x)^\top (K + \rho I)^{-1} f^\star(\boldsymbol{X})$$

which allows us to recognise the first term in Eq. (5.2.5) as $\mathbb{E}[\|(\mathrm{id} - \mathcal{O})\iota\Lambda_{n,\rho} f^\star\|_\mu^2]$. We now analyse the second term in Eq. (5.2.5). Note that $J^\perp$ is positive semi-definite because it is a Gram matrix, so we can apply Lemmas A.2.1 and A.2.2 with the bound on the kernel to get

$$\operatorname{Tr}\left(J^\perp (K + \rho I)^{-2}\right) \geq \gamma_{\min}\left((K + \rho I)^{-2}\right) \operatorname{Tr}(J^\perp) \geq \frac{\operatorname{Tr}(J^\perp)}{(M_k n + \rho)^2}.$$

Taking expectations and using the fact that $X_1, \ldots, X_n$ are i.i.d. gives

$$\mathbb{E}\left[\operatorname{Tr}\left(J^\perp (K + \rho I)^{-2}\right)\right] \geq \frac{\mathbb{E}[J^\perp_{11}]}{(M_k \sqrt{n} + \rho/\sqrt{n})^2}.$$

The following completes the proof

$$\mathbb{E}[J^\perp_{11}] = \mathbb{E}[\|k^\perp_X\|_\mu^2] = \int_{\mathcal{X}} k^\perp(x,y)^2 \,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) = \|k^\perp\|^2_{L_2(\mu\otimes\mu)}.$$

□

Remark 5.3. The orthogonal projection on to the span of representers $\{k_{X_1}, \ldots, k_{X_n}\}$ can be obtained by defining $\Pi_n : \mathcal{H} \to \mathcal{H}$ by $\Pi_n f = \lim_{\rho\to 0} \Lambda_{n,\rho} f$ for all $f \in \mathcal{H}$. In particular $\Pi_n f(x) = k_{\boldsymbol{X}}(x)^\top K^+ f(\boldsymbol{X})$ where K^+ is the Moore-Penrose pseudo-inverse of K, so $\Pi_n^2 f(x) = k_{\boldsymbol{X}}(x)^\top K^+ K K^+ f(\boldsymbol{X}) = \Pi_n f(x)$ and $\langle \Pi_n f, h\rangle_{\mathcal{H}} = \sum_{i,j=1}^n h(X_i) K^+_{ij} f(X_j) = \langle f, \Pi_n h\rangle_{\mathcal{H}}$.

Corollary 5.4. Assume the setting and notation of Theorem 5.2 and its proof. If $\gamma_{\min}(J^\perp) \geq c$ almost surely then Theorem 5.2 reduces to

$$\mathbb{E}\left[R[f] - R[f']\right] \geq \frac{\|f^\star\|_\mu^2 c + \sigma^2 \|k^\perp\|^2_{L_2(\mu\otimes\mu)}}{(\sqrt{n} M_k + \rho/\sqrt{n})^2}.$$

Proof. The bias term in Eq. (5.2.5) can be written as

$$\begin{aligned}\mathbb{E}[(k^{\perp}_{\boldsymbol{X}}(X)^{\top}(K+\rho I)^{-1}u)^2 \mid \boldsymbol{X}] &= u^{\top}(K+\rho I)^{-1}J^{\perp}(K+\rho I)^{-1}u \\ &\geq \|u\|_2^2\gamma_{\min}\left((K+\rho I)^{-1}J^{\perp}(K+\rho I)^{-1}\right) \\ &\geq \frac{c\|u\|_2^2}{(nM_k+\rho)^2}.\end{aligned}$$

Using $\mathbb{E}[\|u\|_2^2] = n\|f^{\star}\|_{\mu}^2$ gives the statement. □

Remark 5.5. One can obtain the upper bound

$$\mathbb{E}\left[R[f]-R[\bar{f}]\right] \leq \mathbb{E}[\|(\mathrm{id}-\mathcal{O})\iota\Lambda_{n,\rho}f^{\star}\|_{\mu}^2] + \sigma^2\rho^{-2}\|k^{\perp}\|_{L_2(\mu\otimes\mu)}^2$$

by applying the upper bound in Lemma A.2.1 to Eq. (5.2.5).

5.2.3 Discussion of Theorem 5.2

The first term in Theorem 5.2 corresponds to an approximation error or bias and the second is the variance term. We examine these terms in the following sections. At a high level, the first term represents the propensity of the learning algorithm to produce a predictor with a large $\mathcal{G}$-anti-symmetric component and can be interpreted as the capacity of the algorithm to represent invariant functions in $L_2(\mu)$. The numerator in the second term can be viewed as the complexity of the $\mathcal{G}$-anti-symmetric projection of $\mathcal{H}$, just as in Theorems 4.1 and 4.7.

We note that as a byproduct of Theorem 5.2 we get a lower bound for KRR with random design. The variance term matches the $O(1/n)$ dependence on the number of examples in the upper bound from Mourtada and Rosasco [119, Section 3].

5.2.3.1 Bias Term

The quantity $\|(\mathrm{id} - \mathcal{O})\iota\Lambda_{n,\rho}f^\star\|_\mu^2$ measures the extent to which $\Lambda_{n,\rho}f^\star$ is $\mathcal{G}$-invariant and vanishes if and only if it is. By invariance $f^\star = \mathcal{O}f^\star$ so we can write

$$(\mathrm{id} - \mathcal{O})\iota\Lambda_{n,\rho}f^\star = [\mathcal{O}, \iota\Lambda_{n,\rho}]f^\star$$

where $[A, B] = AB - BA$ denotes the commutator of operators A and B. The bias term $\mathbb{E}[\|[\mathcal{O}, \iota\Lambda_{n,\rho}]f^\star\|_\mu^2]$ measures the extent to which, on average, solving Eq. (5.2.4) maps invariant functions in $L_2(\mu)$ to invariant functions in $\mathcal{H}$. Note that $\Lambda_{n,n\rho}f^\star$ solves

$$\operatorname*{argmin}_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - f^\star(X_i))^2 + \rho\|f\|_{\mathcal{H}}^2.$$

We see in Theorem 5.8 that, under some mild additional conditions on $\mathcal{X}$, $\mathcal{Y}$ and ρ, $\iota\Lambda_{n,n\rho}f^\star \to \iota\Lambda_\rho f^\star$ in probability as $n \to \infty$ where for $\alpha > 0$ $\Lambda_\alpha f^\star$ solves

$$\operatorname*{argmin}_{f\in\mathcal{H}} \|\iota f - f^\star\|_\mu^2 + \alpha\|f\|_{\mathcal{H}}^2. \tag{5.2.6}$$

In this limit the bias term in Theorem 5.2 becomes

$$\mathbb{E}[\|(\mathrm{id} - \mathcal{O})\iota\Lambda_{n,\rho}f^\star\|_\mu^2] \to \|[\mathcal{O}, \iota\Lambda_\rho]f^\star\|_\mu^2$$

as $n \to \infty$. Hence, under the conditions of Theorem 5.8, bias term vanishes in general as $n \to \infty$ if and only if

$$\lim_{n\to\infty}[\mathcal{O}, \iota\Lambda_\rho]f^\star = 0$$

for all invariant $f^\star \in L_2(\mu)$, where we have used the continuity of $\mathcal{O}$ from Proposition 3.4 (recall that ρ can depend on n). We interpret this as a fundamental requirement for learning invariant functions with kernel ridge regression: the risk is strictly positive in the limit $n \to \infty$ if it is not satisfied.

The above motivates us to consider the operator $[\mathcal{O}, \iota\Lambda_\rho]$ for arbitrary $\rho > 0$. In

Lemma 5.6 and Theorem 5.7, we translate this into a condition on the kernel, finding that $[\mathcal{O}, \iota\Lambda_\rho] = 0$ is equivalent to the kernel satisfying

$$\int_{\mathcal{G}} k(gx, t) \, \mathrm{d}\lambda(g) = \int_{\mathcal{G}} k(x, g^{-1}t) \, \mathrm{d}\lambda(g)$$

for all $x, t \in \mathcal{X}$. In respect of the above asymptotic analysis, this is a sufficient condition on k such that KRR can approximate invariant targets to arbitrary accuracy.[1] Note that the above does not imply that the kernel is invariant. We will return to this condition in Section 5.3 where we see that it implies a version of Lemma 3.1 for $\mathcal{H}$. We also outline some examples of kernels that satisfy this condition.

Before Theorem 5.7 we need the following result, from which the takeaway is that $[\mathcal{O}, \iota\Lambda_\rho] = 0$ if and only if $[\mathcal{O}, T_k] = 0$.

Lemma 5.6. Let $\rho > 0$ and $f \in L_2(\mu)$, then

$$[\mathcal{O}, \iota\Lambda_\rho]f = 0 \quad \text{if and only if} \quad [\mathcal{O}, T_k](T_k + \rho \, \mathrm{id})^{-1} f = 0$$

which immediately implies that following are equivalent

(a) $\exists \rho > 0 \; [\mathcal{O}, \iota\Lambda_\rho] = 0$

(b) $\forall \rho > 0 \; [\mathcal{O}, \iota\Lambda_\rho] = 0$

(c) $[\mathcal{O}, T_k] = 0$.

Proof. By Lemma A.2.7

$$\iota\Lambda_\rho = (T_k + \rho \, \mathrm{id})^{-1} T_k.$$

Also note that for linear operators A and B with B invertible,

$$B[A, B^{-1}]B = B(AB^{-1} - B^{-1}A)B = BA - AB = -[A, B].$$

[1] What's missing in saying that it's also a necessary condition is that the asymptotic argument concludes that $[\mathcal{O}, \iota\Lambda_\rho]$ should restrict to 0 on invariant functions, while the condition on the kernel is equivalent to the commutator vanishing altogether.

Then

$$\begin{aligned}
\mathcal{O}\iota\Lambda_\rho &= \mathcal{O}(T_k + \rho\,\mathrm{id})^{-1}T_k \\
&= [\mathcal{O}, (T_k + \rho\,\mathrm{id})^{-1}]T_k + (T_k + \rho\,\mathrm{id})^{-1}\mathcal{O}T_k \\
&= -(T_k + \rho\,\mathrm{id})^{-1}[\mathcal{O}, T_k](T_k + \rho\,\mathrm{id})^{-1}T_k + (T_k + \rho\,\mathrm{id})^{-1}\mathcal{O}T_k \\
&= -(T_k + \rho\,\mathrm{id})^{-1}[\mathcal{O}, T_k]\iota\Lambda_\rho + (T_k + \rho\,\mathrm{id})^{-1}\mathcal{O}T_k \\
&= -(T_k + \rho\,\mathrm{id})^{-1}[\mathcal{O}, T_k]\iota\Lambda_\rho + (T_k + \rho\,\mathrm{id})^{-1}[\mathcal{O}, T_k] + \iota\Lambda_\rho\mathcal{O}
\end{aligned}$$

so for any particular $\rho > 0$ and any $f \in L_2(\mu)$, $[\mathcal{O}, \iota\Lambda_\rho]f = 0$ is equivalent to

$$[\mathcal{O}, T_k]\iota\Lambda_\rho f = [\mathcal{O}, T_k]f.$$

Without loss of generality we can write $f = (T_k + \rho\,\mathrm{id})h$ for some $h \in L_2(\mu)$, then the above is equivalent to

$$[\mathcal{O}, T_k](T_k + \rho\,\mathrm{id})^{-1}T_k(T_k + \rho\,\mathrm{id})h = [\mathcal{O}, T_k](T_k + \rho\,\mathrm{id})h$$

which in turn is equivalent to

$$\rho[\mathcal{O}, T_k]h = 0.$$

The proof is complete. □

Theorem 5.7. The following are equivalent

(a) $[\mathcal{O}, T_k] = 0$

(b) k satisfies

$$\int_{\mathcal{G}} k(gx, t)\,\mathrm{d}\lambda(g) = \int_{\mathcal{G}} k(x, g^{-1}t)\,\mathrm{d}\lambda(g)$$

for all $x, t \in \mathcal{X}$

(c) $\overline{k}$ is a kernel function (it is symmetric and positive definite).

Proof. The condition $[\mathcal{O}, T_k] = 0$ means $\mathcal{O}T_kf = T_k\mathcal{O}f$ for all $f \in L_2(\mu)$, which is captured by the following

$$\begin{aligned}\int_{\mathcal{G}}\int_{\mathcal{X}} k(gx,t)f(t)\,\mathrm{d}\mu(t)\,\mathrm{d}\lambda(g) &\overset{\text{a.s.}}{=} \int_{\mathcal{X}} k(x,t)\int_{\mathcal{G}} f(gt)\,\mathrm{d}\lambda(g)\,\mathrm{d}\mu(t)\\ &= \int_{\mathcal{G}}\int_{\mathcal{X}} k(x,t)f(gt)\,\mathrm{d}\mu(t)\,\mathrm{d}\lambda(g)\\ &= \int_{\mathcal{G}}\int_{\mathcal{X}} k(x,g^{-1}t)f(t)\,\mathrm{d}\mu(t)\,\mathrm{d}\lambda(g)\end{aligned}$$

where the first line is the condition for the commutators to vanish, the second line uses Fubini's theorem and the third line uses the invariance of μ. So, applying Fubini's theorem again, $[\mathcal{O}, T_k] = 0$ is equivalent to

$$\int_{\mathcal{X}}\left(\int_{\mathcal{G}} k(gx,t)\,\mathrm{d}\lambda(g) - \int_{\mathcal{G}} k(x,g^{-1}t)\,\mathrm{d}\lambda(g)\right) f(t)\,\mathrm{d}\mu(t) \overset{\text{a.s.}}{=} 0$$

for all $f \in L_2(\mu)$. In turn this is equivalent to

$$\int_{\mathcal{G}} k(gx,t)\,\mathrm{d}\lambda(g) = \int_{\mathcal{G}} k(x,g^{-1}t)\,\mathrm{d}\lambda(g)$$

$(\mu \otimes \mu)$-almost-everywhere and indeed pointwise by the injectivity of ι. Moreover, one can check that $\overline{k}$ is a kernel function if and only if the above display holds, it is positive definite since for all $n \in \mathbb{N}$, $a \in \mathbb{R}^n$ and $x_1, \dots, x_n \in \mathcal{X}$

$$\sum_{i,j=1}^{n} a_i a_j \overline{k}(x_i, x_j) = \int_{\mathcal{G}} \sum_{i,j=1}^{n} a_i a_j k(x_i, gx_j)\,\mathrm{d}\lambda(g) \geq 0$$

and symmetry is equivalent to the starred equality holding for all $x, x' \in \mathcal{X}$

$$\overline{k}(x,x') = \int_{\mathcal{G}} k(x,gx')\,\mathrm{d}\lambda(g) \overset{\star}{=} \int_{\mathcal{G}} k(gx,x')\,\mathrm{d}\lambda(g) = \int_{\mathcal{G}} k(x',gx)\,\mathrm{d}\lambda(g) = \overline{k}(x',x).$$

□

The following result leans very heavily on [178].

Theorem 5.8. Let $\mathcal{X} \subset \mathbb{R}^d$ be compact and $\mathcal{Y} \subset [-B, B]$. Let $f^\star \in L_2(\mu)$. Let

$\delta \in (0,1)$, then with probability at least $1-\delta$

$$\|\iota\Lambda_{n,n\rho} f^\star - \iota\Lambda_\rho f^\star\|_\mu \leq \frac{BM_k^2}{8\rho} q\left(\frac{8}{n}\log(4/\delta)\right) + \frac{BM_k}{4\sqrt{\rho}} q\left(\frac{8}{n}\log(4/\delta)\right).$$

where $q(t) = \frac{1}{2}(t + \sqrt{t^2+4t})$ and $q(t) = \sqrt{t} + o(t)$ as $t \to 0$. In particular if $\rho = \omega(n^{-l})$ for $l < 1/2$ then

$$\iota\Lambda_{n,n\rho} f^\star \to \iota\Lambda_\rho f^\star$$

in probability as $n \to \infty$.

Proof. The proof is pieced together from [178]. Endow $\mathbb{R}^n$ with the inner product $\langle u, v\rangle_n = \frac{1}{n} u^\top v$ and write the induced norm as $\|\cdot\|_n$. Define the data dependent evaluation operator $E_{\boldsymbol{X}} : \mathcal{H} \to \mathbb{R}^n$ by $(E_{\boldsymbol{X}} f) = f(\boldsymbol{X})$. Its adjoint is $E_{\boldsymbol{X}}^* : \mathbb{R}^n \to \mathcal{H}$ with values $E_{\boldsymbol{X}}^* v(x) = \langle k_{\boldsymbol{X}}(x), v\rangle_n$. The noiseless KRR problem in Theorem 5.2 solved by $\Lambda_{n,n\rho} f^\star$ can then be written

$$\operatorname*{argmin}_{f\in\mathcal{H}} \|E_{\boldsymbol{X}} f - u\|_n^2 + \rho\|f\|_{\mathcal{H}}^2$$

where $u = f^\star(\boldsymbol{X})$ and then Lemma A.2.7 shows that the unique solution is

$$\hat{f} = (E_{\boldsymbol{X}}^* E_{\boldsymbol{X}} + \rho\,\mathrm{id})^{-1} E_{\boldsymbol{X}}^* u$$

so $\hat{f} = \Lambda_{n,\rho} f^\star$. Analogously, the unique solution to Eq. (5.2.6) is

$$f = (S_k \iota + \rho\,\mathrm{id})^{-1} S_k f^\star$$

and $f = \Lambda_\rho f^\star$. Then, following [178, Eq. 18],

$$\begin{aligned}
\hat{f} - f &= (E_{\boldsymbol{X}}^* E_{\boldsymbol{X}} + \rho \,\mathrm{id})^{-1} E_{\boldsymbol{X}}^* u - (S_k \iota + \rho \,\mathrm{id})^{-1} S_k f^\star \\
&= \left[(E_{\boldsymbol{X}}^* E_{\boldsymbol{X}} + \rho \,\mathrm{id})^{-1} - (S_k \iota + \rho \,\mathrm{id})^{-1}\right] E_{\boldsymbol{X}}^* u + (S_k \iota + \rho \,\mathrm{id})^{-1} (E_{\boldsymbol{X}}^* u - S_k f^\star) \\
&= (S_k \iota + \rho \,\mathrm{id})^{-1} (S_k \iota - E_{\boldsymbol{X}}^* E_{\boldsymbol{X}})(E_{\boldsymbol{X}}^* E_{\boldsymbol{X}} + \rho \,\mathrm{id})^{-1} E_{\boldsymbol{X}}^* u \qquad (5.2.7) \\
&\quad + (S_k \iota + \rho \,\mathrm{id})^{-1} (E_{\boldsymbol{X}}^* u - S_k f^\star). \qquad (5.2.8)
\end{aligned}$$

The derivations in [178, Eq. 19 and Eq. 20] (when corrected slightly) give

$$\begin{aligned}
\|\iota (S_k \iota + \rho \,\mathrm{id})^{-1}\|_{\mathrm{op}} &\le \frac{1}{2\sqrt{\rho}} \\
\|(E_{\boldsymbol{X}}^* E_{\boldsymbol{X}} + \rho \,\mathrm{id})^{-1} E_{\boldsymbol{X}}^*\|_{\mathrm{op}} &\le \frac{1}{2\sqrt{\rho}}.
\end{aligned}$$

Then with $\|u\|_n \le B$, Eq. (5.2.8) becomes

$$\|\iota \hat{f} - \iota f\|_\mu \le \frac{B}{4\rho} \|S_k \iota - E_{\boldsymbol{X}}^* E_{\boldsymbol{X}}\|_{\mathrm{op}} + \frac{1}{2\sqrt{\rho}} \|E_{\boldsymbol{X}}^* u - S_k f^\star\|_{\mathcal{H}}.$$

We apply Theorem 5.9 to get that with probability at least $1 - \delta$

$$\|\iota \hat{f} - \iota f\|_\mu \le \frac{B M_k^2}{8\rho} q\left(\frac{8}{n} \log(4/\delta)\right) + \frac{B M_k}{4\sqrt{\rho}} q\left(\frac{8}{n} \log(4/\delta)\right).$$

The proof is complete. □

Theorem 5.9 ([178, Theorem 3])**.** Assume the setting and notation of Theorem 5.8 and its proof. With probability at least $1 - \delta$ we have, simultaneously,

$$\begin{aligned}
\|S_k \iota - E_{\boldsymbol{X}}^* E_{\boldsymbol{X}}\|_{\mathrm{op}} &\le \frac{M_k^2}{2} q\left(\frac{8}{n} \log(4/\delta)\right) \\
\|E_{\boldsymbol{X}}^* u - S_k f^\star\|_{\mathcal{H}} &\le \frac{B M_k}{2} q\left(\frac{8}{n} \log(4/\delta)\right)
\end{aligned}$$

where $q(t) = \frac{1}{2}(t + \sqrt{t^2 + 4t})$ and $q(t) = \sqrt{t} + o(t)$ as $t \to 0$.

5.2.3.2 Variance Term

We write $N[k^\perp] = \|k^\perp\|^2_{L_2(\mu\otimes\mu)}$ and make the same definition for k and $\overline{k}$. Theorem 5.2 shows that the solution f to Eq. (5.2.1) can only learn to be invariant if the number of examples dominates $N[k^\perp]$. Intuitively, we think of $N[k^\perp]/M_k^2$ as an effective dimension of the space of $\mathcal{G}$-anti-symmetric functions in $\mathcal{H}$. With this interpretation, the variance term in Theorem 5.2 plays the same role as the $\dim A$ term in Theorem 4.1. Taking the ridgeless limit $\rho \to 0$, we view Theorem 5.2 as a generalisation of Theorem 4.1.

This interpretation of $N[k]$ as a measure of dimension or capacity will appear again for the linear kernel in Section 5.2.4. In Section 5.3 we associate $\mathcal{G}$-symmetric and $\mathcal{G}$-anti-symmetric RKHSs with $\overline{k}$ and $k^\perp$ respectively, which somewhat solidifies this perspective. In particular, the decomposition in Lemma 5.10 below will correspond to a decomposition of RKHSs $\mathcal{H} = \mathcal{O}\mathcal{H} \oplus (\mathrm{id} - \mathcal{O})\mathcal{H}$, the same as Lemma 3.1 but with the RKHS inner product. However, the interpretation of $N[k]$ as a dimension requires some refinement: $N[k]$ depends on the scale of k while any capacity measure of $\mathcal{H}$ should not, and the scale invariant $N[k]/M_k$ does not give the decomposition in Lemma 5.10.

Lemma 5.10.

$$N[k] = N[\overline{k}] + N[k^\perp]$$

Proof.

$$\begin{aligned}
N[k] &= \|k\|^2_{L_2(\mu\otimes\mu)} \\
&= \int_{\mathcal{X}} k(x,y)^2 \,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) \\
&= \int_{\mathcal{X}} \|k_x\|^2_\mu \,\mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} \|\overline{k_x}\|^2_\mu \,\mathrm{d}\mu(x) + \int_{\mathcal{X}} \|k_x^\perp\|^2_\mu \,\mathrm{d}\mu(x) \\
&= N[\overline{k}] + N[k^\perp]
\end{aligned}$$

□

5.2.4 Special Case: The Linear Kernel

In this section we explore the relation of Theorem 5.2 to Theorem 4.1 by specialising to the linear kernel. We begin with a calculation $N[k]$ for the linear kernel, which agrees with the interpretation as a dimension from the previous section. For the rest of this section we refer to the action ϕ of $\mathcal{G}$ explicitly, writing $\phi(g)x$ instead of gx.

Example 5.11 (Linear Kernel)**.** Let X and Y be mean zero isotropically distributed random vectors $\mathbb{R}^d$ whose co-ordinates have unit variance and let k be the linear kernel on $\mathbb{R}^d$ with $k(x,y) = x^\top y$, then

$$N[k] = \mathbb{E}[k(X,Y)^2] = \mathbb{E}[X_i X_j Y_i Y_j] = d.$$

Recall the matrix $\Phi = \int_{\mathcal{G}} \phi(g)\,\mathrm{d}\lambda(g)$ defined in Section 4.1 along with the $\mathcal{G}$-symmetric and $\mathcal{G}$-anti-symmetric subspaces S and A of $\mathbb{R}^d$ respectively. We have

$$\overline{k}(x,y) = \int_{\mathcal{G}} x^\top \phi(g) y\,\mathrm{d}\lambda(g) = x^\top \Phi y$$

so

$$N[\overline{k}] = \mathbb{E}\left[(X^\top \Phi Y)^2\right] = \|\Phi\|_{\mathrm{F}}^2 = \dim S$$

and $N[k^\perp] = d - \dim S = \dim A$ by Lemma 5.10

Theorem 5.12. Assume the setting and notation of Theorem 5.2 with $\sigma^2 = 1$. In addition, let $\mathcal{X} = \mathbb{S}_{d-1}(\sqrt{d})$ be the $(d-1)$-sphere of radius $\sqrt{d}$ with $d > 1$ and let $\mu = \operatorname{Unif}\mathcal{X}$. Let $\mathcal{G}$ act via an orthogonal representation ϕ on $\mathcal{X}$ and define the matrix $\Phi = \int_{\mathcal{G}} \phi(g)\,\mathrm{d}\lambda(g)$. Let $k(x,y) = x^\top y$ be the linear kernel and suppose $f^\star(x) = \theta^\top x$ for some $\theta \in \mathbb{R}^d$. Let K be the kernel Gram matrix $K_{ij} = k(X_i, X_j)$ and let $\gamma_1, \dots, \gamma_n$ be its eigenvalues. Define

$$\zeta_1(\rho) = \mathbb{E}\left[\sum_{i=1}^n \frac{\gamma_i^2}{(\gamma_i + \rho)^2}\right] \quad \text{and} \quad \zeta_2(\rho) = \mathbb{E}\left[\left(\sum_{i=1}^n \frac{\gamma_i}{\gamma_i + \rho}\right)^2\right].$$

Then the bound in Theorem 5.2 becomes

$$\mathbb{E}[R[f] - R[f']] \geq \frac{\|\theta\|_2^2(d\zeta_1(\rho) - \zeta_2(\rho))(d - \|\Phi\|_{\mathrm{F}}^2)}{d(d+2)(d-1)} + \frac{d - \|\Phi\|_{\mathrm{F}}^2}{(d\sqrt{n} + \rho/\sqrt{n})^2}$$

where the first and second terms are exactly the bias and variance terms in Theorem 5.2 respectively and each term is non-negative.

Proof. We will calculate each term in the bound in Theorem 5.2 separately. We make use of the Einstein summation convention, implicitly summing over repeated indices. The variance term is straightforward. We know from Example 5.11 that[2]

$$0 \leq \|k^{\perp}\|_{L_2(\mu\otimes\mu)}^2 = d - \|\Phi\|_{\mathrm{F}}^2$$

and the denominator comes from $M_k = \sup_x k(x,x) = \|x\|_2^2 = d$. The rest of the work is on the bias term. Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ have components $\boldsymbol{X}_{ij} = (X_i)_j$. Then

$$K_{ij} \coloneqq k(X_i, X_j) = X_i^\top X_j = (X_i)_l (X_j)_l = (\boldsymbol{X}\boldsymbol{X}^\top)_{ij}$$

and

$$k_{\boldsymbol{X}}(y)_i = k_{X_i}(y) = X_i^\top y = \boldsymbol{X}_{ij} y_j = (\boldsymbol{X}y)_i$$

along with $f^\star(\boldsymbol{X})_i = f^\star(X_i) = X_i^\top \theta = (\boldsymbol{X}\theta)_i$. So

$$\begin{aligned} \Lambda_{n,\rho} f^\star(y) &= k_{\boldsymbol{X}}(y)^\top (K + \rho I_n)^{-1} f^\star(\boldsymbol{X}) \\ &= (\boldsymbol{X}y)^\top (\boldsymbol{X}\boldsymbol{X}^\top + \rho I_n)^{-1} \boldsymbol{X}\theta \\ &= y^\top \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top + \rho I_n)^{-1} \boldsymbol{X}\theta \end{aligned}$$

where for $m \in \mathbb{N}$ we write I_m for the $m \times m$ identity matrix. With $\Phi^\perp = I_d - \Phi$ we

[2] As a sanity check, direct calculation gives

$$\|\Phi\|_{\mathrm{F}}^2 = \mathrm{Tr}(\Phi^\top \Phi) = \int_{\mathcal{G}} \mathrm{Tr}(\phi(g_1)^\top \phi(g_2)) \,\mathrm{d}\lambda(g_1)\,\mathrm{d}\lambda(g_2) \leq \left(\int_{\mathcal{G}} \|\phi(g)\|_{\mathrm{F}} \,\mathrm{d}\lambda(g)\right)^2 = d$$

using Cauchy-Schwarz on the Frobenius inner product and with $\|\phi(g)\|_{\mathrm{F}} = \sqrt{d}$ because the representation is orthogonal.

get (by linearity)

$$\begin{aligned}(\operatorname{id}-\mathcal{O})\Lambda_{n,\rho}f^\star(y) &= y^\top \Phi^\perp \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \rho I_n)^{-1}\boldsymbol{X}\theta \\ &= y^\top \Phi^\perp (\boldsymbol{X}^\top\boldsymbol{X} + \rho I_d)^{-1}\boldsymbol{X}^\top\boldsymbol{X}\theta\end{aligned}$$

which follows from

$$\begin{aligned}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \rho I_n)^{-1}\boldsymbol{X} &= (\boldsymbol{X}^\top\boldsymbol{X} + \rho I_d)^{-1}(\boldsymbol{X}^\top\boldsymbol{X} + \rho I_d)\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \rho I_n)^{-1}\boldsymbol{X} \\ &= (\boldsymbol{X}^\top\boldsymbol{X} + \rho I_d)^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + I_n)(\boldsymbol{X}\boldsymbol{X}^\top + I_n)^{-1}\boldsymbol{X} \\ &= (\boldsymbol{X}^\top\boldsymbol{X} + \rho I_d)^{-1}\boldsymbol{X}^\top\boldsymbol{X}.\end{aligned}$$

Write $P = (\boldsymbol{X}^\top\boldsymbol{X} + \rho I_d)^{-1}\boldsymbol{X}^\top\boldsymbol{X}$, so

$$\|(\operatorname{id}-\mathcal{O})\Lambda_{n,\rho}f^\star(y)\|_\mu^2 = \|\Phi^\perp P\theta\|_2^2$$

and

$$\mathbb{E}[\|(\operatorname{id}-\mathcal{O})\Lambda_{n,\rho}f^\star(y)\|_\mu^2] = \theta_i\theta_j\Phi^\perp_{lq}\Phi^\perp_{mq}\,\mathbb{E}[P_{il}P_{jm}]$$

We show that the 4-tensor $\mathbb{E}[P_{ab}P_{ce}]$ is isotropic. For any rotation matrix $R \in \mathsf{SO}_d$

$$R_{\alpha a}R_{\beta b}R_{\gamma c}R_{\epsilon e}\,\mathbb{E}[P_{ab}P_{ce}] = \mathbb{E}[(RPR^\top)_{\alpha\beta}(RPR^\top)_{\gamma\epsilon}]$$

then isotropy of the covariates implies $\boldsymbol{X}^\top\boldsymbol{X} \overset{\mathrm{d}}{=} R\boldsymbol{X}^\top\boldsymbol{X}R^\top$ for all $R \in \mathsf{SO}_d$ and hence

$$P \overset{\mathrm{d}}{=} (R\boldsymbol{X}^\top\boldsymbol{X}R^\top + \rho I)^{-1}R\boldsymbol{X}^\top\boldsymbol{X}R^\top = RPR^\top.$$

The general form of a real, isotropic 4-tensor is

$$\mathbb{E}[P_{ab}P_{ce}] = \alpha\delta_{ab}\delta_{ce} + \beta\delta_{ac}\delta_{be} + \gamma\delta_{ae}\delta_{cb}$$

for $\alpha, \beta, \gamma \in \mathbb{R}$ [79]. We can ignore α as will be clear in a moment and $P^\top = P$ so

$\beta = \gamma$. We have

$$\zeta_1(\rho) \coloneqq \mathbb{E}[\operatorname{Tr}(P^2)] = d\alpha + d(d+1)\beta$$
$$\zeta_2(\rho) \coloneqq \mathbb{E}[\operatorname{Tr}(P)^2] = d^2\alpha + 2d\beta$$

so

$$\beta = \frac{d\zeta_1(\rho) - \zeta_2(\rho)}{d(d+2)(d-1)}$$

and the bias term is

$$\begin{aligned}
\mathbb{E}[\|(\operatorname{id} - \mathcal{O})\Lambda_{n,\rho} f^\star(y)\|_\mu^2] &= \theta_a \theta_c \Phi^\perp_{bq} \Phi^\perp_{eq} \,\mathbb{E}[P_{ab} P_{ce}] \\
&= (\alpha + \beta)\|\Phi^\perp \theta\|_2^2 + \beta \|\theta\|_2^2 \|\Phi^\perp\|_{\mathrm{F}}^2 \\
&= \frac{\|\theta\|_2^2 \|\Phi^\perp\|_{\mathrm{F}}^2 (d\zeta_1(\rho) - \zeta_2(\rho))}{d(d+2)(d-1)}
\end{aligned}$$

where in the final line we used $\Phi^\perp \theta = 0$ by assumption that $f^\star$ is invariant. Note that $\|\Phi^\perp\|_{\mathrm{F}}^2 = d - \|\Phi\|_{\mathrm{F}}^2$. Finally, it's immediate by diagonalising that

$$\zeta_1(\rho) = \mathbb{E}\left[\sum_{i=1}^n \frac{\gamma_i^2}{(\gamma_i + \rho)^2}\right]$$

and

$$\zeta_2(\rho) = \mathbb{E}\left[\left(\sum_{i=1}^n \frac{\gamma_i}{\gamma_i + \rho}\right)^2\right].$$

The inequality $\zeta_2(\rho) \leq \min\{n, d\}\zeta_1(\rho)$ follows from $\|v\|_1^2 \leq m\|v\|_2^2$ for all $v \in \mathbb{R}^m$ and the number of non-zero eigenvalues of $K = \boldsymbol{X}\boldsymbol{X}^\top$. □

Theorem 5.12 is a special case of Theorem 5.2 with the linear kernel. From Example 5.11 we have $d - \|\Phi\|_{\mathrm{F}}^2 = \dim A$, highlighting a similarity between Theorems 4.1 and 5.12. These results are similar but not quite the same: the input spaces and distributions are different, while Theorem 5.12 has a ridge penalty but Theorem 4.1 does not. Furthermore, the linear kernel is unbounded on $\mathbb{R}^d$, so a direct application of Theorem 5.2 to the setting of Theorem 4.1 is not possible.

On the other hand, in the ridgeless limit $\rho \to 0$ Theorem 5.12 becomes

$$\mathbb{E}[R[f] - R[f']] \geq \frac{\|\theta\|_2^2 (d \min\{n, d\} - \min\{n, d\}^2) \dim A}{d(d+2)(d-1)} + \frac{\dim A}{(d\sqrt{n} + \rho/\sqrt{n})^2}.$$

So the bias term turns out to match Theorem 4.1 exactly: when $n \geq d$ it vanishes and when $n < d$ we recover the $n < d - 1$ case in Theorem 4.1. This is natural, as the bias term was calculated exactly in each case. The variance term in Theorem 5.2 is estimated which means there is no opportunity for double descent behaviour to appear in Theorem 5.12. One would hope that as $d \to \infty$ the two results look qualitatively similar because in this limit the normally distributed covariates in Theorem 4.1 become uniform on the sphere as in Theorem 5.12. Indeed if $d \to \infty$ and n remains fixed then the results match, but if $n, d \to \infty$ with $r = n/d$ finite the variance term above looks like $\frac{1}{r}\frac{\dim A}{d^3}$ while the variance term in Theorem 4.1, which gives an equality, goes like $\frac{r}{1-r}\frac{\dim A}{d}$ when $r < \frac{d-1}{d}$ and $\frac{1}{r-1}\frac{\dim A}{d}$ when $r > \frac{d+1}{d}$. This discrepancy is due to the M_k in the denominator. It's possible that it can be improved with a more sophisticated approach in the proof of Theorem 5.2.

5.3 Structure of $\mathcal{H}$

In this section we present Theorem 5.14, which is a version of Lemma 3.1 for RKHSs. If k satisfies, for all $x, x' \in \mathcal{X}$,

$$\int_{\mathcal{G}} k(gx, x') \, \mathrm{d}\lambda(g) = \int_{\mathcal{G}} k(x, gx') \, \mathrm{d}\lambda(g). \tag{5.3.1}$$

then $\mathcal{H}$ is an orthogonal direct sum of two subspaces, each of which is an RKHS, one of invariant functions and another of those that vanish when averaged over $\mathcal{G}$. Moreover, the kernels turn out to be $\overline{k}$ and $k^{\perp}$ respectively.

For Eq. (5.3.1) to hold, it is sufficient to have $k(gx, y) = k(x, g^{-1}y)$ and use the change of variables $g^{-1} \mapsto g$. Any kernel for which this stronger condition holds is known as a *unitary kernel*. Highlighting two special cases: any inner product kernel

$k(x, x') = \kappa(\langle x, x' \rangle)$ where the action of $\mathcal{G}$ is unitary with respect to $\langle \cdot, \cdot \rangle$ satisfies Eq. (5.3.1), as does any stationary kernel $k(x, x') = \kappa(\|x - x'\|)$ with a norm that is *preserved* by $\mathcal{G}$ in the sense that $\|gx - gx'\| = \|x - x'\|$ for all $g \in \mathcal{G}$ and all $x, x' \in \mathcal{X}$. Respectively, examples are the Euclidean inner product with orthogonal representations and permutations with any p-norm.

As it happens, if the kernel satisfies Eq. (5.3.1) then $\overline{k}$ qualifies as a Haar integration kernel, introduced by Haasdonk, Vossen, and Burkhardt [70] and defined by

$$\tilde{k}(x, x') = \int_{\mathcal{G}} k(gx, g'x') \,\mathrm{d}\lambda(g) \,\mathrm{d}\lambda(g')$$

for any kernel k. Simply apply Eq. (5.3.1) and use the invariance of λ to get $\overline{k}$. Note that this means that Eq. (5.3.1) implies that $\overline{k}$ is $\mathcal{G}$-invariant in both arguments.

Before we go further, for all $h \in \mathcal{H}$ we define the function $\mathcal{O}h$ by

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

Implicit in Theorem 5.14 is that $\mathcal{O} : \mathcal{H} \to \mathcal{H}$ is well defined. Note the slight abuse of notation, simultaneously writing $\mathcal{O}$ for the operator on different spaces. The map $\mathcal{O} : \mathcal{H} \to \mathcal{H}$ is not well-defined in general. Indeed there are trivial examples where it is not, such as Example 5.13.

Example 5.13. Define $k : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ by $k(x, y) = x^\top A y$ with $A = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$. It's easy to check that k is symmetric and positive definite, so there is a (unique) RKHS $\mathcal{H}$ with reproducing kernel k by the Moore-Aronszajn theorem. The RKHS consists of linear functions of the following form and the completion thereof

$$f(x) = \sum_{i=1}^{n} \alpha_i k(z_i, x) = \left(\sum_{i=1}^{n} \alpha_i A z_i \right)^\top x$$

for $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, $z_1, \ldots, z_n \in \mathbb{R}^2$ and some $n \in \mathbb{N}$. Any $f \in \mathcal{H}$ has $f(e_2) = 0$ where $e_2 = (0, 1)^\top$. Now consider $\mathcal{G} = \mathsf{S}_2$ acting on $\mathbb{R}^2$ by permutation of the coordinates. Then $\mathcal{O}f(x) = \frac{1}{2}f(x) + \frac{1}{2}f(Bx)$ where $B = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. If $f(x) = e_1^\top Ax$ where $e_1 = (1, 0)^\top$, then $\mathcal{O}f(e_2) = \frac{1}{2}$ so $\mathcal{O}f \notin \mathcal{H}$. It's easy to check that k does not satisfy Eq. (5.3.1).

Theorem 5.14. Suppose the kernel satisfies Eq. (5.3.1) and let $\overline{\mathcal{H}} = \{f \in \mathcal{H} : f \text{ is } \mathcal{G}\text{-invariant}\}$ and $\mathcal{H}_\perp = \{f \in \mathcal{H} : \mathcal{O}f = 0\}$, then:

- $\mathcal{H}$ admits the orthogonal decomposition $\mathcal{H} = \overline{\mathcal{H}} \oplus \mathcal{H}_\perp$
- $\overline{\mathcal{H}}$ is an RKHS with kernel $\overline{k}(x, y) = \int_\mathcal{G} k(x, gy)\,\mathrm{d}\lambda(g)$
- $\mathcal{H}_\perp$ is an RKHS with kernel $k^\perp(x, y) = k(x, y) - \overline{k}(x, y)$

Proof. First we show that $\mathcal{O} : \mathcal{H} \to \mathcal{H}$ is well defined. Let the image of $L_2(\mu)$ under S_k be $\mathcal{H}_2$. By Lemma 5.1 the completion of $\mathcal{H}_2$ in $\|\cdot\|_\mathcal{H}$ is $\mathcal{H}$. Then for all $f \in L_2(\mu)$

$$\begin{aligned}
\mathcal{O}S_k f(x) &= \int_\mathcal{G}\int_\mathcal{X} k(gx, t) f(t)\,\mathrm{d}\mu(t)\,\mathrm{d}\lambda(g) \\
&= \int_\mathcal{X}\int_\mathcal{G} k(gx, t) f(t)\,\mathrm{d}\lambda(g)\,\mathrm{d}\mu(t) \\
&= \int_\mathcal{X}\int_\mathcal{G} k(x, gt) f(t)\,\mathrm{d}\lambda(g)\,\mathrm{d}\mu(t) \\
&= \int_\mathcal{G}\int_\mathcal{X} k(x, gt) f(t)\,\mathrm{d}\mu(t)\,\mathrm{d}\lambda(g) \\
&= \int_\mathcal{G}\int_\mathcal{X} k(x, t) f(g^{-1}t)\,\mathrm{d}\mu(t)\,\mathrm{d}\lambda(g) \\
&= \int_\mathcal{X} k(x, t) \int_\mathcal{G} f(g^{-1}t)\,\mathrm{d}\lambda(g)\,\mathrm{d}\mu(t) \\
&= \int_\mathcal{X} k(x, t) \int_\mathcal{G} f(gt)\,\mathrm{d}\lambda(g)\,\mathrm{d}\mu(t) \\
&= S_k \mathcal{O} f(x)
\end{aligned}$$

where we used Eq. (5.3.1), Fubini's theorem, the invariance of μ and the compactness of $\mathcal{G}$ to substitute $g^{-1} \mapsto g$. This shows that $\mathcal{O} : \mathcal{H}_2 \to \mathcal{H}_2$ is well defined and that $\mathcal{O}$ and S_k commute. Let $a, b \in \mathcal{H}_2$ with preimages $a', b' \in L_2(\mu)$ such that $a = S_k a'$ and $b = S_k b'$, then, making use of the fact that $\mathcal{O}$ is self-adjoint on $L_2(\mu)$ from

Proposition 3.7 and that $[T_k, \mathcal{O}] = 0$ by Theorem 5.7,

$$\begin{aligned}\langle \mathcal{O}a, b\rangle_{\mathcal{H}} &= \langle \mathcal{O}S_k a', S_k b'\rangle_{\mathcal{H}} = \langle S_k \mathcal{O}a', S_k b'\rangle_{\mathcal{H}} = \langle \mathcal{O}a', T_k b'\rangle_{\mu} \\ &= \langle a', \mathcal{O}T_k b'\rangle_{\mu} = \langle a', T_k \mathcal{O}b'\rangle_{\mu} = \langle S_k a', S_k \mathcal{O}b'\rangle_{\mathcal{H}} \\ &= \langle S_k a', \mathcal{O}S_k b'\rangle_{\mathcal{H}} = \langle a, \mathcal{O}b\rangle_{\mathcal{H}}.\end{aligned}$$

From this it follows that, for all $f \in \mathcal{H}_2$,

$$\|\mathcal{O}f\|_{\mathcal{H}}^2 = \langle \mathcal{O}f, \mathcal{O}f\rangle_{\mathcal{H}} = \langle f, \mathcal{O}f\rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}}\|\mathcal{O}f\|_{\mathcal{H}}$$

so $\|\mathcal{O}f\|_{\mathcal{H}} \leq \|f\|_{\mathcal{H}}$. We can use this to show that $\mathcal{O} : \mathcal{H} \to \mathcal{H}$ is well-defined. Since $\mathcal{H}_2$ is dense in $\mathcal{H}$, for any $f \in \mathcal{H}$ there is a sequence $\{f_n\} \subset \mathcal{H}_2$ converging to f in $\|\cdot\|_{\mathcal{H}}$. We need to show that $\mathcal{O}f \in \mathcal{H}$. From above $\|\mathcal{O}f_n - \mathcal{O}f_m\|_{\mathcal{H}} \leq \|f_n - f_m\|_{\mathcal{H}}$ so the sequence $\{\mathcal{O}f_n\}$ is Cauchy in $\mathcal{H}$, so must converge to some $\tilde{f} \in \mathcal{H}$ by completeness. Note that convergence in $\|\cdot\|_{\mathcal{H}}$ implies uniform convergence since for all $h_1, h_2 \in \mathcal{H}$

$$|h_1(x) - h_2(x)| = |\langle k_x, h_1 - h_2\rangle_{\mathcal{H}}| \leq \sqrt{M_k}\|h_1 - h_2\|_{\mathcal{H}}.$$

Then for all $x \in \mathcal{X}$ and all $n \in \mathbb{N}$

$$\begin{aligned}|\mathcal{O}f(x) - \tilde{f}(x)| &\leq |\mathcal{O}f(x) - \mathcal{O}f_n(x)| + |\mathcal{O}f_n(x) - \tilde{f}| \\ &\leq |\mathcal{O}f(x) - \mathcal{O}f_n(x)| + \sqrt{M_k}\|\mathcal{O}f_n - \tilde{f}\|_{\mathcal{H}}\end{aligned}$$

and the first term can be controlled by

$$|\mathcal{O}f(x) - \mathcal{O}f_n(x)| \leq \int_{\mathcal{G}} |f(gx) - f_n(gx)| \,\mathrm{d}\lambda(g) \leq \sup_{t\in\mathcal{X}} |f(t) - f_n(t)| \leq \sqrt{M_k}\|f - f_n\|_{\mathcal{H}}$$

so $\tilde{f} = \mathcal{O}f$ and $\mathcal{O}f \in \mathcal{H}$.

Now we prove the orthogonal decomposition of $\mathcal{H}$. We show that $\mathcal{O}$ is self-adjoint on $\mathcal{H}$. Let $f, h \in \mathcal{H}$, then there are sequences $\{f_m\}$ and $\{h_n\}$ in $\mathcal{H}_2$ with limits f

and h respectively. We just showed that $\mathcal{O}$ commutes with taking limits of these sequences. So, taking $m \to \infty$ first,

$$\begin{aligned}\langle \mathcal{O} f, h\rangle_{\mathcal{H}} = \langle \mathcal{O} \lim_{m\to\infty} f_m, \lim_{n\to\infty} h_n\rangle_{\mathcal{H}} &= \lim_{n\to\infty}\lim_{m\to\infty} \langle \mathcal{O} f_m, h_n\rangle_{\mathcal{H}} \\ &= \lim_{n\to\infty}\lim_{m\to\infty} \langle f_m, \mathcal{O} h_n\rangle_{\mathcal{H}} = \langle h, \mathcal{O} f\rangle_{\mathcal{H}}.\end{aligned}$$

Now we know that $\mathcal{O} : \mathcal{H} \to \mathcal{H}$ is self-adjoint, it's easy to check that it's idempotent so is an orthogonal projection on $\mathcal{H}$ (see Section 3.1.1). Let h_S have eigenvalue 1 and h_A have eigenvalue 0 under $\mathcal{O}$, then $\langle h_S, h_A\rangle_{\mathcal{H}} = \langle \mathcal{O} h_S, h_A\rangle_{\mathcal{H}} = \langle h_S, \mathcal{O} h_A\rangle_{\mathcal{H}} = 0$. Therefore, by linearity, for all $f \in \mathcal{H}$ we can write $f = \bar{f} + f^{\perp}$ where $\bar{f} = \mathcal{O} f \in \overline{\mathcal{H}}$ is $\mathcal{G}$-invariant and $f^{\perp} = f - \mathcal{O} f \in \mathcal{H}_{\perp}$ and these terms are orthogonal.

We conclude by showing that $\overline{\mathcal{H}}$ and $\mathcal{H}_{\perp}$ are RKHSs with kernels $\overline{k}$ and $k^{\perp}$ respectively. By the linearity of $\mathcal{O}$, $\overline{\mathcal{H}} = \mathcal{O}\mathcal{H} \subset \mathcal{H}$ is a subspace, so also an inner product space with $\langle \cdot, \cdot\rangle_{\mathcal{H}}$. We check the completeness of $\overline{\mathcal{H}}$. Above we already showed that if $\{f_n\}$ be a sequence in $\overline{\mathcal{H}}$ which has limit $f \in \mathcal{H}$ by completeness of $\mathcal{H}$, then $\lim_{n\to\infty} \mathcal{O} f_n = \mathcal{O} \lim_{n\to\infty} f_n = \mathcal{O} f \in \overline{\mathcal{H}}$, which shows that $\overline{\mathcal{H}}$ is complete. The same argument works to show that $\mathcal{H}_{\perp}$ is complete. The evaluation functional is continuous on each of these subspaces so each is an RKHS. Now for all $h_S \in \overline{\mathcal{H}}$ we have

$$h_S(x) = \langle h_S, k_x\rangle_{\mathcal{H}} = \langle h_S, \mathcal{O} k_x\rangle_{\mathcal{H}}$$

and the uniqueness afforded by the Riesz representation theorem for Hilbert spaces [143, Theorem 4.12] tells us that the reproducing kernel for $\overline{\mathcal{H}}$ is

$$\langle \mathcal{O} k_x, \mathcal{O} k_y\rangle_{\mathcal{H}} = \langle k_x, \mathcal{O} k_y\rangle_{\mathcal{H}} = \int_{\mathcal{G}} k(x, gy)\, \mathrm{d}\lambda(g) = \overline{k}(x, y).$$

From the above it can be seen that $\overline{k}$ is positive definite and symmetric. The same steps using $\mathrm{id} - \mathcal{O}$ show that $k^{\perp}$ is the reproducing kernel of $\mathcal{H}_{\perp}$ and also verify that it is positive definite and symmetric. □

Remark 5.15. It is straightforward to use the orthogonality in Theorem 5.14 to derive an invariant version of the representer theorem. That is, that any invariant minimiser of the of the appropriate risk functional is an element of the linear span of the evaluations of $\overline{k}$ at the training data. The same statement but for equivariance was shown by Reisert and Burkhardt [139].

Chapter 6

General Theory II: Orbit Representatives

Summary

In previous chapters we studied invariance and equivariance by considering the action of certain averaging operators on a function space. In this chapter we take a different approach, leveraging the observation that an invariant function can be specified by its values on one representative from each orbit of $\mathcal{X}$ under $\mathcal{G}$. We show rigorously how learning with invariant or equivariant hypotheses reduces to learning on a set of orbit representatives. In addition, we show how to use these equivalences to derive a sample complexity bound for learning invariant/equivariant classes with empirical risk minimisation in terms of the geometry of the input and output spaces.

6.1 Preliminaries

In this chapter, $\mathcal{Y}$ is not restricted to be $\mathbb{R}^k$ and $(\mathcal{Y}, \mathcal{S}_\mathcal{Y})$ can be any standard Borel space. The action ψ is an arbitrary action of $\mathcal{G}$ on $\mathcal{Y}$ and not necessarily a linear representation unless explicitly stated. We will write $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and the action of $\mathcal{G}$ on $\mathcal{Z}$ is defined by $g(x, y) = (gx, gy)$. A set of functions $\mathcal{F}$ is *invariant* or *equivariant* if all its elements are invariant or equivariant respectively. We say the action of $\mathcal{G}$ on $\mathcal{X}$ is *free* if the non-trivial elements of $\mathcal{G}$ have no fixed points, that is $\forall x \in \mathcal{X} : gx = x \implies g = e$ the identity element. If the action is free then ϕ is injective as a function from $\mathcal{G}$ to bijections on $\mathcal{X}$. Let P and Q be random variables, we use the notation $A \sim (P, Q)^n$ to mean the tuple $A = ((P_1, Q_1), \dots, (P_n, Q_n))$ where the (P_i, Q_i) are i.i.d. and distributed independently of and identically to (P, Q).

Definition 6.1. A *task* is a tuple $\mathsf{T} = (X, Y, \ell)$ where X is a random element of $\mathcal{X}$, Y is a random element of $\mathcal{Y}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is an integrable function called the *loss function.*

We will assume that for any function class $\mathcal{F}$ and task $\mathsf{T} = (X, Y, \ell)$ considered in this chapter $\mathbb{E}[\ell(f(X), Y)] < \infty$.

6.1.1 Learning

The PAC framework [155, Definition 3.3], originally due to Valiant [170], provides a precise definition of learning from data. We use a distribution dependent relaxation of the agnostic formulation from Haussler [76]. What we arrive at can be considered a form of uniform learning [172]. A *hypothesis class* $\mathcal{F}$ is a set of functions from $\mathcal{X} \to \mathcal{Y}$, its elements are *hypotheses.* An *algorithm* $\texttt{alg} : \cup_{i \in \mathbb{N}} \mathcal{Z}^i \to \mathcal{F}$ is a map that associates with any tuple of observations an element of $\mathcal{F}$. We say that $\texttt{alg}$ *learns* $\mathcal{F}$ with respect to a task $\mathsf{T} = (X, Y, \ell)$ if $\exists m : (0,1)^2 \to \mathbb{N}$ such that $\forall \epsilon, \delta \in (0, 1)$, if $n \geq m(\epsilon, \delta)$ then

$$\mathbb{P}\left(\mathbb{E}[\ell(f_S(X), Y) \mid S] \leq \inf_{f \in \mathcal{F}} \mathbb{E}[\ell(f(X), Y)] + \epsilon\right) \geq 1 - \delta$$

where $f_S = \texttt{alg}(S)$ and $S \sim (X, Y)^n$. Throughout this chapter, we assume that T, $\mathcal{F}$ and $\texttt{alg}$ are such that the expectations in the above exist and are finite. Suppose $\texttt{alg}$ learns $\mathcal{F}$ with respect to T and let the set of all m satisfying the above be $\mathcal{M}$, we define the *sample complexity* of $\texttt{alg}$ on $\mathcal{D}$ as the pointwise minimum $m_{\texttt{alg},\mathsf{T}}(\epsilon, \delta) = \min_{m \in \mathcal{M}} m(\epsilon, \delta)$.

6.1.2 Invariant Algorithms

We say that an algorithm $\texttt{alg}$ is $\mathcal{G}$-*invariant* if

$$\texttt{alg}((g_1x_1, g_1y_1), \dots, (g_nx_n, g_ny_n)) = \texttt{alg}((x_1, y_1), \dots, (x_n, y_n))$$

$\forall (x_i, y_i) \in \mathcal{Z}$, $\forall g_i \in \mathcal{G}$ and $\forall n \in \mathbb{N}$. Often we will just say *invariant.* If the hypothesis class is invariant and $\mathcal{G}$ acts trivially on $\mathcal{Y}$, then any algorithm that depends on the training inputs only through the values of the hypotheses satisfies this property. For instance, any form of empirical risk minimisation is covered by this. Similarly, if the hypothesis class is equivariant and the algorithm depends on the data only through a loss that is *preserved* by $\mathcal{G}$, meaning $\ell(gy, gy') = \ell(y, y')$ $\forall g \in \mathcal{G}$ $\forall y, y' \in \mathcal{Y}$, then it will be an invariant algorithm. Once again, any form of empirical risk minimisation is covered. Highlighting two cases: if $\mathcal{G}$ acts via a unitary representation, then $\ell(y, y') = l(\langle y, y' \rangle)$ is preserved by $\mathcal{G}$ for any l. The same goes for $\ell(y, y') = l(\|y - y'\|_2)$.

Most of this chapter is about establishing the equivalence of learning problems when the hypothesis class is invariant or equivariant. With this in mind, we clarify what we mean by equivalence. Intuitively, two tasks are $(\mathcal{F}, \mathcal{G})$-equivalent if it's equally difficult to learn $\mathcal{F}$ with respect to them using any $\mathcal{G}$-invariant algorithm.

Definition 6.2. We say the tasks T and T' are $(\mathcal{F}, \mathcal{G})$-*equivalent* if, for all $\mathcal{G}$-invariant algorithms $\texttt{alg}$, $\texttt{alg}$ learns $\mathcal{F}$ with respect to T if and only if $\texttt{alg}$ learns $\mathcal{F}$ with respect to T', and the sample complexities are equal, i.e., $m_{\texttt{alg},\mathsf{T}} = m_{\texttt{alg},\mathsf{T}'}$.

6.1.3 Orbit Representatives

Definition 6.3 ([49, Definition 4.1])**.** Let $\mathcal{G}$ be a group acting measurably on a Borel space $(\mathcal{X}, \mathcal{S}_{\mathcal{X}})$. The set $\mathcal{X}_\pi \subseteq \mathcal{X}$ is a *measurable cross-section* of $\mathcal{X}$ with respect to $\mathcal{G}$ if the following conditions hold:

1) $\mathcal{X}_\pi$ is measurable.

2) $\mathcal{X}_\pi$ contains exactly one element from each orbit of each point $x \in \mathcal{X}$, say x_π.

3) The function $\pi : \mathcal{X} \to \mathcal{X}_\pi$ defined by $\pi(x) = x_\pi$ is $\mathcal{S}_{\mathcal{X}}$-measurable when $\mathcal{X}_\pi$ has the σ-algebra $\{B \cap \mathcal{X}_\pi : B \in \mathcal{S}_{\mathcal{X}}\}$. We call π the *projection.*

A measurable cross-section provides a natural way of identifying a set of orbit representatives with suitable measurability properties. Measurable cross-sections are not necessarily unique.

We maintain some measurable cross-section $\mathcal{X}_\pi$ with projection π in the background throughout this chapter. It is arbitrary insofar as any additional assumptions are satisfied.

6.1.4 Assumption on the Law of X

Departing from the rest of this thesis, in this chapter we do not assume that X is invariant in distribution unless explicitly stated. However, we will often make the following (weaker) assumption.

Assumption 1. We assume X is such that there exists a probability measure ν on $\mathcal{G}$ such that $X \overset{\mathrm{d}}{=} GX_\pi$ where $X_\pi = \pi(X)$, $G \sim \nu$ and $G \perp\!\!\!\perp X_\pi$.

Essentially, Assumption 1 says that the orbit that X belongs to and where it is in the orbit are independent. For intuition, if $\mathcal{G} = \mathsf{SO}_2$ acts by rotation on $\mathbb{R}^2$ then any density $f(r, \theta)$ that is separable in polar co-ordinates as $f(r, \theta) = f_{\mathtt{rad}}(r) f_{\mathtt{ang}}(\theta)$ gives the required independence.

The distribution ν and the choice of cross-section are not independent in general. In particular, if the action is free, under the map $\mathcal{X}_\pi \mapsto g\mathcal{X}_\pi$ we have $\nu(A) \mapsto \nu(Ag)$ for all measurable $A \subset \mathcal{G}$.[1] The dependence goes away only if ν is right-invariant, which by compactness of $\mathcal{G}$ and the uniqueness from Theorem 1.4 means $\nu = \lambda$,

[1] Let $g \in \mathcal{G}$ and $\mathcal{X}_{\pi'} = g\mathcal{X}_\pi$. Using Assumption 1 we can write $X = G\pi(X) = G'\pi'(X) = G'g\pi(X)$ so, because the action is free, for all measurable $A \subset \mathcal{G}$

$$G'^{-1}(A) = \{\omega \in \Omega : G'(\omega) \in A\} = \{\omega \in \Omega : G(\omega)g^{-1} \in A\} = \{\omega \in \Omega : G(\omega) \in Ag\} = G^{-1}(Ag).$$

where λ is the Haar measure on $\mathcal{G}$. Conversely, there is the following result.

Theorem 6.4 ([49, Theorem 4.4])**.** Let $G \sim \lambda$ and let $\mathcal{X}_\pi$ be a measurable cross-section of $\mathcal{X}$ with respect to $\mathcal{G}$, then the following are equivalent:

(i) $X \overset{\mathrm{d}}{=} gX$ for all $g \in \mathcal{G}$.

(ii) There exists a random variable X_π taking values on $\mathcal{X}_\pi$ such that $X \overset{\mathrm{d}}{=} GX_\pi$ and $G \perp\!\!\!\perp X_\pi$.

6.2 Learning with Invariant Models

In this section we present Theorem 6.5, which is a rigorous version of the common intuition that learning with an invariant model is equivalent to learning on a space of orbit representatives. We show that these learning problems have the same sample complexity. The class of orbit representatives may have much smaller dimension or complexity than the input space, which could result in a reduction in sample complexity for invariant models.

Theorem 6.5. Suppose $\mathcal{G}$ acts measurably on $\mathcal{X}$ and trivially on $\mathcal{Y}$. Let $\mathcal{X}_\pi$ be a measurable cross-section of $\mathcal{X}$ with projection π. Let $\mathcal{F}$ be a hypothesis class of $\mathcal{G}$-invariant functions. Let $\mathsf{T} = (X, Y, \ell)$ be any task, write $X_\pi = \pi(X)$ and define $\mathsf{T}_\pi = (X_\pi, Y, \ell)$. Let X satisfy Assumption 1. Then T and T_π are $(\mathcal{F}, \mathcal{G})$-equivalent.

Proof. For any $S = ((X_1, Y_1), \ldots, (X_n, Y_n))$ define $S_\pi = ((\pi(X_1), Y_1), \ldots, (\pi(X_n), Y_n))$. If $S \sim (X, Y)^n$ then $S_\pi \sim (X_\pi, Y)^n$. Let `alg` be a $\mathcal{G}$-invariant algorithm. Set $f_S = \mathtt{alg}(S)$ and $f_{S_\pi} = \mathtt{alg}(S_\pi)$. The invariance of `alg` implies $f_S = f_{S_\pi}$. By the invariance of $\mathcal{F}$, $f(X) = f(X_\pi)$ for all $f \in \mathcal{F}$. Together with the independence in Assumption 1, this means that $\mathbb{E}[\ell(f_S(X), Y) \mid S] = \mathbb{E}[\ell(f_{S_\pi}(X_\pi), Y) \mid S_\pi]$.

We can then conclude that

$$\begin{aligned}&\mathbb{P}\left(\mathbb{E}[\ell(f_S(X),Y) \mid S] \leq \inf_{f\in\mathcal{F}} \mathbb{E}[\ell(f(X),Y)] + \epsilon\right) \\ &= \mathbb{P}\left(\mathbb{E}[\ell(f_{S_\pi}(X_\pi),Y) \mid S_\pi] \leq \inf_{f\in\mathcal{F}} \mathbb{E}[\ell(f(X_\pi),Y)] + \epsilon\right)\end{aligned}$$

which completes the proof. □

Example 6.6. Let $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Y} = \{0,1\}$. Theorem 6.5 tells us that learning with a rotationally invariant hypothesis class, such as discs about the origin $\mathcal{F} = \{(x,y) \mapsto \mathbb{1}\left\{x^2+y^2 \leq r\right\} : r \in \mathbb{R}_+\}$, is equivalent to learning on the reduced space $\mathcal{X} = \mathbb{R}_+$.

Example 6.7 (Deep Sets, [195])**.** Zaheer, Kottur, Ravanbakhsh, Poczos, Salakhutdinov, and Smola [195] and Bloem-Reddy and Teh [22] consider learning functions $f : [0,1]^d \to \mathbb{R}$ that are S_d-invariant, where S_d is the group of permutations on d elements. It is shown that any such continuous function must be of the form $f(T) = f_1\left(\sum_{t\in T} f_2(t)\right)$ where $T \in [0,1]^d$ and $f_1, f_2 : \mathbb{R} \to \mathbb{R}$. By Theorem 6.5 we can see, as was shown by Sannai, Imaizumi, and Kawano [145], that learning permutation invariant functions is equivalent to learning the same class of functions restricted to the domain $\{x_1 \geq x_2 \geq \cdots \geq x_d; x_i \in [0,1]\}$. In high dimensions this is a much smaller space than $[0,1]^d$, it is a factor of $d!$ smaller in volume.

Example 6.8 (G-CNN [34])**.** Cohen and Welling [34] present a convolutional layer that is equivariant to various discrete groups of transformations. These layers are used to generate features for an invariant classifier. Consider the simple case of the group $p4$ of rotations about the origin in $\mathbb{R}^2$ through an angle of $\frac{\pi}{2}$. Theorem 6.5 shows that training a $p4$-invariant G-CNN based classifier (e.g., by gradient descent on the empirical loss) is equivalent to learning the network restricted to a single quadrant of the plane.

For the final corollary we need an additional definition.

Definition 6.9 (Packing, packing number)**.** Let (T,τ) be a metric space and $\epsilon > 0$.

$E \subset T$ is an ϵ-*packing* of T (with respect to τ) if for all $x, y \in E$, $\tau(x, y) > \epsilon$. The ϵ-*packing* number is the largest cardinality of all the ϵ-packings, $\texttt{Pack}(T, \tau, \epsilon) = \sup_{E \in \mathcal{E}} |E|$ where $\mathcal{E}$ is the set of all ϵ-packings. If the supremum doesn't exist, then we say the packing number is infinite.

Corollary 6.10. Let $(\mathcal{X}, \tau)$ be a compact metric space, let $\mathcal{Y} = \{0, 1\}$ and let $\mathcal{F}$ be a hypothesis class (a set of binary classifiers). Assume that the functions in $\mathcal{F}$ are γ-*robust*, meaning $\forall f \in \mathcal{F} : f(x) \neq f(x') \implies \tau(x, x') \geq \gamma$. This implies that the VC-dimension of $\mathcal{F}$ is bounded $\mathrm{VC}(\mathcal{F}) \leq \texttt{Pack}(\mathcal{X}, \tau, \gamma)$. If in addition we assume that $\mathcal{F}$ is invariant then we know from Theorem 6.5 that $\mathrm{VC}(\mathcal{F}) \leq \texttt{Pack}(\mathcal{X}_\pi, \tau, \gamma) \leq \texttt{Pack}(\mathcal{X}, \tau, \gamma)$, suggesting a distribution independent sample complexity improvement for invariant hypotheses.

6.3 Learning with Equivariant Models

Before we present our result for equivariant learning problems, we introduce an additional assumption. We relate Assumption 2 to the work of Bloem-Reddy and Teh [22] in Section 6.3.1.

Assumption 2. We assume that there exists a measurable $f : \mathcal{X} \times [0, 1] \to \mathcal{Y}$ such that $Y \overset{\mathrm{d}}{=} f(X, \eta)$, where $\eta \sim \mathrm{Unif}[0, 1]$, $\eta \perp\!\!\!\perp X$ and f satisfies the equivariance property $f(gX, \eta) \overset{\mathrm{d}}{=} g f(X, \eta)$.

One can verify that, with reference to Assumption 1, $\eta \perp\!\!\!\perp G$ and $\eta \perp\!\!\!\perp X_\pi$.

A special case of Assumption 2 is an equivariant target function with independent additive noise $f(X, \eta) = f^*(X) + \xi$ where ξ is such that $g\xi \overset{\mathrm{d}}{=} \xi$ for all $g \in \mathcal{G}$ (e.g., orthogonal representation on $\mathcal{Y}$ and Gaussian noise). Assumption 2 is more general and allows for stochastic equivariant functions and noise corruption that is not necessarily additive. This is inspired by *noise outsourcing*, which typically appears with almost sure equality rather than equality in distribution. It is also known as *transfer* [85, Theorem 6.10]. See [22] for an application to invariance/equivariance.

Theorem 6.11. Let $\mathcal{G}$ act measurably on both $\mathcal{X}$ and $\mathcal{Y}$. Let $\mathcal{F}$ be a $\mathcal{G}$-equivariant hypothesis class. Let (X, Y) satisfy Assumption 1 and Assumption 2. Then the tasks $T = (X, Y, \ell)$ and $T_\pi = (X_\pi, Y_\pi, \bar{\ell})$ are $(\mathcal{F}, \mathcal{G})$-equivalent, where $X_\pi = \pi(X)$, $Y_\pi = f(X_\pi, \eta)$ with f and η as in Assumption 2 and $\bar{\ell}(y, y') = \int_{\mathcal{G}} \ell(gy, gy')\, \mathrm{d}\nu(g)$.

Proof. Let $S \sim (X, Y)^n$ and $S_\pi \sim (X_\pi, Y_\pi)^n$. We have from Assumption 1 that $X \overset{\mathrm{d}}{=} GX_\pi$ and from Assumption 2 that

$$Y \overset{\mathrm{d}}{=} f(X, \eta) \overset{\mathrm{d}}{=} f(GX_\pi, \eta) \overset{\mathrm{d}}{=} Gf(X_\pi, \eta) = GY_\pi.$$

Let `alg` be an invariant algorithm, then we have $f_S \coloneqq \texttt{alg}(S) = \texttt{alg}(S_\pi) \eqqcolon f_{S_\pi}$ and hence by Assumption 1 we have $\mathbb{E}[\ell(f_S(X), Y) \mid S] = \mathbb{E}[\ell(f_{S_\pi}(X), Y) \mid S_\pi]$. Then, for all $h \in \mathcal{F}$, Assumption 1, Assumption 2 and equivariance gives

$$\begin{aligned}
\mathbb{E}[\ell(h(X), Y)] &= \mathbb{E}[\ell(h(GX_\pi), f(GX_\pi, \eta))] \\
&= \mathbb{E}[\ell(Gh(X_\pi), Gf(X_\pi, \eta))] \\
&= \mathbb{E}[\bar{\ell}(h(X_\pi), Y_\pi)]
\end{aligned}$$

where we applied Fubini's theorem. The function $(g, y, y') \mapsto \ell(gy, gy')$ is measurable by composition (see the proof of Proposition 3.4 for analogous details) so every $(\mathcal{Y} \times \mathcal{Y})$-section $\ell_{y,y'}(g) = \ell(gy, gy')$ is ν-measurable by Lemma A.1.2, so $\bar{\ell}$ exists. Furthermore, all of the above expectations must be finite because the left hand side is finite by assumption. Putting everything together concludes the proof

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{E}[\ell(f_S(X), Y) \mid S] \leq \inf_{f \in \mathcal{F}} \mathbb{E}[\ell(f(X), Y)] + \epsilon\right) \\
&= \mathbb{P}\left(\mathbb{E}[\bar{\ell}(f_{S_\pi}(X_\pi), Y_\pi) \mid S_\pi] \leq \inf_{f \in \mathcal{F}} \mathbb{E}[\bar{\ell}(f(X_\pi), Y_\pi)] + \epsilon\right).
\end{aligned}$$

□

The loss function $\bar{\ell}$ is the average of ℓ over the orbits of $\mathcal{G}$, weighted by the probability of each $g \in \mathcal{G}$. If ℓ is preserved by $\mathcal{G}$, that is $\ell(gy, gy') = \ell(y, y')$ for all $y, y' \in \mathcal{Y}$, we have $\bar{\ell} = \ell$. One can think of Y_π as the canonical target corresponding to the canonical input X_π. The task (X_π, Y_π, ℓ) can be thought of as a canonical version of the original task, with any nuisance information from the group transformations removed. This canonical task depends the choice of cross-section.

Example 6.12 (Deep Sets, [195])**.** Returning to [195], the authors consider neural network layers $f : \mathbb{R}^d \to \mathbb{R}^d$ with $f(x) = \sigma(\Theta x)$, where $\Theta \in \mathbb{R}^{d\times d}$ and σ is an element-wise non-linearity. It is shown that f is S_d-equivariant if and only if $\Theta_{ij} = a\delta_{ij} + b$ for scalars $a, b \in \mathbb{R}$. In this case $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and S_d acts on each space by permutation. Assume that the marginal distribution on the inputs is exchangeable, so ν exists and is uniform on S_d as outlined at the end of Section 6.1.4, then Theorem 6.11 says that learning the class of equivariant f is equivalent to learning on the restricted domain $\{x_1 \geq x_2 \geq \cdots \geq x_d ; x_i \in \mathcal{X}\}$ with the averaged loss function $\bar{\ell}(y, y') = \frac{1}{d!}\sum_{s\in\mathsf{S}_d} \ell(y_s, y'_s)$ where $(y_s)_i = y_{s(i)}$ and the same for y'.

6.3.1 Discussion of Assumption 2

It is known that, under certain conditions, a functional representation for Y similar to the one in Assumption 2 is equivalent to the conditional equivariance of Y, in the sense that $gY \mid gX \overset{\mathrm{d}}{=} Y \mid X \ \forall g \in \mathcal{G}$ [22]. We describe this setting below.

Assume that the distribution of X is $\mathcal{G}$-invariant, so $X \overset{\mathrm{d}}{=} gX \ \forall g \in \mathcal{G}$. We adapt the following definition from Bloem-Reddy and Teh [22].

Definition 6.13 (Representative Equivariant)**.** Let $\mathcal{G}$ be a group acting freely on a set $\mathcal{X}$. A *representative equivariant* is an equivariant function $\tau : \mathcal{X} \to \mathcal{G}$. That is, $\tau(gx) = g\tau(x) \ \forall g \in \mathcal{G} \ \forall x \in \mathcal{X}$.

The following result from Bloem-Reddy and Teh [22] then gives us Assumption 2.

Theorem (Bloem-Reddy and Teh [22, Theorem 9])**.** Let $\mathcal{G}$ be a compact group acting measurably on Borel spaces $\mathcal{X}$ and $\mathcal{Y}$ such that there exists a measurable representative equivariant $\tau : \mathcal{X} \to \mathcal{G}$. Let X be a $\mathcal{G}$-invariant random element of $\mathcal{X}$. Then Y is conditionally $\mathcal{G}$-equivariant if and only if there's a measurable $\mathcal{G}$-

equivariant function $f : \mathcal{X} \times [0,1] \to \mathcal{Y}$ such that $Y \overset{\text{a.s.}}{=} f(X, \eta)$ where $\eta \sim \text{Unif}[0,1]$ and $\eta \perp\!\!\!\perp X$.

6.4 Implication for Empirical Risk Minimisation

We apply the equivalences derived in previous sections to derive improved sample complexity guarantees for equivariant models when learning with empirical risk minimisation. In particular, Theorem 6.15 links the *generalisation error*, defined to be the difference between the risk and the empirical risk, to the geometries of the input and output spaces. Using invariant or equivariant hypotheses reduces the learning problem to one on a cross-section $\mathcal{X}_\pi$. Often, $\mathcal{X}_\pi$ will be smaller than $\mathcal{X}$ and, because the sample complexity depends a notion of the size of the input space, we get a reduction for invariant/equivariant models. We use covering numbers for this notion of size.

Definition 6.14 (Covering, covering number)**.** Let (T, τ) be a metric space and let $U \subset T$. $K \subset T$ is an *ϵ-cover* of U if $\forall u \in U \; \exists k \in K$ with $\tau(u, k) \leq \epsilon$. Let $\mathcal{K}$ be the set of all ϵ-covers of U, the *ϵ-covering number* of U is the smallest cardinality of all the ϵ-covers $\texttt{Cov}(U; \tau, \epsilon) = \inf_{K \in \mathcal{K}} |K|$.

Theorem 6.15 relies upon Proposition 6.16 which is deferred to Section 6.4.1. The basic idea of this proof is to use the Lipschitz property to bound coverings of the hypothesis class in terms of coverings of the input space. Similar ideas appeared in [165, 145].

Theorem 6.15. Let $(\mathcal{X}, \tau)$ be a metric space and let $\mathcal{Y} \subset \mathbb{R}^d$ be convex. Let $\mathcal{F}$ be a hypothesis class of functions $f : \mathcal{X} \to \mathcal{Y}$ that are L-Lipschitz $\|f(x) - f(x')\|_\infty \leq L\tau(x, x') \; \forall x, x' \in \mathcal{X}$. Let $\ell(y, y') : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be measurable loss function that's bounded $\ell(y, y') \leq M$ and Lipschitz in its first argument $|\ell(y_1, y') - \ell(y_2, y')| \leq C_\ell \|y_1 - y_2\|_\infty$ for some $C_\ell \in \mathbb{R}_+$ that's independent of y_1, y_2 and y'. Let $S \sim (X, Y)^n$

be a training sample. Then, for all $\epsilon \in (0,1)$,

$$\begin{aligned}&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\mathbb{E}[\ell(f(X),Y)] - \frac{1}{n}\sum_{i=1}^{n}\ell(f(X_i),Y_i)\right| \geq \epsilon\right)\\ &\leq 2\inf_{\alpha\in(0,1)}\exp\left(D_{\alpha\epsilon}(\mathcal{X},\mathcal{F}) - \frac{1}{8}(1-\alpha)^2 n\epsilon^2 M^{-2}\right)\end{aligned}$$

where

$$D_t(\mathcal{X},\mathcal{F}) = \texttt{Cov}\left(\mathcal{X},\tau,\frac{t}{12LC_\ell}\right)\sup_{x\in\mathcal{X}}\log\left(\texttt{Cov}\left(\mathcal{F}(x),\|\cdot\|_\infty,\frac{t}{12L^2C_\ell}\right)\right)$$

and $\mathcal{F}(x) = \{f(x) : f \in \mathcal{F}\} \subset \mathcal{Y}$.

Proof. We estimate the covering number of $\mathcal{F}$ and plug it into Proposition 6.16. Let E be a minimal δ-cover for $\mathcal{X}$. We may assume that $|E| < \infty$, because otherwise the result is trivial. For all $x \in \mathcal{X}$ let F_x be a minimal κ-cover of $\{f(x) : f \in \mathcal{F}\} \subset \mathcal{Y}$ in the vector norm $\|\cdot\|_\infty$ on $\mathcal{Y}$. Let H_E be the set of all functions $h_E : E \to \mathcal{Y}$ such that $\forall x \in E$, $h_E(x) \in F_x$. For all $x \in \mathcal{X}$ let the set of closest elements in E be $A(x) = \{x' \in E : \tau(x,x') = \min_{\tilde{x}\in E}\tau(x,\tilde{x})\}$. For any $h_E \in H_E$ define its extension $h : \mathcal{X} \to \mathcal{Y}$ by

$$h(x) = \begin{cases} h_E(x) & x \in E \\ \frac{1}{|A(x)|}\sum_{x'\in A(x)} h_E(x') & x \notin E \end{cases}$$

and let H be the set of all such extensions for $h_E \in H_E$. Let $f \in \mathcal{F}$, $x \in \mathcal{X}$, $x' \in A(x)$ and $h \in H$ then

$$\begin{aligned}\|f(x) - h(x)\|_\infty &= \|f(x) - f(x') + f(x') - h(x') + h(x') - h(x)\|_\infty\\ &\leq L\tau(x,x') + \|f(x') - h(x')\|_\infty + \|h(x') - h(x)\|_\infty.\end{aligned}$$

By assumption there's an $h_f \in H$ such that $\|f(x') - h_f(x')\|_\infty \leq \kappa$ because $x' \in E$. Any such h_f gives

$$\|f(x) - h_f(x)\|_\infty \leq L\delta + \kappa + \|h_f(x') - h_f(x)\|_\infty. \tag{6.4.1}$$

We will make use of the following fact about h_f. Let $x_1, x_2 \in A(x)$, then

$$\begin{aligned}\|h_f(x_1) - h_f(x_2)\|_\infty &= \|h_f(x_1) - f(x_1) + f(x_1) - f(x_2) + f(x_2) - h_f(x_2)\|_\infty \\ &\leq 2\kappa + L\tau(x_1, x_2) \\ &\leq 2\kappa + 2L\delta\end{aligned}$$

because $\tau(x_1, x_2) \leq \tau(x_1, x) + \tau(x_2, x) \leq 2\delta$. This then gives

$$\begin{aligned}\|h_f(x') - h_f(x)\|_\infty &= \left\| h_f(x') - \frac{1}{|A(x)|} \sum_{\tilde{x} \in A(x)} h_f(\tilde{x}) \right\|_\infty \\ &\leq \frac{1}{|A(x)|} \sum_{\tilde{x} \in A(x) \setminus \{x'\}} \|h_f(\tilde{x}) - h_f(x')\|_\infty \\ &\leq 2\kappa + 2L\delta.\end{aligned}$$

Putting this into Eq. (6.4.1) gives

$$\|f(x) - h_f(x)\|_\infty \leq 3\kappa + 3L\delta$$

which shows that H is a $(3\kappa + 3L\delta)$-cover for $\mathcal{F}$ in the function norm $\|\cdot\|_\infty$. So setting $\kappa = L\delta$ gives

$$\mathtt{Cov}\left(\mathcal{F}, \|\cdot\|_\infty, 6L\delta\right) \leq |H| = \prod_{x \in E} |F_x| \leq \sup_{x \in \mathcal{X}} \mathtt{Cov}\left(\mathcal{F}(x), \|\cdot\|_\infty, \delta/L\right)^{\mathtt{Cov}(\mathcal{X}, \tau, \delta)}.$$

Using this in Proposition 6.16 gives the statement. □

The above result says, taking $\alpha = 1/2$ for simplicity, that

$$n = \Omega\left(\frac{D_{\frac{\epsilon}{2}}(\mathcal{X}, \mathcal{F}) + \log(1/\delta)}{\epsilon^2}\right)$$

examples are sufficient to have generalisation error at most ϵ with probability at least $1-\delta$. The significance of this for equivariant models is as follows. Suppose that $\mathcal{G}$ acts on $\mathcal{X}$ and $\mathcal{Y}$, that the loss is preserved by $\mathcal{G}$, e.g., squared-error loss and an orthogonal representation on $\mathcal{Y}$, and that (X, Y) satisfy Assumption 1 and Assumption 2.[2] In this setting, we can consider the change to the bound in Theorem 6.15 that arises from taking $\mathcal{F}$ to be $\mathcal{G}$-equivariant. Specifically, Theorem 6.11 tells us that the sample complexity of learning on the task $\mathsf{T} = (X, Y, \ell)$ is the same as learning on the task $\mathsf{T}' = (X_\pi, Y_\pi, \ell)$. This means that

$$n = \Omega\left(\frac{D_{\frac{\epsilon}{2}}(\mathcal{X}_\pi, \mathcal{F}) + \log(1/\delta)}{\epsilon^2}\right)$$

examples are sufficient for an equivariant model to have generalisation error at most ϵ, potentially much less than above. The quantity $D_t(\mathcal{X}, \mathcal{F})$ is, in a sense, simultaneously measuring the size of $\mathcal{X}$ and the outputs of $\mathcal{F}$ at a scale t. The benefit of equivariance depends on the geometry of $\mathcal{X}$ and $\mathcal{X}_\pi$. Since $\mathcal{X}_\pi \subset \mathcal{X}$ we get $D_t(\mathcal{X}_\pi, \mathcal{F}) \leq D_t(\mathcal{X}, \mathcal{F})$ and, informally, the reduction will be large if $\mathcal{X}_\pi$ is much smaller than $\mathcal{X}$. If $\mathcal{F}(\mathcal{X}_\pi)$ is much smaller than $\mathcal{F}(\mathcal{X})$ then the same applies.

6.4.1 Concentration of Measure

The following is adapted from [118, Exercise 3.31].

Proposition 6.16. Let $\mathcal{X}$ be a set and let $\mathcal{Y} \subset \mathbb{R}^d$ be convex. Let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{X} \to \mathcal{Y}$. Let $\ell(y, y') : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be measurable loss function that's bounded $\ell(y, y') \leq M$ and Lipschitz in its first argument $|\ell(y_1, y') - \ell(y_2, y')| \leq C_\ell \|y_1 - y_2\|_\infty$ for some $C_\ell \in \mathbb{R}_+$ that's independent of y_1, y_2 and y'. Let $S = ((X_1, Y_1), \dots, (X_n, Y_n)) \sim (X, Y)^n$ be a training sample where X and Y are arbitrary random elements of $\mathcal{X}$ and $\mathcal{Y}$ respectively. Then,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \mathbb{E}[\ell(f(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i) \right| \geq \epsilon\right) \\
&\leq 2 \inf_{\alpha \in (0,1)} \mathtt{Cov}\left(\mathcal{F}, \|\cdot\|_\infty, \frac{\alpha\epsilon}{2C_\ell}\right) \mathsf{e}^{-\frac{1}{8}(1-\alpha)^2 n\epsilon^2 M^{-2}}.
\end{aligned}$$

Proof. For any probability measure ς on $\mathcal{X} \times \mathcal{Y}$ define $R_\varsigma[f] = \mathbb{E}[\ell(f(A), B)]$ where

[2] If $\mathcal{Y} = B_r(d)$ the closed Euclidean ball in $\mathbb{R}^d$ and $\ell(y, y') = \|y - y'\|_2^2$ is the squared-error loss, then this satisfies the statement of Theorem 6.15 with $M = r^2$ and $C_\ell = 4r\sqrt{d}$.

$(A, B) \sim \varsigma$. Then for all $f, f' \in \mathcal{F}$

$$\begin{aligned} R_\varsigma[f] - R_\varsigma[f'] &= \mathbb{E}[\ell(f(A), B) - \ell(f'(A), B)] \\ &\leq C_\ell \, \mathbb{E}[\|f(A) - f'(A)\|_\infty] \\ &\leq C_\ell \|f - f'\|_\infty \end{aligned}$$

Now let

$$L_S(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) - \mathbb{E}[\ell(f(X), Y)]$$

which has mean 0 for all f. By setting ς as the empirical measure on S and then as the distribution of (X, Y), one finds that

$$|L_S(f) - L_S(f')| \leq 2C_\ell \|f - f'\|_\infty. \tag{6.4.2}$$

Now let $\mathcal{K}$ be a κ-cover of $\mathcal{F}$ in $\|\cdot\|_\infty$. Define the sets $D(k) = \{f \in \mathcal{F} : \|f - k\|_\infty \leq \kappa\}$. Then

$$\begin{aligned} \mathbb{P}\left(\sup_{f \in \mathcal{F}} |L_S(f)| \geq \epsilon \right) &\leq \mathbb{P}\left(\bigcup_{k \in \mathcal{K}} \left\{ \sup_{f \in D(k)} |L_S(f)| \geq \epsilon \right\} \right) \\ &\leq \sum_{k \in \mathcal{K}} \mathbb{P}\left(\sup_{f \in D(k)} |L_S(f)| \geq \epsilon \right). \end{aligned}$$

Set $\kappa = \frac{\alpha\epsilon}{2C_\ell}$ for $0 < \alpha < 1$. Using Eq. (6.4.2), for all $f \in D(k)$ we have

$$L_S(f) \leq 2C_\ell \kappa + L_S(k) = \alpha\epsilon + L_S(k),$$

hence

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |L_S(f)| \geq \epsilon \right) \leq \sum_{k \in \mathcal{K}} \mathbb{P}(|L_S(k)| \geq (1 - \alpha)\epsilon).$$

Then Hoeffding's inequality gives

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |L_S(f)| \geq \epsilon \right) \leq 2|\mathcal{K}| \exp\left(-\frac{2(1 - \alpha)^2 n\epsilon^2}{16M^2} \right)$$

where we used the bound on ℓ in the statement. □

Chapter 7

Connections and Extensions

Summary

This chapter is more high-level than the others. In Section 7.1 we relate the orbit averaging viewpoint on invariance from Chapter 3 to the orbit representative viewpoint from Chapter 6. In Section 7.2 we apply some of the ideas in this thesis to neural networks. We outline some connections to other works in Section 7.3. Finally, in Section 7.4, we give some suggestions for future work.

7.1 Connections between Orbit Averaging and Orbit Representatives

Chapter 6 operates with a different perspective on invariant functions than developed in Chapter 3. The former expresses invariance of functions in terms of their dependence only on their values on a cross-section (one element from each orbit of $\mathcal{X}$ under the action of $\mathcal{G}$), while the latter uses orbit averaging. The purpose of this section is to give an intuitive connection between these viewpoints in the case of invariance. We leave equivariance to future work.

Consider $f \in L_2(\mu)$. By Proposition 3.6, f is invariant if and only if $\mathcal{O}f = f$. On the other hand, fixing a measurable cross-section $\mathcal{X}_\pi$ with projection π we can get an invariant function by $f_\pi = f \circ \pi$, and conversely Theorem 2.1 says that for all choices of $\mathcal{X}_\pi$ and all invariant f there's a function h such that $f = h_\pi$. However, although the map $f \mapsto f_\pi$ is a projection into invariant functions, it's not equivalent to $\mathcal{O}$. Decomposing $f = \mathcal{O}f + f^\perp$ gives $f_\pi = \mathcal{O}f + f_\pi^\perp$ so $f_\pi = \mathcal{O}f$ would require $f^\perp$ to vanish on $\mathcal{X}_\pi$. This is not the case for many choices of $\mathcal{X}_\pi$, for instance see the example in Fig. 3.1.

Instead, orbit averaging with $\mathcal{O}$ can be interpreted as averaging over all images of the cross-section under the action of $\mathcal{G}$, i.e., $\{g\mathcal{X}_\pi : g \in \mathcal{G}\}$. Define $g_\pi : \mathcal{X} \to \mathcal{G}$ such that for all $x \in \mathcal{X}$ $g_\pi(x)$ is any solution to $g_\pi(x)\pi(x) = x$, then, using the invariance of λ,

$$\begin{aligned}
\mathcal{O}f(x) &= \int_{\mathcal{G}} f(gx)\,\mathrm{d}\lambda(g) \\
&= \int_{\mathcal{G}} f(gg_\pi\pi(x))\,\mathrm{d}\lambda(g) \\
&= \int_{\mathcal{G}} f(g\pi(x))\,\mathrm{d}\lambda(g) \\
&= \int_{\mathcal{G}} f((g\circ\pi)(x))\,\mathrm{d}\lambda(g) \\
&= \int_{\mathcal{G}} f_{g\circ\pi}(x)\,\mathrm{d}\lambda(g)
\end{aligned}$$

where $g \circ \pi$ projects onto $g\mathcal{X}_\pi$.

A related view from [49, Section 5.1] is as follows. Let $\mathcal{X}$ be a locally compact, second countable and Hausdorff topological space. Let $\mathcal{K}$ be the set of all continuous $f : \mathcal{X} \to \mathbb{R}$ with compact support.[1] Then for all $f \in \mathcal{K}$ there's a unique function on the quotient space $\tilde{f} : \mathcal{X}/\mathcal{G} \to \mathbb{R}$ (also continuous with compact support) such that for all $x \in \mathcal{X}$

$$\mathcal{O}f(x) = \tilde{f}(p(x))$$

where $p : \mathcal{X} \to \mathcal{X}/\mathcal{G}$ is the map that sends each point to its orbit $p(x) = \{gx : g \in \mathcal{G}\}$.

[1] By compact support, we mean the set $\{x \in \mathcal{X} : f(x) \neq 0\}$ is compact.

The relation can be summarised in the following commutative diagram.

$$\begin{tikzcd}
\mathcal{X} \arrow[r, "p"] \arrow[dr, "\mathcal{O}f"'] & \mathcal{X}/\mathcal{G} \arrow[d, "\tilde{f}"] \\
& \mathbb{R}
\end{tikzcd}$$

This reinforces the interpretation of $\mathcal{O}$ as averaging over all possible cross-sections. Indeed, this viewpoint is somewhat more elegant, since it deals directly with $\mathcal{X}/\mathcal{G}$ and doesn't require the choice of a cross-section.

7.2 Applications to Neural Networks

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 $W^i \in \mathbb{R}^{\kappa_{i+1} \times \kappa_i}$ for $i = 1, \dots, L$ where $\kappa_1 = d$ and $\kappa_{L+1} = k$. We will assume F has the form

$$F(x) = W^L \sigma(W^{L-1} \sigma(\cdots \sigma(W^1 x) \cdots)) \tag{7.2.1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linearity applied element-wise. We refer to this architecture as the multi-layer perceptron (MLP).

7.2.1 Invariant and Equivariant Networks

The standard method for engineering MLPs to be invariant/equivariant to the action (specifically, a finite-dimensional representation) of a group on the inputs is weight tying. This method has been around for a while [191] but has come to recent attention via Ravanbakhsh, Schneider, and Póczos [138]. We outline this approach, give an estimate of the VC dimension of the resulting networks and apply the ideas of this work to propose a new method for learned invariance/equivariance.

The basic idea in [191, 138] can be summarised as follows. Let $\mathcal{G}$ be a compact group admitting the necessary representations. For each $i = 2, \dots, L+1$, the user chooses a matrix representation $\psi_i : \mathcal{G} \to \mathsf{GL}_{\kappa_i}$ of $\mathcal{G}$ that acts on the inputs for each layer $i = 2, \dots, L$ and on the outputs of the network when $i = L + 1$.[2] To avoid degeneracies we assume that the representations are finite, in the sense that for each i the set $\{\psi_i(g) : g \in \mathcal{G}\}$ forms a finite matrix group under multiplication. For $i = 2, \dots, L$, these representations must be chosen such that they commute with the activation function

$$\sigma(\psi_i(g)\cdot) = \psi_i(g)\sigma(\cdot) \tag{7.2.2}$$

for all $g \in \mathcal{G}$.[3] One then chooses weights for the network such that at each layer and for all $g \in \mathcal{G}$

$$W^i \psi_i(g) = \psi_{i+1}(g) W^i \tag{7.2.3}$$

(where there is no implicit sum of i). By induction on the layers, satisfying Eqs. (7.2.2) and (7.2.3) ensures that the network is equivariant. The network is invariant if ψ_{L+1} is the trivial representation.

The condition in Eq. (7.2.3) 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 possible weight matrices in layer i as $U = \mathbb{R}^{\kappa_{i+1} \times \kappa_i}$, the space of intertwiners is immediately recognisable as $\Psi^i(U)$ where Ψ^i is the linear map with components

$$\Psi^i_{abce} = \int_{\mathcal{G}} \psi_{i+1}(g^{-1})_{ac} \psi_i(g)_{eb} \, \mathrm{d}\lambda(g) \tag{7.2.4}$$

which acts by $\Psi^i(W)_{ab} = \Psi^i_{abce} W_{ce}$. This first appeared in Eq. (4.2.1), although defined slightly differently, where Ψ was introduced as the restriction of the operator $\mathcal{Q}$ to linear maps.

We emphasise that the definition of Ψ^i does not require the representations to be orthogonal. Strictly, according to our definition of $\mathcal{Q}$, this means that Ψ^i defined in Eq. (7.2.4) is not the restriction of $\mathcal{Q}$ to linear maps $\mathbb{R}^{\kappa_i} \to \mathbb{R}^{\kappa_{i+1}}$. However, it is still

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

[3] This condition is somewhat restrictive, but note that a permutation representation will commute with any element-wise non-linearity. For a description of admissible combinations of representations and activation functions see [191, Theorem 2.4].

a linear projection onto equivariant elements and we can still write $W = \overline{W} + W^{\perp}$ where $\overline{W} = \Psi(W)$ and $\Psi(W^{\perp}) = 0$, we just lose the orthogonality in Lemma 3.1. It remains well defined because the representations are finite.

7.2.1.1 VC Dimension

The following result is a corollary of [15, Theorem 6] which gives an estimate of the VC dimension of invariant MLPs with ReLU activation of the kind discussed above. Similar results are possible for other activation functions. It applies to the case where the network is layer-wise equivariant and ψ_{L+1} is the trivial representation.

Proposition 7.1. Consider a $\mathcal{G}$-invariant MLP architecture with ReLU activations and weights constrained to intertwine the representations as described above. Let $\mathcal{F}$ be the set of binary classifiers of the form $\operatorname{sign}(F)$ for all functions F computed by this architecture. Then

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

where $\alpha = \log_2\left(4\mathsf{e}\log_2\left(\sum_{i=1}^{L} 2\mathsf{e}i\kappa_i\right)\sum_{i=1}^{L} i\kappa_i\right)$, $\chi_i(g) = \mathrm{Tr}(\psi_i(g))$ are the characters of the representations and $(\chi_i|\chi_{i+1}) = \int_{\mathcal{G}} \chi_i(g)\chi_{i+1}(g)\,\mathrm{d}\lambda(g)$ is their inner product.

Proof. For a ReLU network with t_i independent parameters at each layer we have

$$\mathrm{VC}(\mathcal{F}) \leq L + \alpha \sum_{i=1}^{L} (L - i + 1)t_i,$$

which is by direct application of [15, Theorem 6]. The condition Eq. (7.2.3) is the statement that the weight matrix W^i belongs to the space of equivariant maps $\mathbb{R}^{\kappa_i} \to \mathbb{R}^{\kappa_{i+1}}$. The number of independent parameters at each layer is at most the dimension of this space, which is $(\chi_i|\chi_{i+1})$. The conclusion follows easily. □

Example 7.2 (Permutation invariant networks)**.** Permutation invariant networks are studied in many other works, see [191, 195, 22] and references therein. In particular, multiple authors have given the form of a permutation equivariant weight matrix as

$$W = \alpha I + \beta \mathbf{1}\mathbf{1}^\top$$

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 $\kappa_i = d$ for all i. Let $\mathcal{F}$ be the class of all functions realisable by this network, then

$$\mathrm{VC}(\mathrm{sign}(\mathcal{F})) = O(L^2 \log(Ld \log(Ld))).$$

7.2.2 Regularisation for Equivariance

Typically, the practitioner will parameterise the weight matrices so that they satisfy Eq. (7.2.3). We use the ideas of this work to suggest an alternative method that allows for learned symmetry using regularisation. We leave an empirical exploration of this approach to future work.

We suggest the obvious thing: regularise $W^{i\perp}$ for each layer $i = 1, \dots, L$, where $W^{i\perp} := W^i - \Psi^i(W^i)$. For instance, one could add a weight decay term to the training objective of the form $\sum_{i=1}^L \|W^{i\perp}\|_{\mathrm{F}}^2$. Note that the operator Ψ^i can be computed before training.

If $W^{i\perp} = 0$ for $i = 1, \dots, L$ and the activation function satisfies Eq. (7.2.2), then the resulting network will be exactly invariant or equivariant (depending on the choice of last layer representation). This method could also allow for approximate symmetry. Indeed, the following result suggests $\|W^{i\perp}\|_{\mathrm{F}}^2$ as a measure of the layer-wise equivariance of the network.

Proposition 7.3. Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}^k$, each with the Euclidean inner product. 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 function σ. Let $\mathcal{G}$ be a compact group with orthogonal representations ϕ and ψ on $\mathcal{X}$ and $\mathcal{Y}$ respectively, and assume that ψ

commutes with σ as in Eq. (7.2.2). Assume $f_W \in L_2(\mathcal{X}, \mathcal{Y}, \mu)$ where μ is a $\mathcal{G}$-invariant probability measure and write $\Sigma = \int_{\mathcal{X}} xx^\top \, \mathrm{d}\mu(x)$ for its covariance matrix which we assume to be finite. Then the distance from f_W to its closest equivariant function is bounded by

$$\|(\mathrm{id} - \mathcal{Q}) f_W\|_\mu^2 \leq 2C^2 \|W^\perp \Sigma^{\frac{1}{2}}\|_\mathrm{F}^2 \leq 2C^2 \|\Sigma^{\frac{1}{2}}\|_\mathrm{F}^2 \|W^\perp\|_\mathrm{F}^2$$

Proof. We do the left hand side inequality first. In Eq. (†) below we apply Jensen's inequality Lemma A.3.1 which requires that the integrand is integrable, we verify this later. We calculate

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

where the cross terms vanish in the final line because we must have

$$\int_{\mathcal{G}} \psi(g^{-1}) W^\perp \phi(g) \, \mathrm{d}\lambda(g) = 0$$

(see Section 4.2). Now let $X \sim \mu$. By inserting traces, applying Fubini's theorem and using the $\mathcal{G}$-invariance of μ we get

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

To complete the proof we address our application of Jensen's inequality in Eq. (†), which requires that for all $x \in \mathcal{X}$

$$\psi(g^{-1})f_W(\phi(g)x) \ - f_W(x) \in L_1(\mathcal{G}, \mathcal{Y}, \lambda). \tag{7.2.5}$$

We can ignore the second term in the above because λ is finite. By Theorem 1.1, $L_2(\mathcal{G}, \lambda) \subset L_1(\mathcal{G}, \lambda)$. Applying this to each coordinate, $L_2(\mathcal{G}, \mathcal{Y}, \lambda) \subset L_1(\mathcal{G}, \mathcal{Y}, \lambda)$. The function $m : (g, x) \mapsto \psi(g^{-1})f_W(\phi(g)x)$ is $(\lambda \otimes \mu)$-measurable by the beginning of the proof of Proposition 3.4, so $g \mapsto \psi(g^{-1})f_W(\phi(g)x)$ is λ-measurable for all $x \in \mathcal{X}$ by using Lemma A.1.1 to apply to Corollary A.1.3 each component. It is sufficient, therefore, to verify that

$$\int_{\mathcal{G}} \|\psi(g^{-1})f_W(\phi(g)x)\|_2^2 \,\mathrm{d}\lambda(g) < \infty.$$

Using the developments after Eq. (†) and using $\|\phi(g)\|_2 = 1$ by orthogonality, it only remains to calculate

$$\begin{aligned}
\int_{\mathcal{G}} x^\top \phi(g^{-1})(W^\perp)^\top W^\perp \phi(g)x \,\mathrm{d}\lambda(g) &= \int_{\mathcal{G}} \|W^\perp \phi(g)x\|_2^2 \,\mathrm{d}\lambda(g) \\
&\leq \int_{\mathcal{G}} \|W^\perp\|_2^2 \|\phi(g)\|_2^2 \|x\|_2^2 \,\mathrm{d}\lambda(g) \\
&= \|W^\perp\|_2^2 \|x\|_2^2 \\
&< \infty.
\end{aligned}$$

The right hand side inequality comes from, for all square matrices A, B,

$$\|AB\|_{\mathrm{F}}^2 = \operatorname{Tr}(B^\top A^\top AB) = \operatorname{Tr}(A^\top ABB^\top) \leq \operatorname{Tr}(A^\top A)\operatorname{Tr}(BB^\top) = \|A\|_{\mathrm{F}}^2\|B\|_{\mathrm{F}}^2$$

by Cauchy-Schwarz. □

Proposition 7.3 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 $\mathcal{G}$-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.

7.3 Connections to Other Works

We discuss some connections to other works. Apart from [116], these were found while researching the literature review. We became aware of [116] following work on [51]. We organise the comparisons by the relevant parts of this work and adapt the results from other authors to the notation of this work.

7.3.1 $\mathcal{Q}$ on Linear Functions

Wood and Shawe-Taylor [191] show that a linear map f is equivariant if and only if $\mathcal{Q}f = f$ [191, Lemma 2.5]. They also define the *characteristic matrix* [191, Definition 3.2], which we denote by Φ in Section 4.1, show that it is the orthogonal projection onto the invariant elements [191, Theorem 3.4] and also show that its trace is the dimension of the invariant subspace [191, Corollary 3.5]. Similarly, Pal, Kannan, Arakalgud, and Savvides [124] show that when $\mathcal{G}$ acts by a unitary (i.e., orthogonal) representation on $\mathbb{R}^d$ then $\mathcal{Q}$ is a self-adjoint projection on $\mathbb{R}^d$ [124, Lemma 2.3].

7.3.2 Generalisation in Kernel Ridge Regression

Mei, Misiakiewicz, and Montanari [116] analyse the generalisation of invariant random feature and kernel regression in a high dimensional (large d) setting. They study the risk in the case where the kernel is invariant by design, whereas we use averaging to isolate the benefit of invariance. In this way their results are more specific to what would be done in practice. However, as is stated in Theorem 5.2, if training with an invariant kernel gives lower risk than averaging, then Theorem 5.2 applies to the former too.

The results in [116] concern a specific scaling between the number of training examples n and the input dimension d, and are restricted to specific input distributions and inner product kernels. On the other hand, Theorem 5.2 places mild assumptions on the kernel and holds for any invariant input distribution.

It's worth mentioning that Theorem 5.2 and [116] are consistent: the representations and kernels in [116] satisfy Eq. (5.3.1) so from the discussion in Section 5.2.3.1 the lower estimate on the risk of kernel ridge regression in Theorem 5.2 vanishes in probability as $d \to \infty$, which matches the comments after [116, Proposition 1] because in their setting $n \geq d^c$ for some constant $c > 0$.

7.3.3 Invariant Kernels and the Decomposition of the RKHS

A *unitary kernel* is one for which $k(gx, gy) = k(x, y)$ for all $x, y \in \mathcal{X}$ and all $g \in \mathcal{G}$. This is similar to Eq. (5.3.1) and has appeared many times in the literature, for instance [124, Definition 2.2], [125, Definition 3.1] and [139]. It turns out that, under this assumption, Risi Kondor proved a weaker version of Theorem 5.14 in his PhD Thesis.

Theorem 7.4 ([93, Theorem 4.4.3])**.** *Let $\mathcal{G}$ be a finite group acting on $\mathcal{X}$ and let $\mathcal{H}$ be a reproducing kernel Hilbert space with unitary kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Assume further that $f \circ g \in \mathcal{H}$ for all $f \in \mathcal{H}$ and all $g \in \mathcal{G}$. Then the invariant functions in $\mathcal{H}$ form a subspace which is a reproducing kernel Hilbert space with kernel*

$$\overline{k}(x, y) = \int_{\mathcal{G}} k(x, gy) \, \mathrm{d}\lambda(g).$$

Separately, Reisert and Burkhardt [139] argue that for unitary kernels, $\mathcal{Q}$ is an orthogonal projection onto the equivariant elements in the RKHS $\mathcal{H}$. However, it is not shown that $\mathcal{Q}\mathcal{H} \subset \mathcal{H}$. Based on this, they present an equivariant version of the representer theorem, which, specialising to invariance and the setting of Chapter 5, says that any invariant solution to Eq. (5.2.1) is in $\overline{\mathcal{H}}$ (defined in Theorem 5.14).

7.4 Ideas for Future Work

Generalisation Gap for Non-Invariant μ

An interesting possibility for future work is to try to remove the assumption that μ is invariant from Lemma 3.12. Let ν be a probability measure on $(\mathcal{X}, \mathcal{S}_{\mathcal{X}})$, let $X \sim \nu$ and $Y = f^\star(X) + \xi$ with $f^\star$ equivariant, $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\xi^2] < \infty$. Let $f \in L_2(\mathcal{X}, \mathcal{Y}, \nu)$ and assume, for the sake of argument, that $\mathcal{Q}f \in L_2(\mathcal{X}, \mathcal{Y}, \nu)$. Then we can write $f = \bar{f} + f^\perp$ where $\bar{f} = \mathcal{Q}f$ and $f^\perp = (\mathrm{id} - \mathcal{Q})f$, but the terms may not be orthogonal. The generalisation gap between f and $\bar{f}$ becomes

$$R[f] - R[\bar{f}] = \|f - f^\star\|_\nu^2 - \|\bar{f}\|_\nu^2 = \|f^\perp\|_\nu^2 + 2\langle \bar{f} - f^\star, f^\perp \rangle_\nu.$$

If ν is invariant, then we know from Lemma 3.1 that $\langle \bar{f} - f^\star, f^\perp \rangle_\nu = 0$. Otherwise, this term would need to be estimated. One approach might be to consider something like

$$\langle \bar{f} - f^\star, f^\perp \rangle_\nu \leq \|\bar{f} - f^\star\|_\nu \|f^\perp\|_\nu \sup_{f,h \in L_2(\mathcal{X},\mathcal{Y},\nu)} \frac{\langle \mathcal{Q}f, (\mathrm{id} - \mathcal{Q})h \rangle_\nu}{\|\mathcal{Q}f\|_\nu \|(\mathrm{id} - \mathcal{Q})h\|_\nu},$$

The supremum term is in the interval $[-1, 1]$ and vanishes if ν is invariant. It may be related to a distance between ν and the invariant measure $\bar{\nu}$ defined by $\bar{\nu}(A) = \int_G \nu(gA) \, \mathrm{d}\lambda(g)$ for $A \in \mathcal{S}_{\mathcal{X}}$.

Application to G-CNNs

Recent work has generalised the standard convolutional layer to obtain new group equivariant neural network layers, for instance [34, 37, 35]. As mentioned in Section 2.2.2, Kondor and Trivedi [95] show that any equivariant neural network layer can be written in terms of a generalisation of the standard convolution. Let $f : \mathcal{X} \to \mathcal{Y}$ and let $\mathcal{X}_\pi$ be a measurable cross-section of $\mathcal{X}$ with respect to $\mathcal{G}$. Then, fixing some $x_\pi \in \mathcal{X}_\pi$, f can be *lifted* to a function on $\mathcal{G}$, $f_{x_\pi} : \mathcal{G} \to \mathcal{Y}$ with values $f_{x_\pi}(g) = f(gx_\pi)$. The analysis in [95] is specialised to *homogeneous spaces*, those such that for all $x, y \in \mathcal{X}$ $\exists g \in \mathcal{G}$ such that $gx = y$. We can apply the result to each orbit independently by lifting. Let $\mathcal{Y} = \mathbb{R}^k$, fix $x_\pi \in \mathcal{X}_\pi$ and suppose that f is the output of an equivariant intermediate layer of a G-CNN that we lift to f_{x_π}, so both f and f_{x_π} are equivariant. Kondor and Trivedi [95] show that the (lifted) linear map in any equivariant convolutional layer must take the form of a *group convolution* of the previous layer with respect to some *filter* $\vartheta : \mathcal{G} \to \mathbb{R}^{m \times k}$, where m is the input dimension of the following layer,

$$(\vartheta * f_{x_\pi})(u) = \int_{\mathcal{G}} \vartheta(v) f_{x_\pi}(uv^{-1}) \, \mathrm{d}\lambda(v).$$

Applying the equivariance of f_{x_π}, the invariance λ and the definition of f_{x_π} gives

$$(\vartheta * f_{x_\pi})(u) = \int_{\mathcal{G}} \vartheta(vu)v^{-1} \, \mathrm{d}\lambda(v) f(x_\pi) = \bar{\vartheta}(u) f(x_\pi)$$

where

$$\bar{\vartheta}(u) = \int_{\mathcal{G}} \vartheta(vu)v^{-1} \, \mathrm{d}\lambda(v)$$

which looks, at least formally, rather like $\mathcal{Q}$. In fact, the map $\vartheta \to \bar{\vartheta}$ is even a projection. Could this allow for an analysis of G-CNN type architectures using the techniques developed in Chapter 3?

Other Loss Functions

At the core of most of our results on generalisation is Lemma 3.12, which quantifies the generalisation gap in regression problems in terms of the $\mathcal{G}$-anti-symmetric component of the predictor. An interesting line of future work could be to extend this to similar results for other loss functions. As a starter, one might consider the following style of argument for classification; although Corollary 7.6 is somewhat weaker than Lemma 3.12 in that it does not provide a strict benefit for invariance.

Theorem 7.5 ([96, Theorem 5])**.** Let $f : \mathcal{X} \to [0, 1]$ and let (X, Y) be a random element of $\mathcal{X} \times \{0, 1\}$. Let $L[f] = \mathbb{P}(\text{sign}(f(X) - 1/2) \neq Y)$ be the 0/1 risk and let $R[f] = \mathbb{E}[(f(X) - Y)^2]$ be the squared-error risk. Define $f^\star = \text{argmin}_{f:\mathcal{X}\to[0,1]}\ R[f]$, then

$$L[f] - L[f^\star] \leq 2\sqrt{R[f] - R[f^\star]}.$$

Corollary 7.6. Let $X \sim \mu$. It's easy to show that $R[f] - R[f^\star] = \|f - f^\star\|_\mu^2$, so Theorem 7.5 gives $L[f] - L[f^\star] \leq 2\|f - f^\star\|_\mu$. If $f^\star$ is invariant, using Lemma 3.1 and writing $f = \bar{f} + f^\perp$ gives

$$L[\bar{f} + f^\perp] - L[f^\star] \leq 2\|\bar{f} - f^\star\|_\mu + \|f^\perp\|_\mu.$$

Hence, our framework suggests an improvement in generalisation in classification if the invariance is correctly specified (i.e., $f^\perp = 0$).

Training First vs. Averaging First

Most of this thesis is spent studying the generalisation gap between an arbitrary predictor and its equivariant projection. This corresponds to comparing a trained model with its averaged version. It might seem more sensible from a practical perspective, for instance in the case of regression, to first average the features and then learn the model on the invariant/equivariant features. We have stressed throughout that our generalisation results still apply if this method of averaging then training produces a predictor with lower risk than training and then averaging. A possible, albeit highly involved avenue for future work is to understand the difference in generalisation of these two methods. In the case of kernel ridge regression, some insight can be gained by comparing our results to those in [116].

Invariance in other Kernel Methods

Lemma 3.12 holds for all predictors in $L_2(\mu)$ so in principle could be applied to calculate the generalisation gap for any algorithm with outputs in $L_2(\mu)$. However, in the case of an RKHS with kernel satisfying Eq. (5.3.1) there is also the RKHS inner product decomposition in Theorem 5.14. Could this be used to study invariance in other kernels methods?

Equivariance in Kernel Regression

A natural extension of Theorem 5.2 is to equivariance. It's possible that the equivariant matrix-valued kernels studied by Reisert and Burkhardt [139] are an equivariant generalisation of $\overline{k}$ from Chapter 5, or possibly of kernels satisfying Eq. (5.3.1). The review by Alvarez, Rosasco, Lawrence, et al. [5] may also be relevant.

Appendix A

Useful Results

We provide some useful results that are relied upon elsewhere in the work. Any proofs given are for fun/completeness and we claim no originality.

A.1 Measure Theory

Lemma A.1.1 (tuples of functions [85, Lemma 1.8])**.** Let T be a finite set and let $(\Omega, \mathcal{A})$ and $(S_t, \mathcal{S}_t)$ for $t \in T$ be measurable spaces. Define $S = \bigtimes_{t \in T} S_t$ and $\mathcal{S} = \bigotimes_{t \in T} \mathcal{S}_t$. Consider any functions $f_t : \Omega \to S_t$ and define $f = (f_t : t \in T) : \Omega \to S$. Then f is $\mathcal{A}/\mathcal{S}$-measurable if and only if f_t is $\mathcal{A}/\mathcal{S}_t$-measurable for all $t \in T$.

Lemma A.1.2 (sections [85, Lemma 1.26])**.** Let $(S, \mathcal{S}, \mu)$ be a σ-finite measure space and let $(T, \mathcal{T}, \nu)$ be a measure space. Let $f : S \times T \to \mathbb{R}_+$ be measurable. Then

(i) $f_t(s) = f(s,t)$ is μ-measurable for each $t \in T$;

(ii) $\bar{f}(t) = \int_S f(s,t) \, \mathrm{d}\mu(s)$ is ν-measurable.

The following corollaries follow by applying Lemma A.1.2 to the decomposition $f = f_+ - f_-$ where $f_+ = \max(f, 0)$ and $f_- = \max(-f, 0)$. Each are integrable when f is integrable and the integral of f is defined to be the difference between the integrals of f_+ and f_- [85, p. 11].

Corollary A.1.3 (sections of real functions)**.** Let $(S, \mathcal{S}, \mu)$ and $(T, \mathcal{T}, \nu)$ be probability spaces. Let $f : S \times T \to \mathbb{R}$ be measurable, then $f_t(s) = f(s, t)$ is μ-measurable for each $t \in T$.

Corollary A.1.4 (sections of integrable functions)**.** Let $(S, \mathcal{S}, \mu)$ and $(T, \mathcal{T}, \nu)$ be probability spaces and let $f : S \times T \to \mathbb{R}$ be $(\mu \otimes \nu)$-integrable, then

$$\bar{f}(t) = \int_S f(s, t) \, \mathrm{d}\mu(s)$$

is ν-measurable.

A.2 Linear Transformations

Lemma A.2.1. Let $A, B \in \mathbb{R}^{n \times n}$ with A symmetric and B positive semi-definite, then

$$\gamma_{\max}(A) \operatorname{Tr}(B) \geq \operatorname{Tr}(AB) \geq \gamma_{\min}(A) \operatorname{Tr}(B)$$

where $\gamma_{\min}$ and $\gamma_{\max}$ denote the minimum and maximum eigenvalue respectively.

Proof. Write $A = \sum_{i=1}^n \gamma_i(A) v_i v_i^\top$ where $\gamma_1(A), \ldots, \gamma_n(A)$ are the eigenvalues of A with multiplicity and $v_1, \ldots, v_n \in \mathbb{R}^n$ are the respective eigenvectors that form an orthonormal basis. Then

$$\begin{aligned}
\operatorname{Tr}(AB) &= \operatorname{Tr}\left(\sum_{i=1}^n \gamma_i(A) v_i v_i^\top B\right) \\
&= \sum_{i=1}^n \gamma_i(A) v_i^\top B v_i \\
&\geq \gamma_{\min}(A) \sum_{i=1}^n v_i^\top B v_i \\
&= \gamma_{\min}(A) \operatorname{Tr}(B)
\end{aligned}$$

and the left hand side inequality follows by the same approach. □

Lemma A.2.2. Let $A \in \mathbb{R}^{n \times n}$, then

$$\|A\|_2 \leq n \max_{ij} |A_{ij}|.$$

Proof. Let $a_i \in \mathbb{R}^n$ be the i^{th} column of A, then

$$\begin{aligned}
\sup_{\|x\|_2=1} \|Ax\|_2 &= \sup_{\|x\|_2=1} \sqrt{\sum_i (a_i^\top x)^2} \\
&\leq \sup_{\|x\|_2=1} \sqrt{\sum_i \|a_i\|_2^2 \|x\|_2^2} \\
&\leq \sqrt{\sum_i \|a_i\|_2^2} \\
&\leq \sqrt{n^2 \max_{ij} A_{ij}^2}.
\end{aligned}$$

□

Lemma A.2.3. 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 [126], namely: i. $AXA = A$, ii. $XAX = X$, iii. $(AX)^\top = AX$ and iv. $(XA)^\top = XA$. It is straightforward to check that this is the case. □

Theorem A.2.4 (Adjoints [142, Theorem 4.10])**.** Let H_1 and H_2 be (possibly infinite dimensional) Hilbert spaces. For every bounded linear operator $T : H_1 \to H_2$ there exists a unique bounded linear operator $T^* : H_2 \to H_1$ such that for all $x \in H_1$ and all $y \in H_2$

$$\langle Tx, y \rangle_{H_2} = \langle x, T^* y \rangle_{H_1}$$

and $\|T^*\|_{\text{op}} = \|T\|_{\text{op}}$.

Theorem A.2.5 ([142, Theorem 4.13])**.** Let H be a Hilbert space and let $T : H \to H$ be a bounded linear operator. Then T is surjective if and only if $\exists \delta > 0$ such that $\forall x \in H$ $\|T^* x\|_H \geq \delta \|x\|_H$.

Lemma A.2.6. Let H_1 and H_2 be Hilbert spaces, let $S : H_1 \to H_2$ be a bounded linear operator with adjoint S^* and set $\Sigma = S^* S$. Then for any $\rho > 0$, $\Sigma + \rho\,\mathrm{id} : H_1 \to H_1$ is invertible.

Proof. For any $v \in H$ we have

$$\langle \Sigma v, v \rangle_{H_1} = \langle S^* S v, v \rangle_{H_1} = \|Sv\|_{H_2}^2 \geq 0$$

so

$$\|(\Sigma + \rho\,\mathrm{id})v\|_{H_1}^2 = \|\Sigma v\|_{H_1} + 2\rho \langle \Sigma v, v \rangle_{H_1} + \rho^2 \|v\|_{H_1} = \|\Sigma v\|_{H_1}^2 + 2\rho \|Sv\|_{H_2}^2 + \rho^2 \|v\|_{H_1}$$

which verifies that the kernel of $\Sigma + \rho\,\mathrm{id}$ is trivial so it is injective. The calculation also shows that $\Sigma + \rho\,\mathrm{id}$ is bounded below, i.e., $\|(\Sigma + \rho\,\mathrm{id})v\|_{H_1} \geq \rho \|v\|_{H_1}$. Clearly $\Sigma + \rho\,\mathrm{id}$ is both bounded and self-adjoint, so it is surjective by Theorem A.2.5. □

Lemma A.2.7. Let H_1 and H_2 be Hilbert spaces, let $S : H_1 \to H_2$ be a bounded linear operator with adjoint S^* and set $\Sigma = S^* S$ and $T = SS^*$. Let $\rho > 0$, then the problem

$$\operatorname*{argmin}_{f \in H_1} \|Sf - f'\|_{H_2}^2 + \rho \|f\|_{H_1}^2$$

is solved uniquely by

$$f = (\Sigma + \rho\,\mathrm{id})^{-1} S^* f'$$

and $Sf = (T + \rho\,\mathrm{id})^{-1} T f'$.

Proof. We want to minimise $L_\rho[f] = \|Sf - f'\|_\mu^2 + \rho \|f\|_{H_1}^2$ over $f \in H_1$. For all non-zero $h \in H_1$

$$\begin{aligned} L_\rho[f+h] &= \|Sf + h - f'\|_{H_2}^2 + \rho\|f+h\|_{H_1}^2 \\ &= L_\rho[f] + 2\langle Sf - f', Sh\rangle_{H_2} + 2\rho\langle f, h\rangle_{H_1} + \|Sh\|_{H_2}^2 + \rho\|h\|_{H_1}^2 \\ &> L_\rho[f] + 2\langle Sf - f', Sh\rangle_{H_2} + 2\rho\langle f, h\rangle_{H_1}. \end{aligned}$$

Any f that solves $\langle Sf - f', Sh\rangle_{H_2} + \rho\langle f, h\rangle_{H_1} = 0$ simultaneously for all $h \in H_1$ is then the unique global minimum of L_ρ. We calculate

$$\begin{aligned} \langle Sf - f', Sh\rangle_{H_2} + \rho\langle f, h\rangle_{H_1} &= \langle S^*(Sf - f'), h\rangle_{H_1} + \rho\langle f, h\rangle_{H_1} \\ &= \langle (\Sigma + \rho\,\mathrm{id})f - S^*f', h\rangle_{H_1} \\ &= \langle (\Sigma + \rho\,\mathrm{id})f - S^*f', h\rangle_{H_1}. \end{aligned}$$

This completes the first part of the proof, with Lemma A.2.6 verifying that $\Sigma + \rho\,\mathrm{id}$ is invertible. Finally, let $h \in H_2$ and set $f = (\Sigma + \rho\,\mathrm{id})^{-1}S^*h$ then

$$S(S^*S + \rho\,\mathrm{id})f = SS^*h$$

which implies $Sf = (T + \rho\,\mathrm{id})^{-1}Th$. Again, Lemma A.2.6 validates the inverse. □

A.3 Probability and Statistics

Lemma A.3.1 (Jensen's inequality [85, Lemma 3.5])**.** Let ξ be an integrable element of $\mathbb{R}^d$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, then

$$\mathbb{E}[f(\xi)] \leq f(\mathbb{E}[\xi]).$$

A.3.1 Inverse Wishart Matrices

Lemma A.3.2 ([69])**.** 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 A.3.3. It is well known that the expectation in Lemma A.3.2 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 $R \in \mathsf{O}_d$ by Lemma A.2.3. 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 $\lambda_1, \ldots, \lambda_d$ 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 [121, 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^{-1}, there is a term from the expansion of 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 A.3.4 ([38, 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 A.3.5. The statement of Lemma A.3.4 in [38, Theorem 2.1] gives the condition $n < d-3$, but this is not necessary for the first moment, which can be seen from the proof. In addition, the proof uses a transformation followed by an application of Lemma A.3.2 with the roles of n and d switched. Using this transformation, it follows (e.g., from the earlier remark) that the expectation diverges when $d \geq n \geq d-1$.

A.3.2 Isotropic Random Projections

Lemma A.3.6 (Will Sawin)**.** Let $E \sim \mathrm{Unif}\, \mathbb{G}_n(\mathbb{R}^d)$ 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 (e.g., by [79])

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

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.4 Reproducing Kernel Hilbert Spaces

RKHS stands for reproducing kernel Hilbert space. See Section 5.1.1 for a brief introduction.

Lemma A.4.1 (RHKS of measurable kernels [166, Lemma 4.24])**.** Let $\mathcal{H}$ be an RKHS of functions $f : \mathcal{X} \to \mathbb{R}$ with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then all $f \in \mathcal{H}$ are measurable if and only if $k(\cdot, x) : \mathcal{X} \to \mathbb{R}$ is measurable for all $x \in \mathcal{X}$.

Lemma A.4.2 (RHKS of continuous functions [166, Lemma 4.28])**.** Let $\mathcal{H}$ be an RKHS of functions $f : \mathcal{X} \to \mathbb{R}$ with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then all $f \in \mathcal{H}$ are bounded and continuous if and only if $k(\cdot, x) : \mathcal{X} \to \mathbb{R}$ is bounded and continuous for all $x \in \mathcal{X}$.

Lemma A.4.3 (separable RKHS [166, Lemma 4.33])**.** Let $(\mathcal{X}, \tau)$ be a separable topological space and let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous kernel, then the RKHS of k is separable.

Theorem A.4.4 (integral operators of kernels [166, Theorem 4.26])**.** Let $(\mathcal{X}, \mathcal{S}_{\mathcal{X}}, \mu)$ be a σ-finite measure space and let $\mathcal{H}$ be a separable RKHS of functions $f : \mathcal{X} \to \mathbb{R}$ with measurable kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $p \in [1, \infty)$ and define the function $m_k : \mathcal{X} \to \mathbb{R}$ by $m_k(x) = \sqrt{k(x, x)}$. If

$$\|m_k\|_{L_p(\mu)} := \left(\int_{\mathcal{X}} m_k(x)^p \, \mathrm{d}\mu(x) \right)^{\frac{1}{p}} < \infty$$

then all $f \in \mathcal{H}$ satisfy $\|f\|_{L_p(\mu)} < \infty$ and the inclusion map $\iota : \mathcal{H} \to L_p(\mu)$ is continuous with operator norm $\|\iota\|_{\mathrm{op}} \leq \|m_k\|_{L_p(\mu)}$. Moreover, the adjoint of the inclusion operator is $S_k : L_q(\mu) \to \mathcal{H}$ where

$$S_k f(x) = \int_{\mathcal{X}} k(x, x') f(x') \, \mathrm{d}\mu(x')$$

and $1/p + 1/q = 1$. In addition:

i) $\iota\mathcal{H}$ is dense in $L_p(\mu)$ if and only if S_k is injective

ii) $S_k L_q(\mu)$ is dense in $\mathcal{H}$ if and only if ι is injective.

Appendix B

Excluded Works

Lottery Tickets in Linear Models: An Analysis of Iterative Magnitude Pruning

Bryn Elesedy, Varun Kanade, Yee Whye Teh

Sparsity in Neural Networks Workshop, 2021

https://arxiv.org/abs/2007.08243

Abstract

We analyse the pruning procedure behind the lottery ticket hypothesis [59], *iterative magnitude pruning* (IMP), when applied to linear models trained by gradient flow. We begin by presenting sufficient conditions on the statistical structure of the features under which IMP prunes those features that have smallest projection onto the data. Following this, we explore IMP as a method for sparse estimation.

Effectiveness and Resource Requirements of Test, Trace and Isolate Strategies

Bobby He*, Sheheryar Zaidi*, **Bryn Elesedy***, Michael Hutchinson*, Andrei Paleyes, Guy Harling, Anne Johnson and Yee Whye Teh, on behalf of Royal Society DELVE group (*equal contribution)

Royal Society Open Science, 2021

https://royalsocietypublishing.org/doi/10.1098/rsos.201491

Abstract

We use an individual-level transmission and contact simulation model to explore the effectiveness and resource requirements of various test-trace-isolate (TTI) strategies for reducing the spread of SARS-CoV-2 in the UK, in the context of different scenarios with varying levels of stringency of non-pharmaceutical interventions. Based on modelling results, we show that self-isolation of symptomatic individuals and quarantine of their household contacts has a substantial impact on the number of new infections generated by each primary case. We further show that adding contact tracing of non-household contacts of confirmed cases to this broader package of interventions reduces the number of new infections otherwise generated by 5–15%. We also explore impact of key factors, such as tracing application adoption and testing delay, on overall effectiveness of TTI.

Efficient Bayesian Inference of Instantaneous Reproduction Numbers at Fine Spatial Scales, with an Application to Mapping and Nowcasting the COVID-19 Epidemic in British Local Authorities.

Yee Whye Teh, Avishkar Bhoopchand*, Peter Diggle*, **Bryn Elesedy***, Bobby He*, Michael Hutchinson*, Ulrich Paquet*, Jonathan Read*, Nenad Tomasev*, Sheheryar Zaidi* (*alphabetical ordering).

Royal Society Special Topic Meeting on R, Local R and Transmission of COVID-19.

https://rss.org.uk/RSS/media/File-library/News/2021/WhyeBhoopchand.pdf

Abstract

The spatio-temporal pattern of Covid-19 infections, as for most infections disease epidemics, is highly heterogeneous as a consequence of local variations in risk factors and exposures. Consequently, the widely quoted national-level estimates of reproduction numbers are of limited value in guiding local interventions and monitoring their effectiveness. It is crucial for national and local policy makers as well as health protection teams that accurate, well-calibrated and timely predictions of Covid-19 incidences and transmission rates are available at fine spatial scales. Obtaining such estimates is challenging, not least due to the prevalence of asymptomatic Covid-19 transmissions, as well as difficulties of obtaining high resolution and frequency data. In addition, low case counts at a local level further confounds the inference for Covid-19 transmission rates, adding unwelcome uncertainty. In this paper we develop a hierarchical Bayesian method for inference of incidence and transmission rates at fine spatial scales. Our model incorporates both temporal and spatial dependencies of local transmission rates in order to share statistical strength and reduce uncertainty. It also incorporates information about population flows to model potential

transmissions across local areas. A simple approach to posterior simulation quickly becomes computationally infeasible, which is problematic if the system is required to provide timely predictions. We describe how to make posterior simulation for the model efficient, so that we are able to provide daily updates on epidemic developments. Real-time estimates from our model can be viewed on our website: localcovid.info.

U-Clip: On-Average Unbiased Stochastic Gradient Clipping

Bryn Elesedy, Marcus Hutter

preprint

https://arxiv.org/abs/2302.02971

Abstract

U-Clip is a simple amendment to gradient clipping that can be applied to any iterative gradient optimization algorithm. Like regular clipping, U-Clip involves using gradients that are clipped to a prescribed size (e.g., with component wise or norm based clipping) but instead of discarding the clipped portion of the gradient, U-Clip maintains a buffer of these values that is added to the gradients on the next iteration (before clipping). We show that the cumulative bias of the U-Clip updates is bounded by a constant. This implies that the clipped updates are unbiased on average. Convergence follows via a lemma that guarantees convergence with updates u_i as long as $\sum_{i=1}^{t}(u_i - g_i) = o(t)$ where g_i are the gradients. Extensive experimental exploration is performed on CIFAR10 with further validation given on ImageNet.

Note

Work performed while on an internship at DeepMind. Unfortunately, after finishing, we found out that the main results are basically a special case of [86].

References

[1] Emmanuel Abbe and Enric Boix-Adserà. "On the non-universality of deep learning: quantifying the cost of symmetry". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: https://openreview.net/forum?id=Leg6spUEFFf (page 28).

[2] Nilin Abrahamsen and Lin Lin. "Anti-symmetric Barron functions and their approximation with sums of determinants". In: *arXiv preprint arXiv:2303.12856* (2023) (page 24).

[3] Yaser S Abu-Mostafa. "Hints and the VC dimension". In: *Neural Computation* 5.2 (1993), pp. 278–288 (page 17).

[4] John Frank Adams. *Lectures on Lie groups*. University of Chicago Press, 1982 (page 11).

[5] Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. "Kernels for vector-valued functions: A review". In: *Foundations and Trends® in Machine Learning* 4.3 (2012), pp. 195–266 (page 115).

[6] Brandon Anderson, Truong Son Hy, and Risi Kondor. "Cormorant: Covariant molecular neural networks". In: *Advances in neural information processing systems* 32 (2019) (page 26).

[7] Fabio Anselmi, Lorenzo Rosasco, and Tomaso Poggio. "On invariance and selectivity in representation learning". In: *Information and Inference: A Journal of the IMA* 5.2 (2016), pp. 134–158 (pages 18, 21).

[8] Fabio Anselmi et al. "Symmetry-adapted representation learning". In: *Pattern Recognition* 86 (2019), pp. 201–208 (page 27).

[9] Fabio Anselmi et al. "Unsupervised learning of invariant representations". In: *Theoretical Computer Science* 633 (2016), pp. 112–121 (page 18).

[10] Fabio Anselmi et al. *Unsupervised Learning of Invariant Representations in Hierarchical Architectures*. 2014. arXiv: 1311.4158 [cs.CV] (page 18).

[11] Martin Anthony and Peter L. Bartlett. *Neural network learning: theoretical foundations*. eng. Cambridge: Cambridge University Press, 1999. ISBN: 9780521573535 (page 4).

[12] Nachman Aronszajn. "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404 (page 62).

[13] Benjamin Aslan, Daniel Platt, and David Sheard. "Group invariant machine learning by fundamental domain projections". In: *NeurIPS Workshop on Symmetry and Geometry in Neural Representations*. PMLR. 2023, pp. 181–218 (page 21).

[14] Waiss Azizian and Marc Lelarge. "Expressive Power of Invariant and Equivariant Graph Neural Networks". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=lxHgXYN4bwl (page 20).

[15] Peter L Bartlett et al. "Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks." In: *J. Mach. Learn. Res.* 20 (2019), pp. 63–1 (page 106).

[16] Arash Behboodi, Gabriele Cesa, and Taco Cohen. "A PAC-Bayesian Generalization Bound for Equivariant Networks". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: https://openreview.net/forum?id=6dfYc2IUj4 (page 19).

[17] Erik J Bekkers et al. "Roto-translation covariant convolutional networks for medical image analysis". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer. 2018, pp. 440–448 (page 26).

[18] Gregory Benton et al. “Learning invariances in neural networks from training data”. In: *Advances in neural information processing systems* 33 (2020), pp. 17605–17616 (page 27).

[19] Alberto Bietti and Julien Mairal. “Group invariance, stability to deformations, and complexity of deep convolutional representations”. In: *The Journal of Machine Learning Research* 20.1 (2019), pp. 876–924 (page 23).

[20] Alberto Bietti, Luca Venturi, and Joan Bruna. “On the Sample Complexity of Learning under Geometric Stability”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021. URL: https://openreview.net/forum?id=vlf0zTKa5Lh (page 19).

[21] Ben Bloem-Reddy. “Lecture notes for STAT 547S: Topics in Symmetry in Statistics and Machine Learning (draft; in progress)”. In: (2023). URL: https://www.stat.ubc.ca/~benbr/assets/notes/stat547s-notes.pdf (page 17).

[22] Benjamin Bloem-Reddy and Yee Whye Teh. “Probabilistic symmetries and invariant neural networks”. In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 3535–3595. URL: http://jmlr.org/papers/v21/19-322.html (pages 17, 22, 25, 43, 92, 93, 95, 107).

[23] Ben Blum-Smith and Soledad Villar. *Equivariant maps from invariant functions*. 2022. arXiv: 2209.14991 [stat.ML] (page 21).

[24] Alexander Bogatskiy et al. “Lorentz group equivariant neural network for particle physics”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 992–1002 (page 26).

[25] Denis Boyda et al. “Sampling using SU (N) gauge equivariant flows”. In: *Physical Review D* 103.7 (2021), p. 074504 (page 20).

[26] Michael M Bronstein et al. “Geometric deep learning: going beyond euclidean data”. In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 18–42 (page 28).

[27] Michael M. Bronstein et al. *Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges.* 2021. arXiv: 2104 . 13478 [cs.LG] (page 28).

[28] Joan Bruna and Stéphane Mallat. "Invariant scattering convolution networks". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1872–1886 (page 20).

[29] Olivier Chapelle and Bernhard Schölkopf. "Incorporating invariances in nonlinear support vector machines". In: *Advances in neural information processing systems* 14 (2001) (page 20).

[30] An Mei Chen, Haw-minn Lu, and Robert Hecht-Nielsen. "On the Geometry of Feedforward Neural Network Error Surfaces". In: *Neural Computation* 5.6 (1993), pp. 910–927. DOI: 10.1162/neco.1993.5.6.910 (page 29).

[31] Shuxiao Chen, Edgar Dobriban, and Jane H Lee. "A group-theoretic framework for data augmentation". In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 9885–9955 (page 29).

[32] Ziyu Chen and Wei Zhu. "On the Implicit Bias of Linear Equivariant Steerable Networks: Margin, Generalization, and Their Equivalence to Data Augmentation". In: *arXiv preprint arXiv:2303.04198* (2023) (page 23).

[33] Louis G Christie and John AD Aston. "Testing for Geometric Invariance and Equivariance". In: *arXiv preprint arXiv:2205.15280* (2022) (page 27).

[34] Taco Cohen and Max Welling. "Group equivariant convolutional networks". In: *International conference on machine learning.* PMLR. 2016, pp. 2990–2999 (pages 2, 22, 92, 112).

[35] Taco S Cohen, Mario Geiger, and Maurice Weiler. "A general theory of equivariant cnns on homogeneous spaces". In: *Advances in neural information processing systems* 32 (2019) (pages 2, 22, 112).

[36] Taco S. Cohen and Max Welling. "Steerable CNNs". In: *International Conference on Learning Representations.* 2017. URL: https://openreview.net/forum?id=rJQKYt5ll (page 22).

[37] Taco S. Cohen et al. “Spherical CNNs”. In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=Hkbd5xZRb (pages 22, 112).

[38] R Dennis Cook, Liliana Forzani, et al. “On the mean and variance of the generalized inverse of a singular Wishart matrix”. In: *Electronic Journal of Statistics* 5 (2011), pp. 146–158 (page 121).

[39] Miles Cranmer et al. “Lagrangian Neural Networks”. In: *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*. 2019. URL: https://openreview.net/forum?id=iE8tFa4Nq (page 28).

[40] Ekin D. Cubuk et al. “AutoAugment: Learning Augmentation Strategies From Data”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 113–123. DOI: 10.1109/CVPR.2019.00020 (page 27).

[41] Felipe Cucker and Steve Smale. “On the mathematical foundations of learning”. In: *Bulletin of the American mathematical society* 39.1 (2002), pp. 1–49 (page 4).

[42] Dennis DeCoste and Bernhard Schölkopf. “Training invariant support vector machines”. In: *Machine learning* 46 (2002), pp. 161–190 (page 20).

[43] Nima Dehmamy et al. “Automatic symmetry discovery with lie algebra convolutional network”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 2503–2515 (page 27).

[44] Krish Desai, Benjamin Nachman, and Jesse Thaler. “Symmetry discovery with deep learning”. In: *Physical Review D* 105.9 (2022), p. 096031 (page 27).

[45] Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. “Exploiting cyclic symmetry in convolutional neural networks”. In: *International conference on machine learning*. PMLR. 2016, pp. 1889–1898 (page 21).

[46] Sander Dieleman, Kyle W Willett, and Joni Dambre. “Rotation-invariant convolutional neural networks for galaxy morphology prediction”. In:

Monthly notices of the royal astronomical society 450.2 (2015), pp. 1441–1459 (page 26).

[47] Zhijian Duan, Yunxuan Ma, and Xiaotie Deng. *Are Equivariant Equilibrium Approximators Beneficial?* 2023. arXiv: 2301.11481 [cs.GT] (page 26).

[48] Nadav Dym and Haggai Maron. "On the Universality of Rotation Equivariant Point Cloud Networks". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=6NFBvWlRXaG (page 25).

[49] Morris L Eaton. "Group invariance applications in statistics". In: *Regional conference series in Probability and Statistics*. JSTOR. 1989, pp. i–133 (pages 17, 89, 91, 103).

[50] Bryn Elesedy. "Group Symmetry in PAC Learning". In: *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*. 2022. URL: https://openreview.net/forum?id=HxeTEZJaxq (pages 9, 18).

[51] Bryn Elesedy. "Provably Strict Generalisation Benefit for Invariance in Kernel Methods". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 17273–17283. URL: https://proceedings.neurips.cc/paper/2021/file/8fe04df45a22b63156ebabbb064fcd5e-Paper.pdf (pages iv, 9, 19, 110).

[52] Bryn Elesedy and Sheheryar Zaidi. "Provably Strict Generalisation Benefit for Equivariant Models". In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 2959–2969. URL: http://proceedings.mlr.press/v139/elesedy21a.html (pages iv, 9, 19, 37).

[53] Carlos Esteves. "Theoretical aspects of group equivariant neural networks". In: *arXiv preprint arXiv:2004.05154* (2020) (page 22).

[54] Carlos Esteves et al. “Learning so (3) equivariant representations with spherical cnns”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 52–68 (page 22).

[55] Alhussein Fawzi and Pascal Frossard. “Manitest: Are classifiers really invariant?” In: *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, 2015, pp. 106.1–106.13. ISBN: 1-901725-53-7. DOI: 10.5244/C.29.106. URL: https://dx.doi.org/10.5244/C.29.106 (page 29).

[56] Marc Finzi, Max Welling, and Andrew Gordon Wilson. “A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 3318–3328 (page 21).

[57] Marc Finzi et al. “Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 3165–3176 (pages 2, 22).

[58] Gerald B Folland. *A course in abstract harmonic analysis*. Vol. 29. CRC press, 2016 (page 14).

[59] Jonathan Frankle and Michael Carbin. “The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks”. In: *ICLR* (2019) (page 124).

[60] William T Freeman, Edward H Adelson, et al. “The design and use of steerable filters”. In: *IEEE Transactions on Pattern analysis and machine intelligence* 13.9 (1991), pp. 891–906 (page 20).

[61] Fabian Fuchs et al. “Se (3)-transformers: 3d roto-translation equivariant attention networks”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1970–1981 (page 20).

[62] Fabian B Fuchs et al. “Iterative se (3)-transformers”. In: *Geometric Science of Information: 5th International Conference, GSI 2021, Paris, France, July 21–23, 2021, Proceedings 5*. Springer. 2021, pp. 585–595 (page 20).

[63] Kunihiko Fukushima. “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”. In: *Biological cybernetics* 36.4 (1980), pp. 193–202 (page 20).

[64] William John Andrew Fyfe. “Invariance Hints and the VC Dimension”. PhD thesis. California Institute of Technology, 1992 (page 17).

[65] Robert Gens and Pedro M Domingos. “Deep symmetry networks”. In: *Advances in neural information processing systems* 27 (2014) (page 21).

[66] Jan E Gerken et al. “Geometric deep learning and equivariant neural networks”. In: *arXiv preprint arXiv:2105.13926* (2021) (page 28).

[67] Samuel Greydanus, Misko Dzamba, and Jason Yosinski. “Hamiltonian neural networks”. In: *Advances in neural information processing systems* 32 (2019) (page 28).

[68] Yulan Guo et al. “Deep learning for 3d point clouds: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 43.12 (2020), pp. 4338–4364 (page 25).

[69] Somesh Das Gupta. “Some aspects of discrimination function coefficients”. In: *Sankhyā: The Indian Journal of Statistics, Series A* (1968), pp. 387–400 (page 121).

[70] B. Haasdonk, A. Vossen, and H. Burkhardt. “Invariance in Kernel Methods by Haar Integration Kernels”. In: *SCIA 2005, Scandinavian Conference on Image Analysis*. Springer-Verlag, 2005, pp. 841–851 (pages 20, 82).

[71] Bernard Haasdonk and Hans Burkhardt. “Invariant kernel functions for pattern analysis and machine learning”. In: *Machine learning* 68 (2007), pp. 35–61 (page 20).

[72] Ward Haddadin. “Invariant polynomials and machine learning”. In: *arXiv preprint arXiv:2104.12733* (2021) (page 20).

[73] Jiequn Han et al. “Universal approximation of symmetric and anti-symmetric functions”. In: *Communications in Mathematical Sciences* 20.5 (2022), pp. 1397–1408 (page 24).

[74] Jason Hartford et al. “Deep models of interactions across sets”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1909–1918 (page 26).

[75] Trevor Hastie et al. “Surprises in high-dimensional ridgeless least squares interpolation”. In: *The Annals of Statistics* 50.2 (2022), pp. 949–986 (pages 6, 51).

[76] David Haussler. “Decision theoretic generalizations of the PAC model for neural net and other learning applications”. In: *Information and computation* 100.1 (1992), pp. 78–150 (pages 4, 88).

[77] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90 (page 44).

[78] Lingshen He et al. “Gauge equivariant transformer”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 27331–27343 (page 20).

[79] Philip G. Hodge. “On Isotropic Cartesian Tensors”. eng. In: *The American Mathematical Monthly* 68.8 (1961), pp. 793–795. ISSN: 00029890 (pages 79, 122).

[80] Peter Holderrieth, Michael J Hutchinson, and Yee Whye Teh. “Equivariant learning of stochastic fields: Gaussian processes and steerable conditional neural processes”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 4297–4307 (page 20).

[81] Richard B Holmes. *Mathematical foundations of signal processing. 2. the role of group theory*. Tech. rep. MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB, 1987 (page 20).

[82] Michael J Hutchinson et al. “Lietransformer: Equivariant self-attention for lie groups”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 4533–4543 (page 20).

[83] Marcus Hutter. “On representing (anti) symmetric functions”. In: *arXiv preprint arXiv:2007.15298* (2020) (page 24).

[84] John Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (2021), pp. 583–589 (pages 3, 26).

[85] Olav Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006 (pages 14–16, 22, 44, 93, 116, 120).

[86] Sai Praneeth Karimireddy et al. “Error feedback fixes signsgd and other gradient compression schemes”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 3252–3261 (page 128).

[87] Dhruva Kashyap, Natarajan Subramanyam, et al. “Robustness to augmentations as a generalization metric”. In: *arXiv preprint arXiv:2101.06459* (2021) (page 29).

[88] Michael J Kearns, Robert E Schapire, and Linda M Sellie. “Toward Efficient Agnostic Learning”. In: *Machine Learning* 17.2 (1994), pp. 115–141 (page 4).

[89] Jinpyo Kim et al. “Cycnn: A rotation invariant cnn using polar mapping and cylindrical convolution layers”. In: *arXiv preprint arXiv:2007.10588* (2020) (page 22).

[90] George S Kimeldorf and Grace Wahba. “A correspondence between Bayesian estimation on stochastic processes and smoothing by splines”. In: *The Annals of Mathematical Statistics* 41.2 (1970), pp. 495–502 (page 65).

[91] Jonas Köhler, Leon Klein, and Frank Noé. “Equivariant flows: exact likelihood generative learning for symmetric densities”. In: *International conference on machine learning*. PMLR. 2020, pp. 5361–5370 (page 20).

[92] Vladimir Koltchinskii and Dmitriy Panchenko. “Rademacher processes and bounding the risk of function learning”. In: *High dimensional probability II*. Springer. 2000, pp. 443–457 (page 4).

[93] Imre Risi Kondor. “Group theoretical methods in machine learning”. PhD thesis. Columbia University, 2008 (page 111).

[94] Risi Kondor, Zhen Lin, and Shubhendu Trivedi. “Clebsch–gordan nets: a fully fourier space spherical convolutional neural network”. In: *Advances in Neural Information Processing Systems* 31 (2018) (pages 2, 22).

[95] Risi Kondor and Shubhendu Trivedi. "On the generalization of equivariance and convolution in neural networks to the action of compact groups". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2747–2755 (pages 22, 113).

[96] Akshay Krishnamurthy. *Lecture 12: Surrogate Losses and Calibration*. https://people.cs.umass.edu/~akshay/courses/cs690m/files/lec12.pdf. Accessed: 2023–04-18. 2017 (page 114).

[97] Daniel Kunin et al. "Neural Mechanics: Symmetry and Broken Conservation Laws in Deep Learning Dynamics". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=q8qLAbQBupm (page 29).

[98] Leon Lang and Maurice Weiler. "A Wigner-Eckart Theorem for Group Equivariant Convolution Kernels". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=ajOrOhQOsYx (page 22).

[99] Hannah Lawrence. "Barron's Theorem for Equivariant Networks". In: *NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representations*. 2022. URL: https://openreview.net/forum?id=1G_dwmEKX-w (page 22).

[100] Hannah Lawrence et al. "Implicit Bias of Linear Equivariant Networks". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 12096–12125 (page 23).

[101] Juho Lee et al. "Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 3744–3753. URL: http://proceedings.mlr.press/v97/lee19d.html (page 23).

[102] Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*. Vol. 3. Springer, 2005 (page 25).

[103] Karel Lenc and Andrea Vedaldi. "Understanding image representations by measuring their equivariance and equivalence". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 991–999 (page 22).

[104] Jan Eric Lenssen, Matthias Fey, and Pascal Libuschewski. "Group equivariant capsule networks". In: *Advances in neural information processing systems* 31 (2018) (page 20).

[105] Zhiyuan Li, Yi Zhang, and Sanjeev Arora. "Why Are Convolutional Nets More Sample-Efficient than Fully-Connected Nets?" In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=uCY5MuAxcxU (page 28).

[106] Julia Ling, Andrew Kurzawski, and Jeremy Templeton. "Reynolds averaged turbulence modelling using deep neural networks with embedded invariance". In: *Journal of Fluid Mechanics* 807 (2016), pp. 155–166 (page 26).

[107] Di Luo et al. "Gauge equivariant neural networks for quantum lattice gauge theories". In: *Physical review letters* 127.27 (2021), p. 276402 (page 26).

[108] Di Luo et al. "Gauge-invariant and anyonic-symmetric autoregressive neural network for quantum lattice models". In: *Physical Review Research* 5.1 (2023), p. 013216 (page 26).

[109] Clare Lyle, Marta Kwiatkowksa, and Yarin Gal. "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 (page 19).

[110] Clare Lyle et al. *On the Benefits of Invariance in Neural Networks*. 2020. arXiv: 2005.00178 [cs.LG] (page 29).

[111] Stéphane Mallat. "Group invariant scattering". In: *Communications on Pure and Applied Mathematics* 65.10 (2012), pp. 1331–1398 (page 20).

[112] Diego Marcos et al. “Rotation equivariant vector field networks”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 5048–5057 (page 21).

[113] Haggai Maron et al. “Invariant and Equivariant Graph Networks”. In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=Syx72jC9tm (page 20).

[114] Haggai Maron et al. “On learning sets of symmetric elements”. In: *International conference on machine learning*. PMLR. 2020, pp. 6734–6744 (page 23).

[115] Haggai Maron et al. “On the universality of invariant networks”. In: *International conference on machine learning*. PMLR. 2019, pp. 4363–4371 (page 22).

[116] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. “Learning with invariances in random features and kernel models”. In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, 2021, pp. 3351–3418. URL: https://proceedings.mlr.press/v134/mei21a.html (pages 4, 19, 110, 111, 114).

[117] Adam M. Meier. *Randomized Benchmarking of Clifford Operators*. 2018. arXiv: 1811.10040 [quant-ph] (page 12).

[118] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. 2nd ed. MIT Press, 2018 (pages 4, 99).

[119] Jaouad Mourtada and Lorenzo Rosasco. “An elementary analysis of ridge regression with random design”. In: *Comptes Rendus. Mathématique* 360.G9 (2022), pp. 1055–1063 (page 69).

[120] Youssef Mroueh, Stephen Voinea, and Tomaso A Poggio. “Learning with Group Invariant Features: A Kernel Perspective.” In: *Advances in Neural Information Processing Systems*. 2015, pp. 1558–1566 (page 18).

[121] Robb J Muirhead. *Aspects of multivariate statistical theory*. Vol. 197. John Wiley & Sons, 2009 (page 121).

[122] Ido Nachum and Amir Yehudayoff. “On symmetry and initialization for neural networks”. In: *LATIN 2020: Theoretical Informatics: 14th Latin American Symposium, São Paulo, Brazil, January 5-8, 2021, Proceedings 14*. Springer. 2020, pp. 401–412 (page 24).

[123] Andrew Y Ng. “Feature selection, L 1 vs. L 2 regularization, and rotational invariance”. In: *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 78 (page 28).

[124] Dipan Pal et al. “Max-margin invariant features from transformed unlabelled data”. In: *Advances in Neural Information Processing Systems* 30 (2017) (pages 110, 111).

[125] Dipan K Pal, Felix Juefei-Xu, and Marios Savvides. “Discriminative invariant kernel features: a bells-and-whistles-free approach to unsupervised face recognition and pose estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 5590–5599 (page 111).

[126] Roger Penrose. “A generalized inverse for matrices”. In: *Mathematical proceedings of the Cambridge philosophical society*. Vol. 51. 3. Cambridge University Press. 1955, pp. 406–413 (page 118).

[127] Philipp Christian Petersen and Anna Sepliarskaia. “VC dimensions of group convolutional neural networks”. In: *arXiv preprint arXiv:2212.09507* (2022) (page 23).

[128] David Pfau, James S Spencer, and Alexander GDG Matthews. “Ab initio solution of the many-electron Schrödinger equation with deep neural networks”. In: *Physical Review Research* 2.3 (2020), p. 033429 (pages 3, 26).

[129] Vasco Portilheiro. *A tradeoff between universality of equivariant models and learnability of symmetries*. 2022. arXiv: 2210.09444 [stat.ML] (pages 22, 27).

[130] Omri Puny et al. “Frame Averaging for Invariant and Equivariant Network Design”. In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=zIUyj55nXR (page 20).

[131] Charles R Qi et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 652–660 (page 25).

[132] Tian Qin, Fengxiang He, and Dacheng Tao. *Intrinsic Computational Complexity of Equivariant Neural Networks*. 2023. URL: https://openreview.net/forum?id=-MQWXqNyoa (page 23).

[133] Tian Qin et al. "Benefits of Permutation-Equivariance in Auction Mechanisms". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: https://openreview.net/forum?id=bI1XXtO-hs2 (page 26).

[134] Michael Ragone et al. *Representation Theory for Geometric Quantum Machine Learning*. 2022. arXiv: 2210.07980 [quant-ph] (page 12).

[135] Jad Rahme et al. "A Permutation-Equivariant Neural Network Architecture For Auction Design". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.6 (2021), 5664–5672. ISSN: 2159-5399. DOI: 10.1609/aaai.v35i6.16711. URL: http://dx.doi.org/10.1609/aaai.v35i6.16711 (page 26).

[136] Anant Raj et al. "Local group invariant representations via orbit embeddings". In: *Artificial Intelligence and Statistics*. PMLR. 2017, pp. 1225–1235 (page 20).

[137] Siamak Ravanbakhsh. "Universal equivariant multilayer perceptrons". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7996–8006 (page 22).

[138] Siamak Ravanbakhsh, Jeff Schneider, and Barnabás Póczos. "Equivariance Through Parameter-Sharing". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 2892–2901. URL: https://proceedings.mlr.press/v70/ravanbakhsh17a.html (pages 21, 104).

[139] Marco Reisert and Hans Burkhardt. "Learning Equivariant Functions with Matrix Valued Kernels". In: *Journal of Machine Learning Research* 8.3 (2007) (pages 31, 86, 111, 112, 115).

[140] Danilo Jimenez Rezende et al. "Equivariant hamiltonian flows". In: *arXiv preprint arXiv:1909.13739* (2019) (page 20).

[141] David W Romero and Suhas Lohit. "Learning partial equivariances from data". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 36466–36478 (page 27).

[142] Walter Rudin. *Functional analysis.* eng. 2nd ed. International series in pure and applied mathematics. New York ; London: McGraw-Hill, 1991. ISBN: 9780070542365 (pages 118, 119).

[143] Walter Rudin. *Real and complex analysis.* eng. 3rd ed., international ed. New York: McGraw-Hill, 1987. ISBN: 9780070542341 (pages 14, 85).

[144] Stefan M Rüger and Arnfried Ossen. "Clustering in weight space of feedforward nets". In: *Artificial Neural Networks—ICANN 96: 1996 International Conference Bochum, Germany, July 16–19, 1996 Proceedings 6.* Springer. 1996, pp. 83–88 (page 29).

[145] Akiyoshi Sannai, Masaaki Imaizumi, and Makoto Kawano. "Improved generalization bounds of group invariant/equivariant deep networks via quotient feature spaces". In: *Uncertainty in Artificial Intelligence.* PMLR. 2021, pp. 771–780 (pages 18, 92, 96).

[146] Akiyoshi Sannai, Yuuki Takai, and Matthieu Cordonnier. "Universal approximations of permutation invariant/equivariant functions by deep neural networks". In: *arXiv preprint arXiv:1903.01939* (2019) (page 24).

[147] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. "E (n) equivariant graph neural networks". In: *International conference on machine learning.* PMLR. 2021, pp. 9323–9332 (page 20).

[148] Victor Garcia Satorras et al. “E(n) Equivariant Normalizing Flows”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021. URL: https://openreview.net/forum?id=N5hQI_RowVA (page 20).

[149] Bernhard Schölkopf, Chris Burges, and Vladimir Vapnik. “Incorporating invariances in support vector learning machines”. In: *Artificial Neural Networks—ICANN 96: 1996 International Conference Bochum, Germany, July 16–19, 1996 Proceedings 6*. Springer. 1996, pp. 47–52 (page 20).

[150] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. “A generalized representer theorem”. In: *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001 Amsterdam, The Netherlands, July 16–19, 2001 Proceedings 14*. Springer. 2001, pp. 416–426 (page 65).

[151] H. Schulz-Mirbach. “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)*. Vol. 2. IEEE. 1994, pp. 387–390 (page 20).

[152] H Schulz-Mirbach. “On the existence of complete invariant feature spaces in pattern recognition”. In: *Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol. II. Conference B: Pattern Recognition Methodology and Systems*. IEEE. 1992, pp. 178–182 (page 20).

[153] Nimrod Segol and Yaron Lipman. “On Universal Equivariant Set Networks”. In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=HkxTwkrKDB (page 24).

[154] Jean-Pierre Serre. *Linear representations of finite groups*. eng. Graduate texts in mathematics ; 42. New York: Springer-Verlag, 1977. ISBN: 9780387901909 (page 9).

[155] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014 (pages 4, 88).

[156] Han Shao, Omar Montasser, and Avrim Blum. "A Theory of PAC Learnability under Transformation Invariances". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: https://openreview.net/forum?id=l1WlfNaRkKw (pages 18, 19, 29).

[157] John Shawe-Taylor. "Building symmetries into feedforward networks". In: *1989 First IEE International Conference on Artificial Neural Networks,(Conf. Publ. No. 313)*. IET. 1989, pp. 158–162 (page 21).

[158] John Shawe-Taylor. "Introducing invariance: a principled approach to weight sharing". In: *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*. Vol. 1. IEEE. 1994, pp. 345–349 (page 21).

[159] John Shawe-Taylor. "Sample sizes for threshold networks with equivalences". In: *Information and Computation* 118.1 (1995), pp. 65–72 (page 17).

[160] John Shawe-Taylor. "Symmetries and discriminability in feedforward network architectures". In: *IEEE Transactions on Neural Networks* 4.5 (1993), pp. 816–826 (page 21).

[161] John Shawe-Taylor. "Threshold network learning in the presence of equivalences". In: *Advances in neural information processing systems* 4 (1991) (page 17).

[162] Patrice Simard et al. "Tangent prop-a formalism for specifying selected invariances in an adaptive network". In: *Advances in neural information processing systems* 4 (1991) (page 21).

[163] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations*. 2015 (page 44).

[164] Berfin Simsek et al. "Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 9722–9732 (page 29).

[165] Jure Sokolic et al. "Generalization error of invariant classifiers". In: *Artificial Intelligence and Statistics*. PMLR. 2017, pp. 1094–1103 (pages 18, 96).

[166] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information science and statistics. Springer, 2008. ISBN: 978-0-387-77241-7 (pages 62, 123).

[167] Ingo Steinwart and Clint Scovel. "Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs". In: *Constructive Approximation* 35.3 (2012), pp. 363–417 (page 64).

[168] Christian Szegedy et al. "Going deeper with convolutions". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594 (page 44).

[169] Behrooz Tahmasebi and Stefanie Jegelka. "The Exact Sample Complexity Gain from Invariances for Kernel Regression on Manifolds". In: *arXiv preprint arXiv:2303.14269* (2023) (page 19).

[170] Leslie G Valiant. "A theory of the learnable". In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142 (pages 4, 88).

[171] Mark van der Wilk et al. "Learning invariances using the marginal likelihood". In: *Advances in Neural Information Processing Systems* 31 (2018) (page 27).

[172] Vladimir N Vapnik and A Ya Chervonenkis. "On the uniform convergence of relative frequencies of events to their probabilities". In: *Measures of complexity*. Springer, 2015, pp. 11–30 (page 88).

[173] VN Vapnik and A Ya Chervonenkis. "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities". In: *Theory of Probability and its Applications* 16.2 (1971), p. 264 (page 4).

[174] Alfonso Villani. "Another Note On The Inclusion $L_p(\mu) \subset L_q(\mu)$". eng. In: *The American mathematical monthly* 92.7 (1985), pp. 485–C76. ISSN: 0002-9890 (page 13).

[175] Soledad Villar et al. *Dimensionless machine learning: Imposing exact units equivariance*. 2022. arXiv: 2204.00887 [stat.ML] (pages 28, 32).

[176] Soledad Villar et al. “Scalars are universal: Equivariant machine learning, structured like classical physics”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021. URL: https://openreview.net/forum?id=ba27-RzNaIv (page 21).

[177] Soledad Villar et al. *The passive symmetries of machine learning*. 2023. arXiv: 2301.13724 [stat.ML] (page 28).

[178] Ernesto De Vito et al. “Learning from Examples as an Inverse Problem”. In: *Journal of Machine Learning Research* 6.30 (2005), pp. 883–904. URL: http://jmlr.org/papers/v6/devito05a.html (pages 73–75).

[179] Simon Wadsley. *Lecture Notes on Representation Theory*. 2012. URL: https://www.dpmms.cam.ac.uk/~sjw47/RepThLecturesMich2012.pdf (page 9).

[180] Edward Wagstaff et al. “On the limitations of representing functions on sets”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6487–6494 (page 24).

[181] Edward Wagstaff et al. “Universal approximation of functions on sets”. In: *Journal of Machine Learning Research* 23.151 (2022), pp. 1–56 (page 24).

[182] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019 (page 39).

[183] Christian Walder and Olivier Chapelle. “Learning with transformation invariant kernels”. In: *Advances in Neural Information Processing Systems* 20 (2007) (page 20).

[184] Dian Wang et al. *A General Theory of Correct, Incorrect, and Extrinsic Equivariance*. 2023. arXiv: 2303.04745 [cs.LG] (page 19).

[185] Renhao Wang, Marjan Albooyeh, and Siamak Ravanbakhsh. “Equivariant networks for hierarchical structures”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 13806–13817 (page 23).

[186] Maurice Weiler and Gabriele Cesa. “General e (2)-equivariant steerable cnns”. In: *Advances in Neural Information Processing Systems* 32 (2019) (pages 2, 22).

[187] Maurice Weiler et al. “Coordinate Independent Convolutional Networks–Isometry and Gauge Equivariant Convolutions on Riemannian Manifolds”. In: *arXiv preprint arXiv:2106.06020* (2021) (page 22).

[188] Marysia Winkels and Taco S. Cohen. “3D G-CNNs for Pulmonary Nodule Detection”. In: *Medical Imaging with Deep Learning*. 2018. URL: https://openreview.net/forum?id=H1sdHFiif (pages 3, 26, 37).

[189] Robin Winter et al. “Unsupervised Learning of Group Invariant and Equivariant Representations”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: https://openreview.net/forum?id=47lpv23LDPr (page 20).

[190] Jeffrey Wood. “Invariant pattern recognition: a review”. In: *Pattern recognition* 29.1 (1996), pp. 1–17 (page 20).

[191] Jeffrey Wood and John Shawe-Taylor. “Representation theory and invariant neural networks”. In: *Discrete applied mathematics* 69.1-2 (1996), pp. 33–60 (pages 21, 104, 105, 107, 110).

[192] Huan Xu and Shie Mannor. “Robustness and generalization”. In: *Machine learning* 86.3 (2012), pp. 391–423 (page 18).

[193] Jin Xu et al. “Group equivariant subsampling”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 5934–5946 (pages 20, 22).

[194] Dmitry Yarotsky. “Universal approximations of invariant maps by neural networks”. In: *Constructive Approximation* 55.1 (2022), pp. 407–474 (page 22).

[195] Manzil Zaheer et al. “Deep sets”. In: *Advances in neural information processing systems* 30 (2017) (pages 24, 92, 95, 107).

[196] Xinhua Zhang, Wee Sun Lee, and Yee Whye Teh. “Learning with invariance via linear functionals on reproducing kernel hilbert space”. In: *Advances in Neural Information Processing Systems* 26 (2013) (page 20).

[197] Allan Zhou, Tom Knowles, and Chelsea Finn. “Meta-learning Symmetries by Reparameterization”. In: *International Conference on Learning Repre-*

sentations. 2021. URL: https://openreview.net/forum?id=-QxT4mJdijq (page 27).

[198] Sicheng Zhu, Bang An, and Furong Huang. "Understanding the generalization benefit of model invariance from a data perspective". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 4328–4341 (page 18).