GROUP SYMMETRY IN PAC LEARNING

Bryn Elesedy
Department of Computer Science, University of Oxford
Wolfson Building, Parks Road, Oxford OX1 3QD
bryn@robots.ox.ac.uk

ABSTRACT

In this paper we show rigorously how learning in the PAC framework with invariant or equivariant hypotheses reduces to learning in a space of orbit representatives. Our results hold for any compact group, including infinite groups such as rotations. In addition, we show how to use these equivalences to derive generalisation bounds for invariant/equivariant models in terms of the geometry of the input and output spaces. To the best of our knowledge, our results are the most general of their kind to date.

1 INTRODUCTION

There is renewed interest in engineering models, such as neural networks, to be invariant or equivariant to the action of a group on their inputs. The intuition being that if the relevant aspects of the task are unchanged (or change predictably) by the action of the group, then models that are hard-coded to be invariant (or equivariant) to this group will require fewer examples to learn. Work in this area has garnered much attention (Cohen & Welling, 2016; Cohen et al., 2019) and enjoys application to domains where the symmetry is known a priori, such as quantum mechanics (Pfau et al., 2019).

We approach the study of symmetry in machine learning from a theoretical perspective, using concepts from statistical learning theory to give sample complexity guarantees. We find that, when the symmetry of the model is correctly specified (is also a symmetry of the target) there is a reduction in sample complexity relative to an arbitrary model. Moreover, the reduction in sample complexity (or, equivalently, the improvement in generalisation) for invariant/equivariant models is in terms of the geometry of a set of orbit representatives and a canonical collection of outputs, relative to that of the entire space of inputs and outputs. Informally, if these canonicalised spaces are much smaller, then we expect a large sample complexity improvement from invariant/equivariant models. This agrees with intuition: if the set of orbit representatives is much smaller than the whole space, then there is a lot of information in the group transformations that is not relevant to the task.

Contributions We begin in Section 2 with our technical setup and notation. In Proposition 3.1, we show how to make precise the common intuition that learning with invariant hypotheses is equivalent to a reduced problem on a set of orbit representatives. We go further in Section 4, giving the first of our main results, Theorem 4.1, in which we show that learning with equivariant hypotheses depends only a space of canonicalised inputs and targets. The second of our main results, Theorem 5.2, allows us to use the aforementioned equivalences to derive sample complexity improvements for equivariant models in terms of a measure of the geometrical quantities we alluded to previously. All of this holds for arbitrary compact (possibly infinite) groups. Proofs are deferred to the appendix.

2 PRELIMINARIES

Notation and Technicalities We write $\mathcal{X}$ and $\mathcal{Y}$ for input and output spaces respectively and set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. There will be a background probability space $(\Omega, \sigma_\Omega, \mathbb{P})$ that is assumed to be rich enough to support our analysis. The input and output spaces are assumed to be standard Borel spaces, which are Polish spaces with the Borel σ-algebra. When dealing with real vector spaces, we use $\|\cdot\|_2$ for the Euclidean norm and $\|\cdot\|_\infty$ for the component-wise max magnitude. On a function $f : \mathcal{X} \to \mathbb{R}^d$, $\|f\|_{L_\infty} = \sup_{x\in\mathcal{X}} \|f(x)\|_\infty$. Throughout the paper, $\mathcal{G}$ will represent an arbitrary Hausdorff, second countable and compact topological group that has measurable actions on both $\mathcal{X}$ and $\mathcal{Y}$. For an action ϕ of $\mathcal{G}$ on $\mathcal{X}$ this means that the map $\phi : \mathcal{G} \times \mathcal{X} \to \mathcal{X}$ is measurable, the same goes for an action ψ on $\mathcal{Y}$. The action of $\mathcal{G}$ on $\mathcal{Z}$ is defined by $g(x, y) = (\phi(g)x, \psi(g)y)$. We often write gx as a shorthand for $\phi(g)x$ and similarly for actions on $\mathcal{Y}$ and $\mathcal{Z}$. Our supervised learning setup will be defined by (X, Y), where X (resp. Y) is random element of $\mathcal{X}$ (resp. $\mathcal{Y}$).

Definition 2.1. 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 $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ (the loss function) is integrable.

PAC Learning The PAC framework (Shalev-Shwartz & Ben-David, 2014, Definition 3.3), originally due to Valiant (1984), provides a precise definition of learning from data. We use a distribution dependent relaxation of the agnostic formulation from Haussler (1992). What we arrive at can be considered a form of uniform learning (Vapnik & Chervonenkis, 2015). Let $\mathcal{H}$ be a user-specified class of measurable functions $h : \mathcal{X} \to \mathcal{Y}$ called the hypothesis class. For example, $\mathcal{H}$ could be a neural network architecture with real parameters. Let the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be integrable. An algorithm $\texttt{alg} : \cup_{i\in\mathbb{N}} \mathcal{Z}^i \to \mathcal{H}$ is a measurable map that associates with each finite sequence of points of $\mathcal{Z}$ a hypothesis from $\mathcal{H}$.[1] We say that $\texttt{alg}$ learns $\mathcal{H}$ 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 $\mathbb{P}\left(\mathbb{E}[\ell(h_S(X), Y)|S] \leq \inf_{h\in\mathcal{H}} \mathbb{E}[\ell(h(X), Y)] + \epsilon\right) \geq 1 - \delta$ where $h_S = \texttt{alg}(S)$ and $S \sim (X, Y)^n$ i.i.d. Suppose $\texttt{alg}$ learns $\mathcal{H}$ 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 point wise minimum $m_{\texttt{alg},\mathsf{T}}(\epsilon, \delta) = \min_{m\in\mathcal{M}} m(\epsilon, \delta)$.

Invariance and Equivariance A function $f : \mathcal{X} \to \mathcal{Y}$ is $\mathcal{G}$-invariant if $f(gx) = f(x)\ \forall x \in \mathcal{X}\ \forall g \in \mathcal{G}$. A function $f : \mathcal{X} \to \mathcal{Y}$ is $\mathcal{G}$-equivariant if $f(gx) = gf(x)\ \forall x \in \mathcal{X}\ \forall g \in \mathcal{G}$. A hypothesis class is $\mathcal{G}$-(invariant/equivariant) if all of its members are $\mathcal{G}$-(invariant/equivariant). A measure μ on $\mathcal{X}$ is $\mathcal{G}$-invariant if $\forall g \in \mathcal{G}$ and any μ-measurable $B \subseteq \mathcal{X}$ the pushforward of μ by the action $\mathcal{G}$ equals μ, i.e. $(g_*\mu)(B) = \mu(B)$. This means that if $X \sim \mu$ then $gX \sim \mu\ \forall g \in \mathcal{G}$. When the group is clear from the context we will sometimes just say invariant/equivariant.

Invariant Algorithms An algorithm $\texttt{alg}$ is $\mathcal{G}$-invariant if $\texttt{alg}(\{(g_1x_1, g_1y_1), \ldots, (g_nx_n, g_ny_n)\}) = \texttt{alg}(\{(x_1, y_1), \ldots, (x_n, y_n)\})\ \forall (x_i, y_i) \in \mathcal{Z}$, $\forall g_i \in \mathcal{G}$ and $\forall n \in \mathbb{N}$. If the hypothesis class is $\mathcal{G}$-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 ERM is covered by this. Similarly, if the hypothesis class is $\mathcal{G}$-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 ERM is covered. Highlighting two cases: if the action of $\mathcal{G}$ is via an orthogonal representation, then $\ell(y, y') = l(\langle y, y'\rangle)$ is preserved by $\mathcal{G}$ for any l. The same goes for $\ell(y, y') = l(\|y - y'\|_2)$.

Definition 2.2 ($(\mathcal{H}, \mathcal{G})$-equivalent)**.** We say the tasks T and T' are $(\mathcal{H}, \mathcal{G})$-*equivalent* if, for any $\mathcal{G}$-invariant algorithm $\texttt{alg}$, $\texttt{alg}$ learns $\mathcal{H}$ with respect to T if and only if $\texttt{alg}$ learns $\mathcal{H}$ with respect to T', and the sample complexities are equal, i.e. $m_{\texttt{alg},\mathsf{T}} = m_{\texttt{alg},\mathsf{T}'}$.

Intuitively, if two tasks are $(\mathcal{H}, \mathcal{G})$-equivalent then they are equally 'difficult' to learn with $\mathcal{H}$.

2.1 Orbit Representatives

In the following, $\sigma_{\mathcal{X}}$ is the Borel σ-algebra on $\mathcal{X}$.

Definition 2.3 (Measurable cross-section (Eaton, 1989, Definition 4.1))**.** Let $\mathcal{G}$ be a group acting measurably on $\mathcal{X}$. 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 $\sigma_{\mathcal{X}}$-measurable when $\mathcal{X}_\pi$ has the subspace σ-algebra $\{B \cap \mathcal{X}_\pi : B \in \sigma_{\mathcal{X}}\}$. We call π the *projection*.

A *measurable cross-section* provides a natural way of identifying a collection of orbit representatives with suitable measurability properties. If $\mathcal{G}$ is compact and acts measurably on $\mathcal{X}$ (with respect to a Borel σ-algebra), then a measurable cross-section always exists (Bloem-Reddy & Teh, 2020). Measurable cross-sections are not necessarily unique.

3 Learning with Invariant Models

Proposition 3.1 is a rigorous version of the common intuition that learning with a $\mathcal{G}$-invariant model is equivalent to learning on a space of orbit representatives. In addition, we show that these learning problems have the same sample complexity. The class of orbit representatives is often of much smaller dimension or complexity than the input space, resulting in an improvement in sample complexity for invariant models. Finally, note that our result requires no conditions on the data distribution, in particular there is no constraint for the input distribution to be invariant.

Proposition 3.1. Let $\mathcal{G}$ act measurably on $\mathcal{X}$ and trivially on $\mathcal{Y}$. Let $\mathcal{X}_\pi$ be a measurable cross-section of $\mathcal{X}$ and let π be its projection. Let $\mathcal{H}$ be a hypothesis class of $\mathcal{G}$-invariant functions, let $\texttt{alg} : \cup_{i\in\mathbb{N}} \mathcal{Z}^i \to \mathcal{H}$ be a $\mathcal{G}$-invariant algorithm and let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be and integrable loss function. Then (X, Y, ℓ) and (X_π, Y, ℓ) are $(\mathcal{H}, \mathcal{G})$-equivalent, where $X_\pi = \pi(X)$ with π as in Definition 2.3.

Example 3.2. Let $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Y} = \{0, 1\}$. Proposition 3.1 tells us that learning with a rotationally invariant hypothesis class, such as discs about the origin $\mathcal{H} = \{(x, y) \mapsto \mathbb{1}\left\{x^2 + y^2 \leq r\right\} : r \in \mathbb{R}_+\}$, is equivalent to learning on the reduced space $\mathcal{X} = \mathbb{R}_+$.

[1] Our analysis can also be applied to randomised algorithms.

Example 3.3 (Deep Sets, (Zaheer et al., 2017))**.** Zaheer et al. (2017); Bloem-Reddy & Teh (2020) consider learning functions $f : [0,1]^M \to \mathbb{R}$ that are S_M-invariant, where S_M is the group of permutations on M elements. It is shown that any such continuous function must be of the form $f(T) = \rho\left(\sum_{t\in T}\phi(t)\right)$ where $T \in [0,1]^M$ and $\rho, \phi : \mathbb{R} \to \mathbb{R}$. By Proposition 3.1 we can see, as was shown by Sannai et al. (2021), that learning permutation invariant functions is equivalent to learning the same class of functions restricted to the domain $\{x_1 \geq x_2 \geq \cdots \geq x_M; x_i \in [0,1]\}$. In high dimensions this is a much smaller space than $[0,1]^M$, it is a factor of $M!$ smaller in volume.

Example 3.4 (G-CNN (Cohen & Welling, 2016))**.** Cohen & Welling (2016) 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 $\frac{\pi}{2}$. Proposition 3.1 shows that training a G-CNN based classifier (e.g. by gradient descent on the empirical loss) that is $p4$-invariant is equivalent to learning the network restricted to a single quadrant of the plane.

4 Learning with Equivariant Models

Before we present our first main result, we introduce some additional assumptions, which we discuss in Appendix A.

Assumption 1 (Structure of the marginal $\mathbb{P}_\mathcal{X}$)**.** We assume the existence of a probability distribution ν on $\mathcal{G}$ such that $X \overset{\text{d}}{=} GX_\pi$ where $X \sim \mathbb{P}_\mathcal{X}$, $X_\pi = \pi(X)$, $G \sim \nu$ and $G \perp\!\!\!\perp X_\pi$. Essentially, this says that the orbit that $X \sim \mathbb{P}_\mathcal{X}$ belongs to and where it is in the orbit are independent. For intuition, if $\mathcal{G} = \mathrm{SO}(2)$ acts by rotation on $\mathbb{R}^2$ then any density $f(r,\theta)$ that is separable in polar co-ordinates as $f(r,\theta) = f_{\texttt{rad}}(r)f_{\texttt{ang}}(\theta)$ gives the required independence.

Assumption 2 (Functional representation of Y)**.** We assume that there exists a measurable, $\mathcal{G}$-equivariant function f such that $Y \overset{\text{d}}{=} f(X,\eta)$ where $\eta \sim \mathrm{Unif}[0,1]$ and $\eta \perp\!\!\!\perp X$.[2] Equivariance of f refers to its first argument only $f(gX,\eta) = gf(X,\eta)$. A special case of this is an equivariant target function with independent additive noise ξ such that $g\xi \overset{\text{d}}{=} \xi \ \forall g \in \mathcal{G}$ (e.g. orthogonal representation and Gaussian noise). This is inspired by *noise-outsourcing*, which typically appears with almost sure equality rather than equality in distribution. It is also known as *transfer*, see Kallenberg (2006, Theorem 6.10) and Bloem-Reddy & Teh (2020) for an application to groups.

Theorem 4.1. Let $\mathcal{G}$ act measurably on both $\mathcal{X}$ and $\mathcal{Y}$. Let $\mathcal{H}$ be a hypothesis class of $\mathcal{G}$-equivariant functions and let $\texttt{alg} : \cup_{i\in\mathbb{N}}\mathcal{Z}^i \to \mathcal{H}$ be a $\mathcal{G}$-invariant algorithm. Let $\mathcal{X}_\pi$ be a measurable cross-section of $\mathcal{X}$ and let π be its projection. Let (X,Y) satisfy Assumption 1 and Assumption 2. Then the tasks (X,Y,ℓ) and $(X_\pi,Y_\pi,\bar{\ell})$ are $(\mathcal{H},\mathcal{G})$-equivalent, where $X_\pi = \pi(X)$, $Y_\pi = f(X_\pi,\eta)$ with f and η as in Assumption 2, and $\bar{\ell}(y,y') = \int_\mathcal{G} \ell(gy,gy')\,\mathrm{d}\nu(g)$.

The loss function $\bar{\ell}$ is the average of ℓ over the orbits of $\mathcal{G}$, weighted by the probability of each $g \in \mathcal{G}$. In any instance in which ℓ is preserved by $\mathcal{G}$, we have $\bar{\ell} = \ell$ (see Section 2). 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. Naturally, this canonical task depends the choice of cross-section.

Example 4.2 (Deep Sets, Zaheer et al. (2017))**.** Returning to Zaheer et al. (2017), the authors consider neural network layers $f : \mathbb{R}^M \to \mathbb{R}^M$ with $f(x) = \sigma(\Theta x)$, where $\Theta \in \mathbb{R}^{M\times M}$ and σ is an element-wise non-linearity. It is shown that f is S_M-equivariant if and only if $\Theta_{ij} = a\delta_{ij} + b$ for scalars $a,b \in \mathbb{R}$. In this case $\mathcal{X} = \mathcal{Y} = \mathbb{R}^M$ and S_M acts on each space by permutation. Assume that the marginal distribution on the inputs is exchangeable (we will see in Appendix A that this means that ν exists and is uniform on S_M), then Theorem 4.1 says that learning the class of equivariant f is equivalent to learning on the restricted domain $\{x_1 \geq x_2 \geq \cdots \geq x_M; x_i \in \mathcal{X}\}$ with the averaged loss function $\bar{\ell}(y,y') = \frac{1}{M!}\sum_{\tau\in S_M}\ell(y_\tau,y'_\tau)$ where $(y_\tau)_i = y_{\tau(i)}$ and the same for y'.

5 Implications for Sample Complexity

Theorem 5.2 links the generalisation error 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 5.1 (Covering, covering number)**.** Let (T,d) be a pseudo-metric space and let $U \subset T$. $K \subset T$ is an ϵ-cover of U (with respect to d) if $\forall u \in U\ \exists k \in K$ with $d(u,k) \leq \epsilon$. The ϵ-covering number of U is the smallest cardinality of all the ϵ-covers, i.e. $\texttt{Cov}(U;d,\epsilon) = \inf_{K\in\mathcal{K}}|K|$ where $\mathcal{K}$ is the set of all ϵ-covers of U.

[2] One can check that, with reference to Assumption 1, $\eta \perp\!\!\!\perp G$ and $\eta \perp\!\!\!\perp X_\pi$.

Theorem 5.2. Let $\mathcal{X}$ be a closed subset of a metric space (T, ρ). Let $\mathcal{Y} = B_d(r)$ be the closed Euclidean ball of radius r in $\mathbb{R}^d$. Let $\mathcal{H}$ be a class of functions $h : \mathcal{X} \to \mathcal{Y}$ that are C-Lipschitz in the sense that $\|h(x) - h(x')\|_\infty \leq C\rho(x, x')$ $\forall x, x' \in \mathcal{X}$. Let $S = ((X_1, Y_1), \ldots, (X_n, Y_n)) \sim (X, Y)^n$ be a training sequence drawn i.i.d. for some random variables X, Y. Then, for any $\epsilon \in (0, 1)$,

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\mathbb{E}[\ell(h(X), Y)] - \frac{1}{n}\sum_{i=1}^{n}\ell(h(X_i), Y_i)\right| \geq \epsilon\right) \leq 2 \inf_{\alpha\in(0,1)} D_{\alpha\epsilon}(\mathcal{X}, \mathcal{H}) \exp(-8^{-1}(1-\alpha)^2 n\epsilon^2 r^{-4})$$

where

$$D_\tau(\mathcal{X}, \mathcal{H}) = \texttt{Cov}\left(\mathcal{X}, \rho, \frac{\tau}{32Crd}\right) \sup_{x\in\mathcal{X}} \texttt{Cov}\left(\mathcal{H}(x), \|\cdot\|_\infty, \frac{\tau}{16rd}\right).$$

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

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

examples are sufficient to generalise. The significance of this for equivariant models is as follows. Suppose that the loss is preserved by $\mathcal{G}$, e.g. an orthogonal representation on $\mathcal{Y}$ and the Euclidean loss, and that (X, Y) satisfy Assumption 1 and Assumption 2. In this setting, we can consider the change to the bound in Theorem 5.2 that arises from taking $\mathcal{H}$ to be $\mathcal{G}$-equivariant. Specifically, Theorem 4.1 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

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

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

6 Related Work

We build on concepts from Eaton (1989); Bloem-Reddy & Teh (2020), which discuss group invariance in statistics and the representation of invariant/equivariant functions.

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

Learning and Generalisation It was noted by Abu-Mostafa (1993) that constraining a model to be invariant cannot increase its VC dimension. A heuristic argument for reduced sample complexity is made by Mroueh et al. (2015) in the case that the input space has finite cardinality. The sample complexity of linear classifiers with invariant representations trained on a simplified image task is discussed briefly by Anselmi et al. (2014), the authors conjecture that a general result may be obtained using wavelet transforms. Sokolic et al. (2017) use the idea of robustness Xu & Mannor (2012) to formulate a generalisation bound for interpolating large-margin classifiers that are invariant to a finite set of transformations, their results contain an implicit margin constraint on the training data. Sannai et al. (2021) consider the generalisation of models invariant or equivariant to finite permutation groups. Each of Lyle et al. (2019; 2020) cover the PAC Bayes approach to generalisation of invariant models, the latter also considers the relative benefits of feature averaging and data augmentation. Recently, Elesedy & Zaidi (2021) proved the first strict benefit in generalisation for equivariant models by taking a function space approach. This was extended to kernel methods in Elesedy (2021). Following a different approach, Mei et al. (2021) study the generalisation of invariant random features and kernel methods by projecting into a space of high degree polynomials. Zhu et al. (2021) study the generalisation benefit of invariance in terms of coverings of the training set. Finally, the concurrent work of Shao et al. (2022) also considers invariance in the PAC framework, but takes a different approach.

ACKNOWLEDGEMENTS

We thank Sheheryar Zaidi for helpful discussion in earlier stages of this project. We thank Varun Kanade and Yee Whye Teh for help and support through this and other projects. We also thank Akiyoshi Sannai and Benjamin Bloem-Reddy for helpful clarifications of their work. The author receives support from the UK EPSRC CDT in Autonomous Intelligent Machines and Systems (grant reference EP/L015897/1).

REFERENCES

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

Fabio Anselmi, Joel Z. Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised learning of invariant representations in hierarchical architectures, 2014.

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

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

Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018.

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

Morris L Eaton. Group invariance applications in statistics. In *Regional conference series in Probability and Statistics*, pp. i–133. JSTOR, 1989.

Bryn Elesedy. Provably strict generalisation benefit for invariance in kernel methods. *arXiv preprint arXiv:2106.02346*, 2021.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2959–2969. PMLR, 2021. URL http://proceedings.mlr.press/v139/elesedy21a.html.

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

David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

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

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*, pp. 2747–2755, 2018.

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.

Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks, 2020.

Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the universality of invariant networks. In *International Conference on Machine Learning*, pp. 4363–4371, 2019.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. *arXiv preprint arXiv:2102.13219*, 2021.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT Press, 2 edition, 2018.

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

David Pfau, James S. Spencer, Alexander G. de G. Matthews, and W. M. C. Foulkes. Spectral inference networks: Unifying spectral methods with deep learning. *arXiv preprint arXiv:1909.02487*, 2019.

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

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*, pp. 771–780. PMLR, 2021.

Bernhard Schölkopf, Chris Burges, and Vladimir Vapnik. Incorporating invariances in support vector learning machines. pp. 47–52. Springer, 1996.

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)*, volume 2, pp. 387–390 vol.2, 1994.

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

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Han Shao, Omar Montasser, and Avrim Blum. A theory of pac learnability under transformation invariances. *arXiv preprint arXiv:2202.07552*, 2022.

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

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pp. 11–30. Springer, 2015.

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

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

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

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in neural information processing systems*, pp. 3391–3401, 2017.

Sicheng Zhu, Bang An, and Furong Huang. Understanding the generalization benefit of model invariance from a data perspective. *Advances in Neural Information Processing Systems*, 34, 2021.

A Discussion of Assumptions

First, a remark on Assumption 1. The distribution ν and the choice of cross-section are not independent. In particular, under the map $\mathcal{X}_\pi \mapsto g\mathcal{X}_\pi$ we have $\nu \mapsto \nu \circ g^{-1}$. The dependence goes away only if $\nu = \nu \circ g \ \forall g \in \mathcal{G}$, which by uniqueness means $\nu = \lambda$, the Haar measure on $\mathcal{G}$. This fact is benign, but worth being aware of.

We now justify Assumption 2. First note that this assumption is much weaker than, for instance, the standard (equivariant) regression setup in which $Y = f^*(X) + \xi$ where $\xi \perp\!\!\!\perp X$, ξ is $\mathcal{G}$-invariant and f^* is equivariant. In particular, Assumption 2 allows for noise corruption that is not necessarily additive. It is known that, under certain conditions, a functional representation for Y is equivalent to the conditional equivariance of Y, in the sense that $gY|gX \overset{\text{d}}{=} Y|X \ \forall g \in \mathcal{G}$ (Bloem-Reddy & Teh, 2020). We describe this setting now.

Assume that the marginal for X is $\mathcal{G}$-invariant, so $X \overset{\text{d}}{=} gX \ \forall g \in \mathcal{G}$. We adapt the following definition from Bloem-Reddy & Teh (2020).

Definition (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 & Teh (2020) then gives us Assumption 2. This result is a form of *noise-outsourcing* or *transfer* (Kallenberg, 2006, Theorem 6.10).

Theorem (Bloem-Reddy & Teh (2020, 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 $\tau : \mathcal{X} \to \mathcal{G}$ (c.f. Appendix A). Let X be a $\mathcal{G}$-invariant random element of $\mathcal{X}$. Then Y is conditionally $\mathcal{G}$-equivariant if and only if $\exists$ 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 $\eta \perp\!\!\!\perp X$.

As it happens, if we assume that the marginal for X is $\mathcal{G}$-invariant, so $X \overset{\text{d}}{=} gX \; \forall g \in \mathcal{G}$, then Eaton (1989, Theorem 4.4) says that $\exists U \sim \lambda$ such that $X = UX_\pi$ and $U \perp\!\!\!\perp X_\pi$. So this setup is sufficient for Assumption 1 too.

B Proofs

B.1 Proof of Proposition 3.1

Proposition (Proposition 3.1)**.** Let $\mathcal{G}$ act measurably on $\mathcal{X}$ and trivially on $\mathcal{Y}$. Let $\mathcal{X}_\pi$ be a measurable cross-section of $\mathcal{X}$ and let π be its projection. Let $\mathcal{H}$ be a hypothesis class of $\mathcal{G}$-invariant functions, let $\texttt{alg} : \cup_{i\in\mathbb{N}} \mathcal{Z}^i \to \mathcal{H}$ be a $\mathcal{G}$-invariant algorithm and let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be and integrable loss function. Then (X, Y, ℓ) and (X_π, Y, ℓ) are $(\mathcal{H}, \mathcal{G})$-equivalent, where $X_\pi = \pi(X)$ with π as in Definition 2.3.

Proof. For any training set S, define $S_\pi = \{(\pi(x), y) : (x, y) \in S\}$. Clearly, if $S \sim (X, Y)^n$ then $S_\pi \sim (X_\pi, Y)^n$. Set $h_S = \texttt{alg}(S)$ and $h_{S_\pi} = \texttt{alg}(S_\pi)$, which with the invariance of $\texttt{alg}$ implies $h_S = h_{S_\pi}$. By invariance, $h(X) = h(X_\pi)$ for any $h \in \mathcal{H}$. Together, this means that

$$\mathbb{E}[\ell(h_S(X), Y)] = \mathbb{E}[\ell(h_{S_\pi}(X_\pi), Y)].$$

We can then conclude that

$$\mathbb{P}\left(\mathbb{E}[\ell(h_S(X), Y)] \leq \inf_{h\in\mathcal{H}} \mathbb{E}[\ell(h(X), Y)] + \epsilon\right) = \mathbb{P}\left(\mathbb{E}[\ell(h_{S_\pi}(X_\pi), Y)] \leq \inf_{h\in\mathcal{H}} \mathbb{E}[\ell(h(X_\pi), Y)] + \epsilon\right)$$

which completes the proof. In the last line, the probabilities are taken over the respective training samples. □

B.2 Proof of Theorem 4.1

Theorem (Theorem 4.1)**.** Let $\mathcal{G}$ act measurably on both $\mathcal{X}$ and $\mathcal{Y}$. Let $\mathcal{H}$ be a hypothesis class of $\mathcal{G}$-equivariant functions and let $\texttt{alg} : \cup_{i\in\mathbb{N}} \mathcal{Z}^i \to \mathcal{H}$ be a $\mathcal{G}$-invariant algorithm. Let $\mathcal{X}_\pi$ be a measurable cross-section of $\mathcal{X}$ and let π be its projection. Let (X, Y) satisfy Assumption 1 and Assumption 2. Then the tasks (X, Y, ℓ) and $(X_\pi, Y_\pi, \bar{\ell})$ are $(\mathcal{H}, \mathcal{G})$-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 the training sets be $S \sim (X, Y)^n$ and $S_\pi \sim (X_\pi, Y_\pi)^n$. We have from Assumption 1 that $X \overset{\text{d}}{=} GX_\pi$ and from Assumption 2 that

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

So by the invariance of a we have $h_S \coloneqq \texttt{alg}(S) = \texttt{alg}(S_\pi) =: h_{S_\pi}$ and hence $\mathbb{E}[\ell(h_S(X), Y)] = \mathbb{E}[\ell(h_{S_\pi}(X), Y)]$. Then, for any equivariant hypothesis $h \in \mathcal{H}$,

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

where we applied Fubini's theorem (Kallenberg, 2006, Theorem 1.27) and $\bar{\ell}(y, y') = \int_{\mathcal{G}} \ell(gy, gy') \, \mathrm{d}\nu(g)$. Note that the function $\ell_{y,y'}(g) = \ell(gy, gy')$ is measurable in g by Kallenberg (2006, Lemma 1.26), so $\bar{\ell}$ exists. Furthermore, the right hand side of the above must be finite because ℓ is integrable (c.f. the left hand side). Putting everything together gives

$$\mathbb{P}\left(\mathbb{E}[\ell(h_S(X), Y)] \leq \inf_{h\in\mathcal{H}} \mathbb{E}[\ell(h(X), Y)] + \epsilon\right) = \mathbb{P}\left(\mathbb{E}[\bar{\ell}(h_{S_\pi}(X_\pi), Y_\pi)] \leq \inf_{h\in\mathcal{H}} \mathbb{E}[\bar{\ell}(h(X_\pi), Y_\pi)] + \epsilon\right),$$

which is exactly what's required. In the last line, the probabilities are taken over the respective training samples. □

B.3 Proof of Theorem 5.2

Theorem (Theorem 5.2)**.** Let $\mathcal{X}$ be a closed subset of a metric space (T, ρ). Let $\mathcal{Y} = B_d(r)$ be the closed Euclidean ball of radius r in $\mathbb{R}^d$. Let $\mathcal{H}$ be a class of functions $h : \mathcal{X} \to \mathcal{Y}$ that are C-Lipschitz in the sense that $\|h(x) - h(x')\|_\infty \leq C\rho(x, x')\ \forall x, x' \in \mathcal{X}$. Let $S = ((X_1, Y_1), \ldots, (X_n, Y_n)) \sim (X, Y)^n$ be a training sequence drawn i.i.d. for some random variables X, Y. Then, for any $\epsilon \in (0, 1)$,

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\mathbb{E}[\ell(h(X), Y)] - \frac{1}{n}\sum_{i=1}^{n}\ell(h(X_i), Y_i)\right| \geq \epsilon\right) \leq 2 \inf_{\alpha\in(0,1)} D_{\alpha\epsilon}(\mathcal{X}, \mathcal{H}) \exp(-8^{-1}(1-\alpha)^2 n\epsilon^2 r^{-4})$$

where

$$D_\tau(\mathcal{X}, \mathcal{H}) = \texttt{Cov}\left(\mathcal{X}, \rho, \frac{\tau}{32Crd}\right) \sup_{x\in\mathcal{X}} \texttt{Cov}\left(\mathcal{H}(x), \|\cdot\|_\infty, \frac{\tau}{16rd}\right).$$

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 (Sokolic et al., 2017; Sannai et al., 2021).

Proof. Let E be a minimal δ-cover for $\mathcal{X}$. We construct a cover for $\mathcal{H}$ in terms of E. Take $h \in \mathcal{H}$ and let $h' : \mathcal{X} \to \mathcal{Y}$ be C-Lipschitz. Since $\mathcal{X}$ is closed, there is an $x \in \mathcal{X}$ such that $\|h - h'\|_{L_\infty} = \|h(x) - h'(x)\|_\infty$. Let x' be an element of E such that $d(x, x') \leq \delta$, then

$$\begin{aligned}
\|h - h'\|_{L_\infty} &= \|h(x) - h'(x)\|_\infty \\
&\leq \|h(x) - h(x')\|_\infty + \|h'(x) - h'(x')\|_\infty + \|h(x') - h'(x')\|_\infty \\
&\leq 2\|x - x'\|_\infty + \|h(x') - h'(x')\|_\infty \\
&\leq 2C\delta + \|h(x') - h'(x')\|_\infty.
\end{aligned}$$

Define $\mathcal{H}(x) = \{h(x) : x \in \mathcal{H}\} \subset \mathcal{Y}$. For any $x \in E$ let K_x be a minimal κ-covering of $\mathcal{H}(x)$ in the vector norm $\|\cdot\|_\infty$. Let F_x be the collection of constant functions $\mathcal{X} \to \mathcal{Y}$ taking the values on K_x

$$F_x = \{f_t(\cdot) = t :\ t \in K_x\}.$$

This means that the set $\{f(x) : x \in E,\ f \in F_x\}$ is a minimal κ-cover for $\mathcal{H}(x)$ in $\|\cdot\|_\infty$. Then the above derivations mean that $\bigcup_{x\in E} F_x$ is a minimal $(2C\delta + \kappa)$-cover for $\mathcal{H}$ in $\|\cdot\|_{L_\infty}$. Thus,

$$\begin{aligned}
\texttt{Cov}(\mathcal{H}, \|\cdot\|_{L_\infty}, 2C\delta + \kappa) &= \left|\bigcup_{x\in E} F_x\right| \\
&\leq \sum_{x\in E} |F_x| \\
&= \sum_{x\in E} \texttt{Cov}(\mathcal{H}(x), \|\cdot\|_\infty, \kappa) \\
&\leq \texttt{Cov}(\mathcal{X}, \rho, \delta) \sup_{x\in\mathcal{X}} \texttt{Cov}(\mathcal{H}(x), \|\cdot\|_\infty, \kappa).
\end{aligned}$$

Then for $\alpha \in (0, 1)$ set $\delta = \frac{\alpha\epsilon}{32Crd}$ and $\kappa = \frac{\alpha\epsilon}{16rd}$ and apply Proposition C.3 to get the conclusion. □

C Additional Results

We need an additional definition.

Definition C.1 (Packing, packing number)**.** Let (T, d) be a pseudo-metric space and $\epsilon > 0$. $E \subset T$ is an ϵ-packing of T (with respect to d) if $\forall x, y \in E$, $d(x, y) > \epsilon$. The ϵ-packing number is the largest cardinality of all the ϵ-packings, i.e. $\texttt{Pack}(T, d, \epsilon) = \sup_{E\in\mathcal{E}} |E|$ where $\mathcal{E}$ is the set of all ϵ-packings. If this supremum doesn't exist, then we say the packing number is infinite.

The following is a straightforward corollary of Proposition 3.1.

Corollary C.2 (Lipschitz classifiers with margin)**.** Let $\mathcal{X}$ be a compact set and let $\mathcal{F}$ be a class of L-Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$. The hypothesis class $\mathcal{H}$ is built by thresholding functions from $\mathcal{F}$, e.g. by $\mathcal{H} = \{h = \text{sign}(f) : f \in \mathcal{F}\}$, and we insist that the output classifier attains margin γ on any training set. Then for any points $x, x' \in \mathcal{X}$ with different labels we must have $\gamma \leq |f(x) - f(x')| \leq L\|x - x'\|$ which means that $\text{VC}(\mathcal{H}) \leq \texttt{Pack}(\mathcal{X}, \|\cdot\|, \gamma/L)$. Alternatively, if $\mathcal{H}$ were invariant then we know from Proposition 3.1 that $\text{VC}(\mathcal{H}) \leq \texttt{Pack}(\mathcal{X}_\pi, \|\cdot\|, \gamma/L) \leq \texttt{Pack}(\mathcal{X}, \|\cdot\|, \gamma/L)$, suggesting a distribution independent sample complexity improvement for invariant hypotheses.

C.1 Concentration of Measure

The following is an adaptation of a standard result.

Proposition C.3. Let $(\mathcal{X}, \rho)$ be a metric space, Let $\mathcal{Y} = B_d(r)$ be the closed Euclidean ball of radius r in $\mathbb{R}^d$. Let $\mathcal{H}$ be a class of functions $h : \mathcal{X} \to \mathcal{Y}$. Let $\ell(y, y') = \|y - y'\|_2^2$ be the squared-error loss and let $S = ((X_1, Y_1), \ldots, (X_n, Y_n)) \sim (X, Y)^n$ be a training sequence drawn i.i.d. for some random variables X, Y. Then,

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\mathbb{E}[\ell(h(X),Y)] - \frac{1}{n}\sum_{i=1}^{n}\ell(h(X_i),Y_i)\right| \geq \epsilon\right) \leq 2 \inf_{\alpha\in(0,1)} \mathtt{Cov}\left(\mathcal{H}, \|\cdot\|_{L_\infty}, \frac{\alpha\epsilon}{8rd}\right) \mathrm{e}^{-\frac{1}{8}(1-\alpha)^2 n\epsilon^2 r^{-4}}.$$

Proof. The proof is based on Mohri et al. (2018, Exercise 3.31). For any probability measure μ on $\mathcal{X} \times \mathcal{Y}$ define

$$R_\mu(h) = \mathbb{E}[\ell(h(A), B)] = \mathbb{E}[\|h(A) - B\|_2^2]$$

where $(A, B) \sim \mu$. Then for any $h, h' \in \mathcal{H}$

$$\begin{aligned}
R_\mu(h) - R_\mu(h') &= \mathbb{E}[\sum_{j=1}^{d}(h(A)_j - B_j)^2 - (h'(A)_j - B_j)^2] \\
&= \mathbb{E}[\sum_{j=1}^{d}(h(A)_j - h'(A)_j)(h(A)_j - B_j + h'(A)_j - B_j)] \\
&\leq \mathbb{E}[|h(A)_j - h'(A)_j|\,(|h(A)_j - B_j| + |h'(A)_j - B_j|)] \\
&\leq 4rd\|h - h'\|_{L_\infty}.
\end{aligned}$$

Now let

$$L_S(h) = \frac{1}{n}\sum_{i=1}^{n}\|h(X_i) - Y_i\|_2^2 - \mathbb{E}[\|h(X) - Y\|_2^2].$$

By setting μ as the empirical measure on S and then as the distribution of (X, Y), one finds that

$$|L_S(h) - L_S(h')| \leq 8rd\|h - h'\|_{L_\infty}.$$

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

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}|L_S(h)| \geq \epsilon\right) = \mathbb{P}\left(\bigcup_{k\in\mathcal{K}}\sup_{h\in D(k)}|L_S(h)| \geq \epsilon\right) \leq \sum_{k\in\mathcal{K}}\mathbb{P}\left(\sup_{h\in D(k)}|L_S(h)| \geq \epsilon\right).$$

Set $\kappa = \frac{\alpha\epsilon}{8rd}$ for $0 < \alpha < 1$. Using the above, for any $h \in D(k)$ we have

$$L_S(h) \leq 8rd\kappa + L_S(k) = \alpha\epsilon + L_S(k),$$

hence

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

Then Hoeffding's Inequality gives

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}|L_S(h)| \geq \epsilon\right) \leq 2\,|\mathcal{K}|\,\exp\left(-\frac{2(1-\alpha)^2 n\epsilon^2}{16r^4}\right)$$

where we used $\|h(X_i) - Y_i\|_2^2 \leq 4r^2$. □