Provably Strict Generalisation Benefit for Invariance in Kernel Methods

Anonymous Author(s)
Affiliation
Address
email

Abstract

It is a commonly held belief that enforcing invariance improves generalisation. Although this approach enjoys widespread popularity, it is only very recently that a rigorous theoretical demonstration of this benefit has been established. In this work we build on the function space perspective of Elesedy and Zaidi [8] to derive a strictly non-zero generalisation benefit of incorporating invariance in kernel ridge regression when the target is invariant to the action of a compact group. We study invariance enforced by feature averaging and find that generalisation is governed by a notion of effective dimension that arises from the interplay between the kernel and the group. In building towards this result, we find that the action of the group induces an orthogonal decomposition of both the reproducing kernel Hilbert space and its kernel, which may be of interest in its own right.

1 Introduction

Recently, there has been significant interest in models that are invariant to the action of a group on their inputs. It is believed that engineering models in this way improves sample efficiency and generalisation. Intuitively, if a task has an invariance, then a model that is constructed to be invariant ahead of time should require fewer examples to generalise than one that must learn to be invariant. Indeed, there are many application domains, such as fundamental physics or medical imaging, in which the invariance is known a priori [28, 30]. Although this intuition is certainly not new (e.g. [31]), it has inspired much recent work (for instance, see [34, 15]).

However, while implementations and practical applications abound, until very recently a rigorous theoretical justification for invariance was missing. As pointed out in [8], many prior works such as [27, 23] provide only worst-case guarantees on the performance of invariant algorithms. It follows that these results do not rule out the possibility of modern training algorithms automatically favouring invariant models, irrespective of the choice of architecture. Steps towards a more concrete theory of the benefit of invariance have been taken by [8, 20] and our work is a continuation along the path set by [8].

In this work we provide a precise characterisation of the generalisation benefit of invariance in kernel ridge regression. In contrast to [27, 23], this proves a *provably strict* generalisation benefit for invariant, feature-averaged models. In deriving this result, we provide insights into the structure of reproducing kernel Hilbert spaces in relation to invariant functions that we believe will be useful for analysing invariance in other kernel algorithms.

The use of feature averaging to produce invariant predictors enjoys both theoretical and practical success [17, 9]. For the purposes of this work, feature averaging is defined as training a model as normal (according to any algorithm) and then transforming the learned model to be invariant. This transformation is done by *orbit-averaging*, which means projecting the model on the space of invariant functions using the operator $\mathcal{O}$ introduced in Section 2.3.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Kernel methods have a long been a mainstay of machine learning (see [29, Section 4.7] for a brief historical overview). Kernels can be viewed as mapping the input data into a potentially infinite dimensional feature space, which allows for analytically tractable inference with non-linear predictors. While modern machine learning practice is dominated by neural networks, kernels remain at the core of much of modern theory. The most notable instance of this is the theory surrounding the *neural tangent kernel* [11], which states that the functions realised by an infinitely wide neural network belong to an RKHS with a kernel determined by the network architecture. This relation has led to many results on the theory of optimisation and generalisation of wide neural networks (e.g. [14, 3]).

1.1 Summary of Contributions

This paper builds towards Theorem 5 in Section 4, which gives a precise characterisation of the benefit of incorporating invariance in kernel methods by feature averaging. We find that for a predictor from a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with kernel k, the benefit is $O(\dim_{\text{eff}}(\mathcal{H}_A)/n)$ where $\dim_{\text{eff}}(\mathcal{H}_A)$ is a notion of effective dimension of an RKHS $\mathcal{H}_A \subset \mathcal{H}$ that arises from the interaction between the kernel k and the group. Lemma 3, given in Section 3, forms the basis of Theorem 5 and shows that $\mathcal{H}$ decomposes into an orthogonal direct sum $\mathcal{H} = \mathcal{H}_S \oplus \mathcal{H}_A$, where $\mathcal{H}_S$ is an RKHS consisting of all of the invariant functions in $\mathcal{H}$. We stress that while Theorem 5 is specialised to kernel ridge regression, Lemma 3 holds regardless of training algorithm and could be used to explore invariance in other kernel methods. In Section 2 we outline our assumptions and the ideas from [8] on which we build. We discuss related works in Section 5.

2 Background and Preliminaries

In this section we provide a brief introduction to reproducing kernel Hilbert spaces (RKHS) and the ideas we borrow from Elesedy and Zaidi [8]. Throughout this paper, $\mathcal{H}$ with be an RKHS with kernel k. In Section 2.2 we state some topological and measurability assumptions that are needed for our proofs. These conditions are benign, and the reader not interested in technicalities need take from Section 2.2 only that μ is $\mathcal{G}$-invariant and that the kernel k is bounded and satisfies Eq. (1). We defer some background results to Appendix A of the Supplementary Material.

2.1 RKHS Basics

A Hilbert space is an inner product space that is complete with respect to the norm topology induced by the inner product. A reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ is 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 $\forall x \in \mathcal{X}$, or, equivalently is a bounded operator. The Reisz Representation Theorem tells us that there is a unique function $k_x \in \mathcal{H}$ such that $\delta_x[f] = \langle k_x, f\rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$, where $\langle \cdot, \cdot\rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is the inner product on $\mathcal{H}$. We 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 shows that for any positive-definite and symmetric kernel k, there is a unique RKHS with reproducing kernel k. In addition, any Hilbert space admitting a reproducing kernel is an RKHS. Finally, another characterisation of $\mathcal{H}$ is as the completion of linear combinations of the form $f_c(x) = \sum_{i=1}^n c_i k(x, x_i)$ for $c_1, \dots, c_n \in \mathbb{R}$ and $x_1, \dots, x_n \in \mathcal{X}$. For (many) more details, see [29, Chapter 4].

2.2 Technical Setup and Assumptions

Let $\mathcal{G}$ be a compact group with Haar measure λ. Let $\mathcal{X}$ be a non-empty Polish space admitting a finite, $\mathcal{G}$-invariant Borel measure μ, with $\operatorname{supp}\mu = \mathcal{X}$. We normalise $\mu(\mathcal{X}) = \lambda(\mathcal{G}) = 1$, the latter is possible because λ is a Radon measure. We assume that $\mathcal{G}$ has a measurable action on $\mathcal{X}$ that we will write as gx for $g \in \mathcal{G}$, $x \in \mathcal{X}$. A function $f : \mathcal{X} \to \mathbb{R}$ is $\mathcal{G}$-invariant if $f(gx) = f(x)\ \forall x \in \mathcal{X}$ $\forall g \in \mathcal{G}$. Similarly, a measure μ on $\mathcal{X}$ is $\mathcal{G}$-invariant if $\forall g \in \mathcal{G}$ and any μ-measurable $B \subset \mathcal{X}$ the pushforward of μ by the action of $\mathcal{G}$ equals μ, i.e. $(g_*\mu)(B) = \mu(B)$. This means that if $X \sim \mu$ then $gX \sim \mu\ \forall g \in \mathcal{G}$.

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 any $x \in \mathcal{X}$. Assume that $\sup_{x\in\mathcal{X}} k(x, x) = M_k < \infty$ and note that this implies that k is bounded since

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

Every $f \in \mathcal{H}$ is μ-measurable, bounded and continuous by [29, Lemmas 4.24 and 4.28] and in addition $\mathcal{H}$ is separable using [29, Lemma 4.33]. These conditions allow the application of [29, Theorem 4.26] to relate $\mathcal{H}$ to $L_2(\mathcal{X}, \mu)$ in the proofs building towards Lemma 3. We assume that the kernel satisfies, for all $x, y \in \mathcal{X}$,

$$\int_{\mathcal{G}} k(gx, y)\,\mathrm{d}\lambda(g) = \int_{\mathcal{G}} k(x, gy)\,\mathrm{d}\lambda(g). \tag{1}$$

For this it is sufficient to have $k(gx, y)$ equal to $k(x, gy)$ or $k(x, g^{-1}y)$ (the latter using unimodularity of $\mathcal{G}$). Highlighting two special cases: any inner product kernel $k(x, x') = \kappa(\langle x, x' \rangle)$ such that the action of $\mathcal{G}$ is unitary with respect to $\langle \cdot, \cdot \rangle$ satisfies Eq. (1), as does any stationary kernel $k(x, x') = \kappa(\|x - x'\|)$ with norm that is preserved by $\mathcal{G}$ in the sense that $\|gx - gx'\| = \|x - x'\|$ for any $g \in \mathcal{G}$, $x, x' \in \mathcal{X}$.

2.3 Invariance from a Function Space Perspective

Given a function $f : \mathcal{X} \to \mathbb{R}$ we can define a corresponding orbit-averaged function $\mathcal{O}f : \mathcal{X} \to \mathbb{R}$ with values

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

$\mathcal{O}f$ will exist whenever f is μ-measurable. Note that $\mathcal{O}$ is a linear operator and $\mathcal{O}f$ is always $\mathcal{G}$-invariant. Interestingly, f is $\mathcal{G}$-invariant *only* if $f = \mathcal{O}f$. Elesedy and Zaidi [8] use these observations to characterise invariant functions and study their generalisation properties. In short, this work extends these insights to kernel methods. Along the way, we will make frequent use of the following (well known) facts about $\mathcal{O}$.

Lemma 1 ([8, Propositions 23 and 24])**.** A function f is $\mathcal{G}$-invariant if and only if $\mathcal{O}f = f$. This implies that $\mathcal{O}$ is idempotent, so can have only two eigenvalues 0 and 1.

Lemma 2 ([8, Lemma 1])**.** $\mathcal{O} : L_2(\mathcal{X}, \mu) \to L_2(\mathcal{X}, \mu)$ is well-defined and self-adjoint. Hence, $L_2(\mathcal{X}, \mu)$ has the orthogonal decomposition

$$L_2(\mathcal{X}, \mu) = S \oplus A$$

where $S = \{f \in L_2(\mathcal{X}, \mu) : f \text{ is } \mathcal{G} \text{ invariant}\}$ and $A = \{f \in L_2(\mathcal{X}, \mu) : \mathcal{O}f = 0\}$.

In [8], S and A are referred to as the symmetric and anti-symmetric parts of $L_2(\mathcal{X}, \mu)$. We will use the same terminology.

The meaning of Lemma 2 is that any $f \in L_2(\mathcal{X}, \mu)$ has a decomposition $f = \bar{f} + f^{\perp}$ where $\bar{f} = \mathcal{O}f$ is $\mathcal{G}$-invariant and $\mathcal{O}f^{\perp} = 0$. A noteworthy consequence of this setup, as discussed in [8], is a provably non-negative generalisation benefit for feature averaging. In particular, for any predictor $f \in L_2(\mathcal{X}, \mu)$, if the target $f^* \in L_2(\mathcal{X}, \mu)$ is $\mathcal{G}$-invariant then the test error $R(f) = \mathbb{E}_{X \sim \mu}[(f(X) - f^*(X))^2]$ satisfies

$$R(f) - R(\bar{f}) = \|f^{\perp}\|^2_{L_2(\mathcal{X}, \mu)} \geq 0.$$

The same holds if the target is corrupted by independent, zero mean (additive) noise.

3 Induced Structure of $\mathcal{H}$

In this section we present Lemma 3, which is an analog of Lemma 2 for RKHSs. Lemma 3 shows that for any compact group $\mathcal{G}$ and RKHS $\mathcal{H}$, if the kernel for $\mathcal{H}$ satisfies the assumptions in Section 2.2, then $\mathcal{H}$ can be viewed as being built from two orthogonal RKHSs, one consisting of invariant functions and another of those that vanish when averaged over $\mathcal{G}$. Later in the paper, this decomposition will allow us to analyse the generalisation benefit of invariant predictors.

It may seem at first glance that Lemma 3 should follow immediately from Lemma 2, but this is not the case. First, it is not obvious that for any $f \in \mathcal{H}$, its orbit averaged version $\mathcal{O}f$ is also in $\mathcal{H}$. Moreover, in contrast with $L_2(\mathcal{X}, \mu)$, an explicit form for the inner product on $\mathcal{H}$ is not immediate, which means that some work is needed to check that $\mathcal{O}$ is self-adjoint on $\mathcal{H}$. These are important requirements for the proofs of both Lemmas 2 and 3 and we establish them, along with $\mathcal{O}$ being continuous on $\mathcal{H}$, in the Supplementary Material. The assumption that the kernel satisfies Eq. (1) plays a central role.

Lemma 3. $\mathcal{H}$ admits an orthogonal decomposition into symmetric and anti-symmetric parts

$$\mathcal{H} = \mathcal{H}_S \oplus \mathcal{H}_A$$

where $\mathcal{H}_S = \{f \in \mathcal{H} : f \text{ is } \mathcal{G}\text{-invariant}\}$ and $\mathcal{H}_A = \{f \in \mathcal{H} : \mathcal{O}f = 0\}$. Moreover, $\mathcal{H}_S$ is an RKHS with kernel

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

and $\mathcal{H}_A$ is an RKHS with kernel

$$k^{\perp}(x, y) = k(x, y) - \bar{k}(x, y).$$

Finally, $\bar{k}$ is $\mathcal{G}$-invariant in both arguments.

Proof. From Lemma 1 we know that $\mathcal{O}$ is a projection operator. Since it is self-adjoint, $\mathcal{O}$ is even an orthogonal projection on $\mathcal{H}$: 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 any $f \in \mathcal{H}$ we can write $f = \bar{f} + f^{\perp}$ where $\bar{f} = \mathcal{O}f \in \mathcal{H}_S$ is $\mathcal{G}$-invariant and $f^{\perp} = f - \mathcal{O}f \in \mathcal{H}_A$ and these terms are mutually orthogonal.

By the linearity of $\mathcal{O}$, it is clear that $\mathcal{H}_S = \mathcal{O}\mathcal{H}$ is an inner product space. It is easy to show that $\mathcal{O}$ being continuous implies $\mathcal{H}_S$ is complete. Thus $\mathcal{H}_S$ is a Hilbert space, and an RKHS since the evaluation functional is clearly continuous on $\mathcal{H}_S \subset \mathcal{H}$. For any $h_S \in \mathcal{H}_S$ we have

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

and the uniqueness afforded by the Reisz representation theorem tells us that the reproducing kernel for $\mathcal{H}_S$ is $\bar{k}(x, y) = \int_{\mathcal{G}} k(x, gy) \,\mathrm{d}\lambda(g)$. We have $\|\mathrm{id} - \mathcal{O}\| \leq 2$ and we can do the same argument to show that $\mathcal{H}_A$ is an RKHS with reproducing kernel $k^{\perp}$ as claimed. Note that one can write $k^{\perp}(x, y) = \langle k^{\perp}_x, k^{\perp}_y \rangle_{\mathcal{H}}$ so it must be positive-definite. The $\mathcal{G}$-invariance of $\bar{k}(x, y)$ in both arguments is immediate from Eq. (1) and Lemma 1. □

As stated earlier, the perspective provided by Lemma 3 will support our analysis of generalisation. Just as with Lemma 2, Lemma 3 says that any $f \in \mathcal{H}$ can be written as $f = \bar{f} + f^{\perp}$ where $\bar{f}$ is $\mathcal{G}$-invariant and $\mathcal{O}f^{\perp} = 0$ with $\langle \bar{f}, f^{\perp} \rangle_{\mathcal{H}} = 0$. As an aside, $\bar{k}$ happens to qualify as a *Haar Integration Kernel*, a concept introduced by Haasdonk, Vossen, and Burkhardt [10]. We will see that a notion of effective dimension of the RKHS $\mathcal{H}_A$ with kernel $k^{\perp}$ governs the generalisation gap between an arbitrary predictor f and its invariant version $\mathcal{O}f$. This effective dimension arises from the spectral theory of an integral operator related to k, which we develop in the next section.

3.1 Spectral Representation and Effective Dimension

In this section we consider the spectrum of an integral operator related to the kernel k. This analysis will ultimately allow us to define a notion of effective dimension of $\mathcal{H}_A$ that we will later see is important to the generalisation of invariant predictors. While the integral operator setup is standard, the use of this technique to identify an effective dimension of $\mathcal{H}_A$ is novel.

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

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

One way of viewing things is that S_k assigns to every element in $L_2(\mathcal{X}, \mu)$ a function in $\mathcal{H}$. On the other hand, every $f \in \mathcal{H}$ is bounded so has $\|f\|_{L_2(\mathcal{X},\mu)} < \infty$ and belongs to some element of $L_2(\mathcal{X}, \mu)$. We write $\iota : \mathcal{H} \to L_2(\mathcal{X}, \mu)$ for the *inclusion map* that sends f to the element of $L_2(\mathcal{X}, \mu)$ that contains f. In the Supplementary Material we show that ι is injective, so any element of $L_2(\mathcal{X}, \mu)$ contains at most one $f \in \mathcal{H}$.

One can define $T_k : L_2(\mathcal{X}, \mu) \to L_2(\mathcal{X}, \mu)$ by $T_k = \iota \circ S_k$, and [29, Theorem 4.27] says that T_k is compact, positive, self-adjoint and trace-class. In addition, $L_2(\mathcal{X}, \mu)$ is separable by [7, Proposition 3.4.5], because $\mathcal{X}$ is Polish and μ is a Borel measure, so has a countable orthonormal basis. Hence, by the Spectral Theorem, there exists a countable orthonormal basis $\{\tilde{e}_i\}$ for $L_2(\mathcal{X}, \mu)$ such that $T_k \tilde{e}_i = \lambda_i \tilde{e}_i$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are the eigenvalues of T_k. Moreover, since ι is injective, for each of the $\tilde{e}_i$ for which $\lambda_i > 0$ there is a unique $e_i \in \mathcal{H}$ such that $\iota e_i = \tilde{e}_i$ and $S_k \tilde{e}_i = \lambda_i e_i$.

Now, since $\iota k_x \in L_2(\mathcal{X},\mu)$ we have

$$\iota k_x = \sum_i \langle \iota k_x, \tilde{e}_i\rangle_{L_2(\mathcal{X},\mu)}\tilde{e}_i = \sum_i (S_k \tilde{e}_i)(x)\tilde{e}_i = \sum_i \lambda_i e_i(x)\tilde{e}_i. \quad (2)$$

From now on we permit ourself to drop the ι to reduce clutter. We use the above to define

$$j(x,y) = \langle k_x, k_y\rangle_{L_2(\mathcal{X},\mu)}, \quad \bar{j}(x,y) = \langle \bar{k}_x, \bar{k}_y\rangle_{L_2(\mathcal{X},\mu)} \quad \text{and} \quad j^\perp(x,y) = \langle k_x^\perp, k_y^\perp\rangle_{L_2(\mathcal{X},\mu)}.$$

These quantities will appear again in our analysis of the generalisation of invariant kernel methods. Indeed, we will see later in this section that $\mathbb{E}[j^\perp(X,X)]$ is a type of effective dimension of $\mathcal{H}_A$. Following Eq. (2), one finds the series representations given below in Lemma 4.

The reader may have noticed that our setup is very similar to the one provided by Mercer's theorem. However, we do not assume compactness of $\mathcal{X}$ and so (the classical form of) Mercer's Theorem does not apply. In particular, the set $\{e_i\}$ (even when scaled appropriately) need not form an orthonormal basis in $\mathcal{H}$. This aspect of our work is a feature, rather than a bug: the loosening of the compactness condition allows application to common settings such as $\mathcal{X} = \mathbb{R}^n$.

Lemma 4. We have

$$j = \bar{j} + j^\perp.$$

Furthermore, let $\bar{e}_i = \mathcal{O}e_i$ and $e_i^\perp = e_i - \bar{e}_i$ then

$$j(x,y) = \sum_i \lambda_i^2 e_i(x)e_i(y), \quad \bar{j}(x,y) = \sum_i \lambda_i^2 \bar{e}_i(x)\bar{e}_i(y), \quad \text{and} \quad j^\perp(x,y) = \sum_i \lambda_i^2 e_i^\perp(x)e_j^\perp(y).$$

Finally, the function $\sum_i \lambda_i^2 \bar{e}_i \otimes e_i^\perp : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ vanishes everywhere.

Proof. We show in the Supplementary Material that $\mathcal{O}$ and S_k commute on $L_2(\mathcal{X},\mu)$ and $\mathcal{O}$ is self-adjoint on $L_2(\mathcal{X},\mu)$ by Lemma 1, so $\mathcal{O}$ and ι (the adjoint of S_k by [29, Theorem 4.26]) must also commute. The first comment is then immediate from the observation that if $a \in \mathcal{H}_S$ and $b \in \mathcal{H}_A$ one has

$$\langle \iota a, \iota b\rangle_{L_2(\mathcal{X},\mu)} = \langle \iota\mathcal{O}a, \iota b\rangle_{L_2(\mathcal{X},\mu)} = \langle \mathcal{O}\iota a, \iota b\rangle_{L_2(\mathcal{X},\mu)} = \langle \iota a, \iota\mathcal{O}b\rangle_{L_2(\mathcal{X},\mu)} = 0.$$

We also have both of

$$\langle \iota\bar{k}_x, \tilde{e}_i\rangle_{L_2(\mathcal{X},\mu)} = \langle \iota k_x, \mathcal{O}\tilde{e}_i\rangle_{L_2(\mathcal{X},\mu)} = S_k\mathcal{O}\tilde{e}_i = \mathcal{O}S_k\tilde{e}_i = \lambda_i\bar{e}_i$$

and

$$\langle \iota k_x^\perp, \tilde{e}_i\rangle_{L_2(\mathcal{X},\mu)} = \langle \iota k_x, (\mathrm{id}-\mathcal{O})\tilde{e}_i\rangle_{L_2(\mathcal{X},\mu)} = S_k(\mathrm{id}-\mathcal{O})\tilde{e}_i = (\mathrm{id}-\mathcal{O})S_k\tilde{e}_i = \lambda_i e_i^\perp.$$

Therefore $\iota\bar{k}_x = \sum_i \lambda_i \bar{e}_i(x)\tilde{e}_i$ and $\iota k_x^\perp = \sum_i \lambda_i e_i^\perp(x)\tilde{e}_i$. Taking inner products on $L_2(\mathcal{X},\mu)$ gives the remaining results. ☐

Before turning to generalisation, we describe how the above quantities can be used to define a measure effective dimension. We define

$$\dim_{\text{eff}}(\mathcal{H}) = \mathbb{E}[j(X,X)]$$

where $X \sim \mu$. Applying Fubini's theorem, we find

$$\dim_{\text{eff}}(\mathcal{H}) = \sum_i \lambda_i^2\, \mathbb{E}[e_i(X)^2] = \sum_i \lambda_i^2 \|\tilde{e}_i\|^2_{L_2(\mathcal{X},\mu)} = \sum_i \lambda_i^2.$$

The series converges by the comparison test because $\lambda_i \geq 0$ and $\sum_i \lambda_i = \mathrm{Tr}(T_k) < \infty$. We have $\dim_{\text{eff}}(\mathcal{H}) = \mathrm{Tr}(T_k^2)$ and we can think of this (very informally) as taking $L_2(\mathcal{X},\mu)$, pushing it through $\mathcal{H}$ twice using T_k and then measuring its size. Now because $j = \bar{j} + j^\perp$ we get

$$\dim_{\text{eff}}(\mathcal{H}) = \dim_{\text{eff}}(\mathcal{H}_S) + \dim_{\text{eff}}(\mathcal{H}_A)$$

with

$$\dim_{\text{eff}}(\mathcal{H}_A) = \sum_i \lambda_i^2 \|\tilde{e}_i^\perp\|^2_{L_2(\mathcal{X},\mu)} = \mathrm{Tr}(T_k^2) - \mathrm{Tr}((\mathcal{O}T_k)^2)$$

where $\tilde{e}_i^\perp = \iota e_i^\perp$. Again, very informally, this can be thought of as pushing $L_2(\mathcal{X},\mu)$ through $\mathcal{H}_A$ twice and measuring the size of the output. In the next section we will consider the generalisation of kernel ridge regression and find that $\dim_{\text{eff}}(\mathcal{H}_A)$ plays a critical role.

4 Generalisation

In this section we apply the theory developed in Section 3 to study the impact of invariance on kernel ridge regression with an invariant target. We analyse the generalisation benefit of feature averaging, finding a strict benefit when the target is $\mathcal{G}$-invariant.

4.1 Kernel Ridge Regression

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

$$\operatorname*{argmin}_{f \in \mathcal{H}} C(f) \quad \text{where} \quad C(f) = \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \rho \|f\|_{\mathcal{H}}^2 \tag{3}$$

and $\rho > 0$ is the regularisation parameter. KRR can be thought of as performing ridge regression with a possibly infinite dimensional feature space $\mathcal{H}$. The representer theorem tells us that 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

$$\operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \left\{ \|\boldsymbol{Y} - K\alpha\|_2^2 + \rho \alpha^\top K \alpha \right\} \tag{4}$$

and $\boldsymbol{Y} \in \mathbb{R}^n$ is the typical row-stacking of the training outputs with $\boldsymbol{Y}_i = y_i$. K is the typical kernel Gram matrix with $K_{ij} = k(x_i, x_j)$. We consider solutions of the form[1] $\alpha = (K + \rho I)^{-1} \boldsymbol{Y}$ which results in the predictor

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

where $k_x(\boldsymbol{X}) \in \mathbb{R}^n$ is the vector with components $k_x(\boldsymbol{X})_i = k_x(x_i)$. We will compare the generalisation performance of this predictor with that of its averaged version

$$\bar{f} = \bar{k}_x(\boldsymbol{X})^\top (K + \rho I)^{-1} \boldsymbol{Y} \in \mathcal{H}_S.$$

To do this we look at the generalisation gap.

4.2 Generalisation Gap

The generalisation gap is a quantity that compares the expected test performances of two predictors on a given task. Given a distribution $(X, Y) \sim \mathbb{P}$ and loss function l defining a supervised learning task, we define the generalisation gap between two predictors f and f' to be

$$\Delta(f, f') = \mathbb{E}[l(f(X), Y)] - \mathbb{E}[l(f'(X), Y)]$$

where the expectation is conditional on the given realisations of f, f' if the predictors are random. In this paper we consider $l(a, b) = (a - b)^2$ the squared-error loss and we will assume $Y = f^*(X) + \xi$ for some target function f^* where ξ is has mean 0 and is independent of X. In this case, the generalisation gap reduces to

$$\Delta(f, f') = \mathbb{E}[(f(X) - f^*(X))^2] - \mathbb{E}[(f'(X) - f^*(X))^2].$$

Clearly, if $\Delta(f, f') > 0$ then we expect strictly better test performance from f than f'.

4.3 Generalisation Benefit of Feature Averaging

We are now in a position to give our main result, which is a characterisation of the generalisation benefit of invariance in kernel methods. This is in some sense a generalisation of [8, Theorem 6] and we will return to this comparison later. We emphasise that Theorem 5 holds under quite general conditions that cover the majority of practical applications.

Theorem 5. Let the training data be $\{(X_i, Y_i) : i = 1, \ldots, n\}$ with $Y_i = f^*(X_i) + \xi_i$ where $X \sim \mu$, $f^* \in L_2(\mathcal{X}, \mu)$ is $\mathcal{G}$-invariant and $\{\xi_i : i = 1, \ldots, n\}$ are independent, with $\mathbb{E}[\xi_i] = 0$ and $\mathbb{E}[\xi_i^2] = 0$. Let $f = \operatorname{argmin}_{f \in \mathcal{H}} C(f)$ be the solution to Eq. (3) and let $\bar{f} = \mathcal{O} f \in \mathcal{H}_S$ be the result of applying feature averaging to f, then the generalisation gap with the squared-error loss satisfies

$$\mathbb{E}[\Delta(f, \bar{f})] \geq \frac{\sigma^2 \dim_{\text{eff}}(\mathcal{H}_A) + \sum_\alpha \lambda_\alpha^2 \langle (f^*)^2, (\tilde{e}_\alpha^\perp)^2 \rangle_{L_2(\mathcal{X}, \mu)}}{(\sqrt{n} M_k + \rho / \sqrt{n})^2}$$

[1] When K is a positive definite matrix this will be the *only* solution. If K is singular then $\exists c \in \mathbb{R}^n$ with $\sum_{ij} K_{ij} c_i c_j = \|\sum_i c_i k_{x_i}\|_{\mathcal{H}}^2 = 0$ so $\sum_i c_i k_{x_i}$ is identically 0 and $\forall f \in \mathcal{H}$ we get $\sum_i c_i f(x_i) = 0$ (see [18, Section 4.6.2]). Clearly, this can't happen if $\mathcal{H}$ is sufficiently expressive. In any case, the chosen α is the minimum in Euclidean norm of all possible solutions.

where the squares are to be interpreted pointwise as $(h)^2(x) = h(x)^2$ and

$$\dim_{\text{eff}}(\mathcal{H}_A) \coloneqq \text{Tr}(T_k^2) - \text{Tr}((\mathcal{O}T_k)^2) = \mathbb{E}[j^\perp(X,X)] = \sum_\alpha \lambda_\alpha^2 \|\tilde{e}_\alpha^\perp\|^2_{L_2(\mathcal{X},\mu)} \geq 0$$

is the *effective dimension* of $\mathcal{H}_A$.

Proof. Let $J^\perp$ be the Gram matrix with components $J^\perp_{ij} = j^\perp(X_i, X_j)$ let $u \in \mathbb{R}^n$ have components $u_i = f^*(x_i)$. We can use Lemma 2 to get

$$\Delta(f, \bar{f}) = \mathbb{E}[(k_X^\perp(\boldsymbol{X})^\top (K + \rho I)^{-1}\boldsymbol{Y})^2 | \boldsymbol{X}, \boldsymbol{Y}]$$

where $k_x^\perp(\boldsymbol{X}) \in \mathbb{R}^n$ with $k_x^\perp(\boldsymbol{X})_i = k_x^\perp(X_i)$. Let $\boldsymbol{\xi} \in \mathbb{R}^n$ have components $\boldsymbol{\xi}_i = \xi_i$ then one finds

$$\begin{aligned}\mathbb{E}[\Delta(f,\bar{f})|\boldsymbol{X}] &= \mathbb{E}[(k_X^\perp(\boldsymbol{X})^\top (K+\rho I)^{-1}u)^2|\boldsymbol{X}] + \mathbb{E}[(k_X^\perp(\boldsymbol{X})^\top(K+\rho I)^{-1}\boldsymbol{\xi})^2|\boldsymbol{X}] \\ &= u^\top (K+\rho I)^{-1} J^\perp (K+\rho I)^{-1} u + \sigma^2 \,\text{Tr}\left(J^\perp (K+\rho I)^{-2}\right)\end{aligned}$$

where the first equality follows because $\boldsymbol{\xi}$ has mean 0 and the second comes from the trace trick.

Let $\lambda_{\text{min}}(A)$ be the smallest eigenvalue of a matrix A. Consider the first term

$$\begin{aligned}u^\top (K+\rho I)^{-1} J^\perp (K+\rho I)^{-1} u &\geq \lambda_{\text{min}}\left((K+\rho I)^{-1}\right)^2 u^\top J^\perp u \\ &\geq \frac{1}{(M_k n + \rho)^2} u^\top J^\perp u \\ &= \frac{1}{(M_k n + \rho)^2} \sum_{ij} \sum_\alpha \lambda_\alpha^2 f^*(X_i) e_\alpha^\perp(X_i) e_\alpha^\perp(X_j) f^*(X_j).\end{aligned}$$

where the second inequality follows from $\|A\|_{\text{op}} \leq n \max_{ij} A_{ij}$ for $n \times n$ matrix A and the last line comes from Lemma 4. Now $\mathbb{E}[f^*(X)e_\alpha^\perp(X)] = \langle \iota f^*, \tilde{e}_\alpha^\perp \rangle_{L_2(\mathcal{X},\mu)}$ so the above terms vanish in expectation when $i \neq j$ and (temporarily suspending suspicion regarding the infinite sum) it follows that

$$\mathbb{E}[u^\top (K+\rho I)^{-1} J^\perp (K+\rho I)^{-1} u] \geq \frac{n}{(nM_k + \rho)^2} \sum_\alpha \lambda_\alpha^2 \langle \iota(f^*)^2, (\tilde{e}_\alpha^\perp)^2 \rangle_{L_2(\mathcal{X},\mu)}$$

where the squares are interpreted pointwise $(h)^2(x) = h(x)^2$.

Interchanging the sum and the expectation above is valid by Fubini's theorem [12, Theorem 1.27] (interpreting the index α as having the counting measure, which is σ-finite) as long as

$$\sum_\alpha \lambda_\alpha^2 \,\mathbb{E}[f^*(X_i) e_\alpha^\perp(X_i) e_\alpha^\perp(X_j) f^*(X_j)] < \infty \tag{5}$$

for all i, j. When $i \neq j$ we have established that this holds, while when $i = j$ we look at each term individually. We have

$$\mathbb{E}[f^*(X)^2 e_\alpha^\perp(X)^2] \leq \|\tilde{e}_\alpha^\perp\|^2_{L_2(\mathcal{X},\mu)} \operatorname*{ess\,sup}_{x\in\mathcal{X}} f^*(x)^2 \leq \operatorname*{ess\,sup}_{x\in\mathcal{X}} f^*(x)^2 < \infty$$

since

$$\|\mathcal{O}\tilde{e}_\alpha\|^2_{L_2(\mathcal{X},\mu)} + \|\tilde{e}_\alpha^\perp\|^2_{L_2(\mathcal{X},\mu)} = \|\tilde{e}_\alpha\|^2_{L_2(\mathcal{X},\mu)} = 1$$

by Lemma 2. Hence the series in Eq. (5) converges if $\sum_\alpha \lambda_\alpha^2$ converges. Recall from Section 3.1 that T_k is positive and trace-class, so $\sum_\alpha \lambda_\alpha < \infty$ and $\lambda_\alpha \geq 0 \;\forall \alpha$, so $\sum_\alpha \lambda_\alpha^2 < \infty$ by the comparison test and we're safe.

Moving to the second term, we have

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

and then

$$\begin{aligned}\frac{1}{n}\mathbb{E}[\mathrm{Tr}(J^{\perp})] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sum_{\alpha}\lambda_{\alpha}^{2}e_{\alpha}^{\perp}(X_i)e_{\alpha}^{\perp}(X_i)\right]\\ &= \sum_{\alpha}\lambda_{\alpha}^{2}\|\tilde{e}_{\alpha}^{\perp}\|_{L_2(\mathcal{X},\mu)}^{2}\\ &= \sum_{\alpha}\lambda_{\alpha}^{2} - \sum_{\alpha}\lambda_{\alpha}^{2}\|\mathcal{O}\tilde{e}_{\alpha}\|_{L_2(\mathcal{X},\mu)}^{2}\\ &= \mathrm{Tr}(T_k^2) - \mathrm{Tr}(T_k^2\mathcal{O})\end{aligned}$$

with the sums converging and the expectations being okay by virtue of the above. □

Theorem 5 shows that feature averaging is provably beneficial in terms of generalisation if the mean of the target distribution is invariant. One might think that, given enough training examples, the solution f to Eq. (3) would *learn* to be $\mathcal{G}$-invariant. Theorem 5 shows that this cannot happen unless the number of examples dominates the effective dimension of $\mathcal{H}_A$.

The role of $\dim_{\text{eff}}(\mathcal{H}_A)$ mirrors that of $\dim A$ in [8, Theorem 6], where A is $\mathcal{H}_A$ when k is the linear kernel. In this sense, Theorem 5 can be seen as a generalisation of [8, Theorem 6]. It is for this reason that we believe that, although the constant M_k in the denominator is likely not optimal, the $O(1/n)$ rate that matches [8] is tight. We leave a more precise analysis of the constants to future work.

Finally, we must be careful to state that our setting does not directly reduce to that of [8, Theorem 6] for two reasons. First, [8, Theorem 6] considers $\mathcal{G}$ invariant linear models without regularisation. This may turn out to be accessible by a $\rho \to 0^+$ limit (the so called ridgeless limit) of Theorem 5. More importantly, linear regression is equivalent to kernel regression with the linear kernel. However, the linear kernel can be unbounded (e.g. on $\mathbb{R}$), so does not meet our technical conditions in Section 2.2. We therefore conjecture that the boundedness assumption on k can be removed.

5 Related Work

Incorporating invariance into machine learning models is not a new idea. The majority of modern applications concern neural networks, but earlier work has used kernels [10], support vector machines [24] and polynomial feature spaces [25, 26]. Indeed, early work also considered invariant neural networks [31], using methods that seem to have been rediscovered in [22]. Modern implementations include invariant/equivariant convolutional architectures [4, 6] that are inspired by concepts from mathematical physics and harmonic analysis [13, 5]. Some of these models even enjoy universal approximation properties [19, 33].

The earliest attempt at theoretical justification for invariance of which we are aware is [1], which roughly states that enforcing invariance cannot increase the VC dimension of a model. Anselmi et al. [2] and Mroueh, Voinea, and Poggio [21] propose heuristic arguments for improved sample complexity of invariant models. Sokolic et al. [27] build on the work of Xu and Mannor [32] to obtain a generalisation bound for certain types of classifiers that are invariant to a finite set of transformations, while Sannai and Imaizumi [23] obtain a bound for models that are invariant to finite permutation groups. The PAC Bayes formulation is considered in [16, 17].

The above works guarantee only a worst-case improvement and it was not until very recently that Elesedy and Zaidi [8] derived a strict benefit for invariant/equivariant models. Our work is similar to [8] in that we provide a provably strict benefit, but differs in its application to kernels and RKHSs as opposed to linear models. Also very recently, Mei, Misiakiewicz, and Montanari [20] analyse the generalisation benefit of invariance in kernels and random feature models. Our results differ from [20] in some key aspects. First, Mei, Misiakiewicz, and Montanari [20] focus kernel ridge regression with an invariant inner product kernel whereas we study symmetrised predictors from general kernels. Second, they obtain an expression for the generalisation error that is conditional on the training data and in terms of the projection of the predictor onto a space of high degree polynomials, while we are able to integrate against the training data and express the generalisation benefit directly in terms of properties of the kernel and the RKHS.

References

[1] Yaser S Abu-Mostafa. "Hints and the VC dimension". In: *Neural Computation* 5.2 (1993), pp. 278–288 (page 8).

[2] Fabio Anselmi et al. *Unsupervised Learning of Invariant Representations in Hierarchical Architectures*. 2014. arXiv: 1311.4158 [cs.CV] (page 8).

[3] Sanjeev Arora et al. "Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer 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. 322–332. URL: http://proceedings.mlr.press/v97/arora19a.html (page 2).

[4] Taco Cohen and Max Welling. "Group equivariant convolutional networks". In: *International conference on machine learning*. 2016, pp. 2990–2999 (page 8).

[5] Taco S Cohen, Mario Geiger, and Maurice Weiler. "A general theory of equivariant cnns on homogeneous spaces". In: *Advances in Neural Information Processing Systems*. 2019, pp. 9145–9156 (page 8).

[6] Taco S Cohen et al. "Spherical cnns". In: *arXiv preprint arXiv:1801.10130* (2018) (page 8).

[7] Cohn, Donald L. *Measure Theory*. 2nd ed. Springer, 2013 (page 4).

[8] Bryn Elesedy and Sheheryar Zaidi. "Provably Strict Generalisation Benefit for Equivariant Models". In: (2021). arXiv: 2102.10333 [stat.ML] (pages 1–3, 6, 8).

[9] Adam Foster, Rattana Pukdee, and Tom Rainforth. "Improving Transformation Invariance in Contrastive Representation Learning". In: *arXiv preprint arXiv:2010.09515* (2020) (page 1).

[10] 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 4, 8).

[11] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in neural information processing systems*. 2018, pp. 8571–8580 (page 2).

[12] Olav Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006 (page 7).

[13] 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*. 2018, pp. 2747–2755 (page 8).

[14] Jaehoon Lee et al. "Wide neural networks of any depth evolve as linear models under gradient descent". In: *Advances in neural information processing systems*. 2019, pp. 8572–8583 (page 2).

[15] 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 1).

[16] 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 8).

[17] Clare Lyle et al. *On the Benefits of Invariance in Neural Networks*. 2020. arXiv: 2005.00178 [cs.LG] (pages 1, 8).

[18] Jonathan H Manton and Pierre-Olivier Amblard. "A primer on reproducing kernel hilbert spaces". In: *arXiv preprint arXiv:1408.0952* (2014) (page 6).

[19] Haggai Maron et al. "On the Universality of Invariant Networks". In: *International Conference on Machine Learning*. 2019, pp. 4363–4371 (page 8).

[20] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "Learning with invariances in random features and kernel models". In: *arXiv preprint arXiv:2102.13219* (2021) (pages 1, 8).

[21] 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 8).

[22] Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. "Equivariance through parameter-sharing". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 2892–2901 (page 8).

[23] Akiyoshi Sannai and Masaaki Imaizumi. *Improved Generalization Bound of Group Invariant / Equivariant Deep Networks via Quotient Feature Space*. 2019. arXiv: 1910.06552 [stat.ML] (pages 1, 8).
[24] Bernhard Schölkopf, Chris Burges, and Vladimir Vapnik. "Incorporating Invariances in Support Vector Learning Machines". In: Springer, 1996, pp. 47–52 (page 8).
[25] 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. 1994, 387–390 vol.2 (page 8).
[26] Hanns Schulz-Mirbach. "On the existence of complete invariant feature spaces in pattern recognition". In: *International Conference On Pattern Recognition*. Citeseer. 1992, pp. 178–178 (page 8).
[27] Jure Sokolic et al. "Generalization error of invariant classifiers". In: *Artificial Intelligence and Statistics*. 2017, pp. 1094–1103 (pages 1, 8).
[28] James S Spencer et al. "Better, Faster Fermionic Neural Networks". In: *arXiv preprint arXiv:2011.07125* (2020) (page 1).
[29] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information science and statistics. Springer, 2008. ISBN: 978-0-387-77241-7 (pages 2, 4, 5).
[30] Marysia Winkels and Taco S Cohen. "3D G-CNNs for pulmonary nodule detection". In: *arXiv preprint arXiv:1804.04656* (2018) (page 1).
[31] Jeffrey Wood and John Shawe-Taylor. "Representation theory and invariant neural networks". In: *Discrete applied mathematics* 69.1-2 (1996), pp. 33–60 (pages 1, 8).
[32] Huan Xu and Shie Mannor. "Robustness and generalization". In: *Machine learning* 86.3 (2012), pp. 391–423 (page 8).
[33] Dmitry Yarotsky. *Universal approximations of invariant maps by neural networks*. 2018. arXiv: 1804.10306 [cs.NE] (page 8).
[34] Manzil Zaheer et al. "Deep sets". In: *Advances in neural information processing systems*. 2017, pp. 3391–3401 (page 1).

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 2.2 which describes our assumptions.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See **??** for general technical conditions not given in statements of results.
   (b) Did you include complete proofs of all theoretical results? [Yes] Proofs are given in the supplementary material
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]