Demystifying Spectral Bias on Real-World Data

Anonymous authors
Paper under double-blind review

Abstract

Machine learning models can appear to learn functions easily, yet generalize poorly out of distribution due to shortcut learning. This gap arises because statistical properties of the data are often entangled with the model's information processing. In this work, we aim to disentangle these two aspects by introducing *cross-dataset learnability*, which measures whether a function learned from one dataset is captured when evaluated on a choice of general distribution. We show that for kernel ridge regression (KRR) and Gaussian processes (GPs), cross-dataset learnability admits an explicit upper bound and that the cross-dataset learnability sample complexity can be predicted on real-world datasets. This framework further allows us to exploit symmetries present in realistic Neural Tangent Kernels (NTKs) and Neural Network Gaussian Processes (NNGPs), enabling a characterization of their spectral bias when learned from real-world and evaluated on symmetric distributions.

1 Introduction

Recent years have brought forth great advances in the understanding of machine learning (ML) models when learning from artificial data. The spectral bias of neural networks in the lazy regimes has been characterized extensively, e.g., MLPs acting on rotationally invariant data distributions (Basri et al., 2019; Yang & Salman, 2020; Bietti & Bach, 2021) or transformers acting on permutation invariant data (Lavie et al., 2024). Other works use Gaussian student-teacher settings to study neural networks in the rich regime (e.g. Arous et al. (2021); Rubin et al. (2023); Damian et al. (2024); Defilippis et al. (2025)), revealing fundamental limits of learnability and a rich phenomenology of phase transitions during learning.

Much less is known about learning from natural data. Achieving the same level of control would generally require a good mathematical description of the data itself, a task that seems to be currently beyond reach. Even in simpler, better understood setting like kernel ridge regression (KRR) where there are known results on universality, Gaussian equivalence, and staircase phenomena for dot product kernels Misiakiewicz (2022); Dubova et al. (2023) they are restricted to i.i.d. coordinates, limiting their applicability to real-world data.

This work aims to mitigate some of this gap by disentangling the information processing performed by the model from the statistical properties of the data. We introduce the concept of *cross-dataset learnability* (x-learnability), where one learns from real data but tests whether a function is learned or not on a synthetic dataset that serves as a standard ruler. For example, in an image classification task, a model may appear to recognize an object category while in fact relying on a correlated background cue—such as water for boats. Evaluating the model on a synthetic dataset in which such correlations are removed or randomized often reveals that the target function was not learned successfully. This possibility of shortcut learning Geirhos et al. (2020); Ye et al. (2025) is part of the challenge in analyzing learning from real data and the x-learnability allows us to bypass this challenge.

We focus on applying the idea of x-learnability to spectral bias in KRR or GP regression, with applications to highly overparameterized neural networks through the NTK Jacot et al. (2018) and NNGP Neal (1996); Lee et al. (2018) correspondences. Spectral bias provides a characterization of a model's inductive bias in terms of sample complexity - which functions are easiest to learn from finite data. This bias is governed by the eigenspectrum of the kernel operator in the infinite-data limit Canatar et al. (2021b). Decomposing the target function into kernel eigenfunctions, learning acts as a spectral filter: components associated with eigenvalues smaller than approximately σ^2/P are effectively suppressed, where P is the sample size and σ^2 is the ridge (or effective ridge) parameter Canatar et al. (2021b). For example, classical results on rotationally

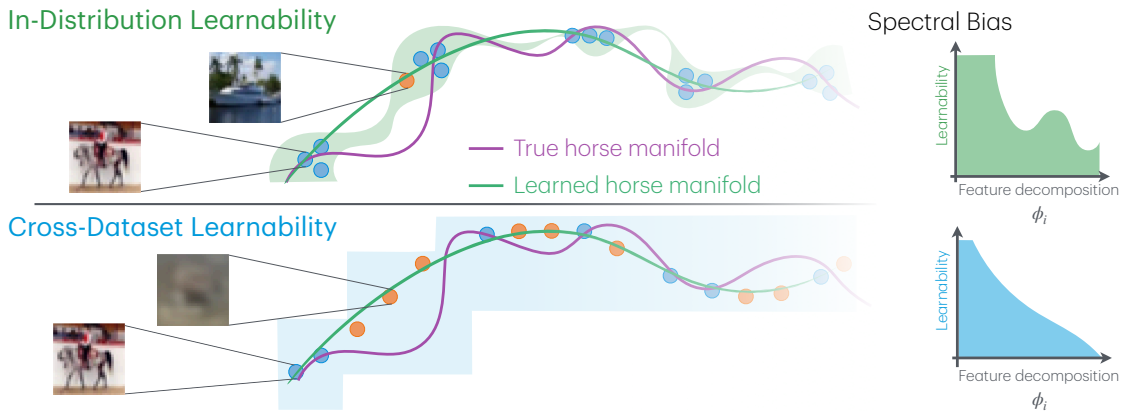

Figure 1: The x-learnability introduces a standard ruler for function similarity and has a clear, analytically tractable spectral bias. On the right, feature decomposition corresponds to eigenfunction ϕ_i as in Eq. 3 solved on a standard ruler measure q ranked by eigenvalues in decreasing order; see Section 2 for details.

invariant kernels and data show order-by-order learning of polynomials, with a degree-ℓ polynomial generically requiring $P \sim d^\ell$ samples in d dimensions.

In this work, we derive an upper bound on the learnability of learning from a general, real-world dataset D and testing on a simple standard ruler distribution q - the x-learnability. We then translate this bound to a lower bound on the sample complexity. We show that our bound predicts the scaling of the cross-dataset sample complexity accurately. Additionally, when the dataset D and the standard ruler distribution q are "closer", the equivalent kernel-like Sollich & Williams (2004) interpolator tracks the true learnability closely, enabling learning curve predictions from elementary dataset statistics.

Our main contributions are:

- We prove an analytically tractable upper bound on the x-learnability, requiring minimal knowledge of the dataset, and translate it to a sample complexity lower bound.
- We show that our sample complexity lower bound predicts the scale at which features are learned from multiple real datasets accurately. Additionally, when the dataset is whitened, the x-learnability tracks the in-distribution learnability closely, effectively enabling learning curve prediction.
- We analyze spectral bias in transformers and show that our method predicts large sample complexity for copying heads, and hence for several elementary in-context learning mechanisms, in line with empirical findings.

1.1 Related Works

Learning & multiple data distributions. There is a large body of work on distributional shift and out-of-distribution generalization Ben-David et al. (2006; 2010); Pan & Yang (2010); Arjovsky (2021); Canatar et al. (2021a); Ma et al. (2023); Feng et al. (2023), however, this setting and its motivation are different from ours. In the study of distributional shifts, one assumes the training (source) distribution does not accurately reflect the test (target) distribution and tries to estimate (bound, guarantee) the performance on the test distribution, which is the true object of interest. This concept fundamentally differs from the setting in this work. Here, we do not assume a difference between the underlying distribution for the test and train set; rather, we are interested in the spectral bias on real-world datasets and use q, a "standard ruler" distribution, solely as a tool. Presenting q allows us to decouple the function implemented by the model from the specific training data used to learn it. Importantly, we assume full knowledge of q, whereas works on distributional shift assume no access to the target distribution.

Kernels, lazy learning & NNs. The correspondence between infinitely wide neural networks and kernels (NNGP correspondence) was first presented in Neal (1996) for a Bayesian setting, and was later extended to deep networks and other architectures Lee et al. (2018); Matthews et al. (2018); Novak et al. (2018); Yang (2019); Hron et al. (2020). In parallel, a correspondence between infinitely wide NNs in a gradient flow setting

(NTK correspondence) was introduced in Jacot et al. (2018) and extended in Arora et al. (2019); Yang (2020); Yang & Littwin (2021). The general phenomenon of lazy learning was further clarified in Chizat et al. (2019). Finally, recent theoretical works Li & Sompolinsky (2021); Seroussi et al. (2023); Fischer et al. (2024); Lauditi et al. (2025) use adaptive kernel approaches to explain the performance gap between NNGP/NTK and NNs. This line of work suggests that even post-training in the rich regime, NNs are still described by a kernel similar to the NNGP/NTK, which is augmented with information on the data and target. Understanding interaction with data in the lazy regime can thus be the first step in understanding it in richer regimes of NNs.

Symmetry in kernels and neural networks & spectral bias. Symmetry has been used extensively to understand kernels and neural networks in the kernel regime from the perspective of spectral bias. Kernels corresponding to fully connected networks with data uniformly distributed on a hypersphere have been studied in Basri et al. (2019); Bietti & Bach (2021); Scetbon & Harchaoui (2021). Data distributed non-uniformly on the hypersphere was studied in Basri et al. (2020), while Gaussian data and data uniformly distributed on the hypercube were studied in Yang & Salman (2020). The eigenspectra of kernels corresponding to convolutional neural networks were studied in Bietti (2021); Xiao (2022); Cagnetta et al. (2023); Geifman et al. (2022). Finally, Lavie et al. (2024) studied transformers in the kernel regime by leveraging their permutation symmetry. Here, we show that spectral bias results have a universal component and show how they can be adapted from idealized symmetric datasets to real-world datasets by choosing those idealized distributions as standard rulers q.

2 Cross-Dataset Learnability

In this section, we present a short introduction to KRR or GP regression and the concept of learnability, followed by a generalization to cross-dataset learnability (x-learnability).

The regression setting includes a kernel function $k(x, y)$, a ridge parameter σ^2, and a dataset $D = \{(x_\mu, y(x_\mu))\}_{\mu=1}^{P}$ of P data points drawn i.i.d. from an underlying distribution p. x_μ denotes the μ'th input and $y(x_\mu)$ is the regression target/label for x_μ. In this case, the predictor is given by

$$\hat{f}_D(x) = \sum_{\nu,\rho=1}^{P} k(x, x_\nu) \left[K + I\sigma^2\right]^{-1}_{\nu\rho} y(x_\rho), \tag{1}$$

where $[K]_{\mu\nu} = k(x_\mu, x_\nu)$, I is the identity matrix and $x_\mu \in D$. In a contemporary context, kernels of particular interest are the NTK and NNGP, which describe infinitely wide neural networks in the lazy regime.

We can define the *learnability* of the target as

$$\mathcal{L} := \frac{\mathbb{E}_{x\sim p}\left[y(x)\hat{f}_D(x)\right]}{\mathbb{E}_{x\sim p}\left[y(x)y(x)\right]} \tag{2}$$

The learnability monotonically increases with P (the dataset size), and takes values in the range $[0, 1)$ such that $\mathcal{L} \overset{P\to\infty}{\longrightarrow} 1$ for $y \in \mathcal{H}_k$ where $\mathcal{H}_k$ is the RKHS of k. Losses such as MSE are also tightly related to the learnability (Simon et al., 2023).

The learnability is known to be controlled by the eigenvalue decomposition $\{(\psi_i(x), \eta_i)\}_{i=1}^{\infty}$ of the kernel operator in the infinite data limit $\hat{K}_p$

$$\begin{aligned} \hat{K}_p\psi_i(x) &:= \mathbb{E}_{x'\sim p}\left[k(x, x')\psi_i(x')\right] = \eta_i\psi_i(x), \\ \mathbb{E}_{x\sim p}\left[\psi_i(x)\psi_j(x)\right] &= \delta_{ij}, \quad x \in \mathrm{supp}(p), \end{aligned} \tag{3}$$

with δ_{ij} the Kronecker delta and $\mathbb{E}_{x\sim p}[\,\cdot\,]$ being expectation value w.r.t the distribution p.

Under the equivalent kernel (EK) approximation Silverman (1984); Sollich & Williams (2004); Cohen et al. (2021), where essentially the regression is done of the full population rather than a specific realization of a dataset D, the learnability of a specific eigenfunction feature ψ_i (setting $y(x) = \psi_i(x)$) takes an especially simple form

$$\mathcal{L}_i = \frac{\mathbb{E}_{x\sim p}\left[\psi_i(x)\hat{f}_D(x)\right]}{\mathbb{E}_{x\sim p}\left[\psi_i(x)y(x)\right]} \overset{\text{EK}}{\approx} \frac{\eta_i}{\eta_i + \sigma^2/P}. \tag{4}$$

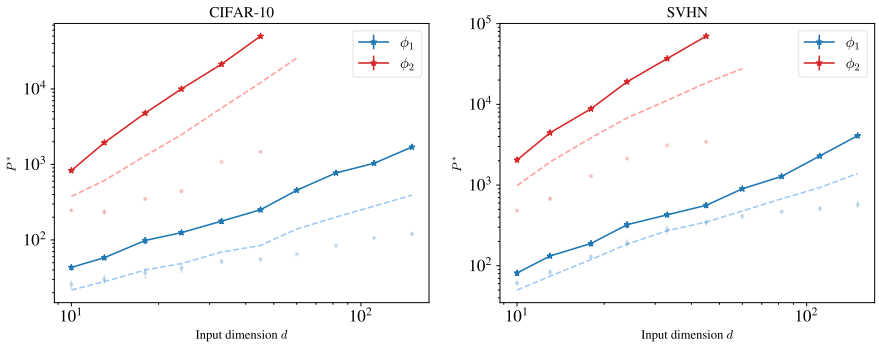

Figure 2: **(The sample complexity lower bound predicts the scaling accurately)** The cross-dataset learnability (x-learnability) sample complexity (taken at $\epsilon = 1/2$) and our bound on the x-learnability (dashed) as a function of the input dimension d. The input dimension is controlled by keeping the top d principal components. **Left:** The model is trained on the linear and quadratic features of a CIFAR-10 classifier that achieves performance comparable to a standard MLP on CIFAR-10 classification task; see main text for details. The in-distribution learnability is indicated by dots, and is very different from the x-learnability, indicating a large gap between how the learn function appears on the dataset and how it appears on our standard ruler. **Right:** same experiment for SVHN dataset.

The EK approximation is exact for large datasets $P \gg 1$ and large ridge σ^2. Beyond this regime, in high dimension, the same form of learnability holds when one replaces the "bare" ridge with an effective ridge $\sigma^2 \to \sigma^2_{\text{eff}}$ (Canatar et al., 2021b; Simon et al., 2023).

Learnability results can be restated as a function of the number of samples, resulting in sample complexity; e.g. requiring learnability $1 - \epsilon$ gives

$$P_i^* = \eta_i^{-1} \sigma^2 \frac{1-\epsilon}{\epsilon}, \tag{5}$$

with P_i^* the sample complexity to achieve learnability that is ϵ close to unity for the feature (eigenfunction) ψ_i. Larger eigenvalues give better (lower) sample complexity, inducing a *spectral bias* - the regression will learn those features first and use them to explain the data. Whether a model needs few or many samples, generalizes or overfits, is largely dependent on the spectral bias and the support of the target on the eigenfunctions. Even under the assumption that the true population distribution and target function are known, solving the eigenvalue problem in Eq.equation 3 remains an intractable task, leaving the spectral bias on rich and complex data obscure. Moreover, even in the special cases where the problem is solvable, the information processing applied by the model remains entangled with the data properties, making effective separation difficult. To bypass these difficulties, we consider a cross-dataset generalization of this setting, where one solves the eigenvalue problem given in Eq.equation 3 on an “standard ruler" dataset with a probability measure q, but performs the regression on the dataset D. Analogous to the learnability in Eq. 2, we introduce *x-learnability* where one replace the population distribution p with the “standard ruler" distribution q

$$\mathcal{L}_i^{D,q} := \frac{\mathbb{E}_{x\sim q}\left[\phi_i(x)\hat{f}_D(x)\right]}{\mathbb{E}_{x\sim q}\left[\phi_i(x)y(x)\right]} \tag{6}$$

with ϕ_i an eigenfunction of the kernel w.r.t. q, i.e., ϕ_i solves Eq. equation 3 with p replaced by q. We note that while the common learnability in Eqs. (2,4) is bounded $\mathcal{L}_i \in [0, 1)$, the x-learnability is unbounded. As a consequence, maximizing x-learnability does not imply good learning; instead, one must require it to be close to unity. In the next section, we present a tractable upper bound on the x-learnability and show empirically that it predicts spectral bias on real-world datasets. The rest of this section gives a more intuitive understanding of x-learnability.

2.1 Interpreting Cross-Dataset Learnability

Intuitively, the change of distribution amounts to learning from D but judging how good the reconstruction is based on functional similarity on q. Importantly, q is not intended to approximate the data distribution

p, nor to represent a realistic testset. Rather, it serves as a diagnostic ruler that allows us to probe which functions are implemented by the predictor in a controlled and interpretable basis: the predictor $\hat{f}_D$ is a function whose definition only requires a domain (support), whereas the loss induces a data-dependent notion of functional similarity by weighting discrepancies according to the data distribution. When the distribution is concentrated near a low-dimensional region, functions that differ primarily in directions orthogonal to that region can therefore appear artificially similar, and evaluating on a symmetric, high-entropy reference distribution q largely mitigates this bias; at the same time, it is analytically convenient because the kernel eigendecomposition on q can often be characterized using symmetry. Choosing a simple q has the advantage of providing a clear ruler for measuring network outputs, as demonstrated in the next example, capturing a notion of out-of-distribution generalization, but with the disadvantage of being uninformed about the details of the specific dataset.

As an illustrative example, consider learning the parity function on the d-dimensional hypercube, defined by $y(x) = \prod_{i=1}^{d} x_i$ for $x \in \{-1, 1\}^d$. Under generic conditions—and in particular under the uniform measure—learning parity is notoriously difficult (e.g., Shalev-Shwartz et al. (2017); Yang & Salman (2020)). However, if the data distribution p induces perfect correlations among the coordinates and d is even, the target collapses to the constant function $y(x) \equiv 1$, rendering the task trivial and susceptible to shortcut learning. In this setting, evaluating the learned predictor under a uniform, symmetric measure on the hypercube (or its hyperspherical extension) reveals that a function achieving perfect performance on p does not implement parity in terms of its underlying computation. When the goal is to assess the information processing performed by the network, x-learnability with respect to a uniform reference distribution q therefore provides a more meaningful notion of function similarity. In Appendix F, we show that achieving good x-learnability for parity with a rotationally invariant kernel while learning from any distribution on the hypercube with a "standard ruler" distribution q that is uniform on the hypersphere requires sample complexity exponential in the dimension.

3 A Bound on the Cross-Dataset Learnability

Theorem 3.1. *Given a kernel $k(x, y)$ and its eigendecomposition $\{(\lambda_i, \phi_i(x))\}$ on a "standard ruler" measure $q(x)$ and a dataset $D = \{(x_\mu, y(x_\mu))\}_{\mu=1}^{P}$ of P samples, the x-learnability (see Eq.6) is bounded from above by*

$$\mathcal{L}_i^{D,q} \leq \underbrace{\frac{\lambda_i P}{\sigma^2}}_{\substack{\textit{Model and ruler} \\ \textit{dependent}}} \underbrace{\frac{\sqrt{\mathbb{E}_{x\sim D}\left[\phi_i^2(x)\right] \mathbb{E}_{x\sim D}\left[y^2(x)\right]}}{\left|\mathbb{E}_{x\sim q(x)}[\phi_i(x) y(x)]\right|}}_{\textit{Dataset and target dependent}} . \tag{7}$$

with $\mathbb{E}_{x\sim D}[\,\cdot\,]$[1] ($\mathbb{E}_{x\sim q(x)}[\,\cdot\,]$) being expectation value w.r.t the dataset D (q). The bound holds as long as the training dataset D is a subset of the support of q – $\mathrm{supp}(q) \supseteq \{x_\mu\}_{\mu=1}^{P}$; note that this is the only requirement on q. The proof is given in Appendix A.

We see that the information about the training dataset in equation 7 is fully contained in the norm of the feature ϕ_i and the target y on the dataset D, with the latter often being $O(1)$. The equation also requires knowledge of the spectral decomposition of the target of the target on the standard ruler q, as is required by spectral bias approaches. The last component, is the model and ruler dependent eigenvalue, which is universal across different datasets.

While the bound in Eq. 7 may exceed unity for large P, its primary utility is in identifying the sample complexity threshold (P^*) discussed below. Additionally, as we show in Appendix B under favorable assumptions, it is the optimal linear approximation and upper bound to the learnability at the small P limit.

Our main use of the bound is predicting the sample complexity of different features and thus characterizing the spectral bias of the model.

[1]This quantity may naturally fluctuate a little as one scans the values of P.

Corollary 3.2. *As a corollary from 3.1 the number of samples P^* required to achieve x-learnability $\mathcal{L}_t^{D,q} = 1-\epsilon$ for a specific target feature $\phi_t(x)$ is bounded from below by*

$$P^* \geq \sigma^2 \lambda_t^{-1}(1-\epsilon) \frac{\left|\mathbb{E}_{x\sim q(x)}[\phi_t(x)y(x)]\right|}{\sqrt{\mathbb{E}_{x\sim D}\left[\phi_t^2(x)\right]\mathbb{E}_{x\sim D}\left[y^2(x)\right]}}, \tag{8}$$

The result in equation 8 can be interpreted as follows. At least P^* samples are required to learn a feature $\phi_i(x)$ from the dataset D, where the eigenvalue/features are found by performing eigenvalue decomposition of the kernel on q. This bound can be seen as a prescription for carrying over sample complexity / spectral bias results found on a favorable "standard ruler" measure q to a rich dataset of interest D. It can be used whenever one is able to perform the eigenvalue decomposition on q, for q that includes D in its support.

Importantly, one is free to choose a favorable measure q. In particular, one may choose q such that it respects all the symmetries of the kernel function k. Such a choice simplifies the eigenproblem and makes a naturally uniform measure that's equally sensitive to all of its support. We dedicate the next subsection to investigating that.

3.1 Symmetries and the "standard ruler" measure q

Here, we motivate choosing the "standard ruler" distribution q based on guiding principles of symmetry.

Definition 3.3 (Kernel Symmetry)**.** We say a kernel $k(x, y)$ has a symmetry group G if

$$\forall g \in G \quad k(T_g x, T_g y) = k(x, y) \tag{9}$$

for T_g a faithful representation of $g \in G$. *Intuitively*, a kernel symmetry means the kernel treats a pair of inputs and their symmetry augmented pairs in the same way.

Definition 3.4 (Dataset Symmetry)**.** We say a dataset measure q is symmetric under the action of a group G if

$$\forall g \in G \quad q(T_g x) = q(x) \tag{10}$$

for T_g a faithful representation of $g \in G$. *Intuitively*, a dataset symmetry means a symmetry-augmented version is as likely to be seen in the dataset as the original one.

When both the kernel and dataset respect a symmetry group G we say it is a symmetry of the kernel *operator*. A symmetry of the kernel operator can be used to asymptotically bound the eigenvalues from above using the dimension of their corresponding irreducible representations (irreps), as shown for the case of fully connected networks (Basri et al., 2019; Yang & Salman, 2020; Bietti & Bach, 2021) and transformers (Lavie et al., 2024). A natural choice for q is therefore one that respects all the kernel symmetries

$$\forall g \in G \quad k(T_g x, T_g y) = k(x, y) \rightarrow q(T_g x) = q(x). \tag{11}$$

Such a choice also ensures that the dataset does not ignore or favor any part of the input space, making it an appropriate standard ruler.

In the last section, we give an example where we analyze the spectral bias of transformers from the lens of symmetry and show that large context size and large vocabulary imply strong spectral bias, predicting large sample complexity for the copying head task.

4 Experiments

Here we experiment with x-learnability and spectral bias. We use CIFAR-10 (Krizhevsky, 2009) and SVHN Netzer et al. (2011), as a standard benchmarks and include results on Fashion MNIST (Xiao et al., 2017) and MNIST (LeCun et al., 2010) in Appendix C. First, (1) we show our sample complexity bound in Eq. equation 8 is able to predict the spectral bias of the x-learnability. Then, (2) we show that x-learnability tracks the regular in-distribution learnability closely when we use PCA whitening to make the data and the

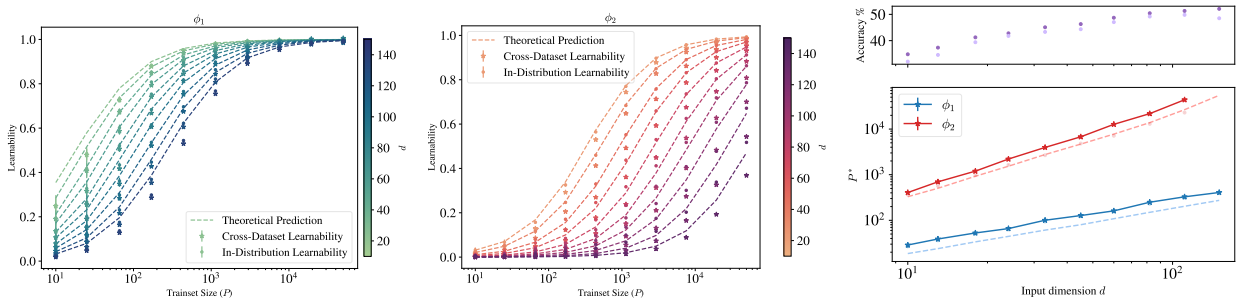

Figure 3: **(Theory predicts performance on whitened data accurately)** The x-learnability (stars) and (test) learnability (dots) together with the bound on the x-learnability bound in Eq. equation 7 (dashed) on whitened CIFAR-10. The first two panels from the left show learning curves for linear and quadratic features (respectively) as a function of the number of samples P, with different colors indicating different input dimensions d. The bottom figure in the right panel is analogous to figure 2, this time with PCA whitening; the dots are often covered by the stars indicating perfect agreement. The top part of the right panel shows the performance in terms of accuracy for the quadratic teacher model for whitened (light colored dots) and unwhitened (full-colored dots) data as a function of the d; the performance is on par with the performance of an MLP on CIFAR-10 classification.

"standard ruler" distribution more similar. The whitening transformation erases first-order correlations but leaves higher statistical dependencies. We emphasize that the second statement (2) is a falsifiable empirical claim supported by experiments across multiple datasets and settings, rather than a theoretical guarantee.

In Figure 2 we show an experiment of KRR with an NNGP kernel corresponding to a single hidden-layer ReLU network learning the linear (ϕ_1) and quadratic (ϕ_2) features of a CIFAR-10 (right: SVHN) classifier. To visualize our result in equation equation 8, we choose $\epsilon = 1/2$, and vary the input dimension by keeping the top d principal components (see Appendix D for visualization). We pre-process each sample by centering to remove the dataset mean and normalizing the Euclidean norm to unity, such that all samples lie on the hypersphere. In line with this choice, we choose the standard ruler distribution q to be uniform on the hypersphere where the eigendecomposition of the kernel is well understood.

The target features are extracted from a trained quadratic teacher model via symmetrization and anti-symmetrization after subtracting the constant component; effectively decomposing the CIFAR-10 (SVHN) labels to hyperspherical harmonics of $\ell = 1, 2$. The corresponding eigenvalues are $\lambda_\ell \sim d^{-\ell}$ for d the input dimension, hence the variation in the input dimension. We note that in this case, the quadratic teacher achieves performance comparable to an MLP on CIFAR-10 of $\sim 50\%$ accuracy (see top bar of Fig. 3 right panel); this level of performance in KRR is made possible by the eigenpro solver Ma & Belkin (2019), allowing us to process datasets that are too large for naive KRR implementation due the memory constraints.

In Figure 2, we can also see the in-distribution learnability indicated by shaded dots. We note that the in-distribution learnability differs substantially from the x-learnability, reflecting the complex structure of the dataset. Notably, on the CIFAR-10 dataset, the model appears to learn a quadratic when it cannot do so on the full hypersphere (as indicated by the fact the dots are below the lower bound).

Finally, we examine the choice of the ruler q and show that when the dataset distribution is close enough to q our bound becomes an accurate predictor of the performance on the testset. We repeat the setting above but use PCA *whitening*. PCA whitening makes the data covariance spherically symmetric, thus making it more similar to the uniform distribution on the hypersphere. The results in Fig. 3 indeed shows the cross-data set learnability (stars) approximated the learnability (dots) well. For better visualization of our sample complexity result, we use the EK-like interpolator, plotting a "prediction" (dashed) for the learnability based on what the eigenvalue λ_i must be to match the same sample complexity under the EK approximation (see equation equation 4). On the left and middle panel, we show learning curves (learnability vs. trainset size) for the linear and quadratic features (resp.). On the bottom right panel, we show a plot analogous to Fig 2 for the whitened data, and at the top, the accuracy achieved by the quadratic teacher, showing performance

on-par with MLPs Lee et al. (2020). We note that dimensional reduction with PCA together with whitening was found to be beneficial to the performance of other NNGP and NTK kernels in Lee et al. (2020).

In Appendix C, we also show results on randomly sampled eigenfunctions with $\ell = 1, 2, 4$, instead of the features extracted from the quadratic teacher.

5 Vignettes

Here we give two examples of implementing our main results in Eqs. (7,8). The first one is a simple linear regression, showing how restricting the inputs to a sub-manifold that is relevant to the target can reduce sample complexity and how this fact enters our bound. The second example is inspired by in-context learning (ICL) in transformers and uses spectral bias to estimate the sample complexity of the copying head task.

5.1 Awareness to Dataset & Linear Regression on a Low Dimensional Data Manifold

The purpose of this example is twofold: first, present a simple setting as an example of how one can use our main result, and second, discuss the ways in which our bound accounts for different training datasets.

Consider the kernel $K(x, y) = \frac{1}{d} x \cdot y$ with $x, y \in R^d$ and $q(x) = \prod_{i=1}^{d} p_N(x_i)$ where $p_N(x_i)$ are standard centered Gaussians. Let D consist of P d-dimensional vectors sampled i.i.d. from the distribution p such that $\vec{x} \sim p$ obeys

$$\vec{x} = (x_1, x_2, ..., x_d), \quad x_1 \sim N(0, d), x_{i>1} = 0. \tag{12}$$

Where the scale was chosen such that the sum of eigenvalues (trace) of the kernel on both measures equals unity

$$\int k(\vec{x}, \vec{x}) \frac{e^{-\frac{x_1^2}{2d}}}{\sqrt{2\pi d}} \prod_{i=2}^{d} \delta(x_i)\, d\vec{x} = \int k(\vec{x}, \vec{x}) \prod_{i=1}^{d} \frac{e^{-\frac{x_i^2}{2}}}{\sqrt{2\pi}}\, d\vec{x} = 1 \tag{13}$$

with $\delta(x)$ the Dirac delta distribution, such that the learnability budget is the same in both distributions. Finally, let $y(\vec{x}) = x_1$.

Let us first estimate the learnability here without using the above bound. The kernel on this dataset coincides with $d^{-1} x_1' \cdot x_1$. From an EK perspective, this has a single non-zero eigenvalue η, associated with the function $\psi = x_1/\sqrt{d}$ (normalized w.r.t. p), given by

$$\int_{-\infty}^{\infty} dx_1' \frac{x_1' x_1}{d} \frac{x_1'}{\sqrt{d}} \frac{e^{-\frac{x_1'^2}{2d}}}{\sqrt{2\pi d}} = \eta \frac{x_1}{\sqrt{d}}; \qquad \eta = 1 \tag{14}$$

hence the standard learnability (under EK approximation) is

$$\mathcal{L} = \frac{\eta}{\eta + \sigma^2/P}. \tag{15}$$

Implying that $P^* = \eta^{-1}\sigma^2 = \sigma^2$, for $\epsilon = 1/2$, the crossover value that allows to learn half the target value.

Next, we calculate the cross-dataset. To this end, we require the eigenfunctions and eigenvalues with respect to q. These can be checked to be

$$\phi_i(x) = x_i \qquad \lambda_i = d^{-1} \tag{16}$$

Our target in these terms is $y(x) = \phi_1(x)$. Using EK approximation to estimate the predictor, we find that the x-learnability is exactly equal to the true learnability, hence implying the sample complexity $P^* = \sigma^2$.

Finally, we apply our bound. Applying the bound in equation equation 8 for $\epsilon = 1/2$

$$P^* = \frac{\sigma^2 \lambda_1^{-1}}{2} \frac{\mathbb{E}_{x\sim q(x)}[\phi_1(x) y(x)]}{\sqrt{\mathbb{E}_{x\sim D}\left[\phi_1^2(x)\right] \mathbb{E}_{z\sim D}\left[y^2(x)\right]}} \approx \frac{\sigma^2}{2} \frac{d}{d} = \frac{\sigma^2}{2} \tag{17}$$

where the last approximation is due to replacing empirical sampling (D) with the expected one. In this case, the resulting bound is within a factor of 2 from the exact result.

Awareness to the training dataset. Clearly, learning the target from the symmetric measure q is a harder task, requiring $P^* = \sigma^2 d$ samples. We see that our bound equation 8 encodes the information about the training dataset D by the expected norms $\sqrt{\mathbb{E}_{x\sim D}\left[\phi_1^2\left(x\right)\right]}$, $\sqrt{\mathbb{E}_{x\sim D}\left[y^2(x)\right]}$ of the target function and feature; which scales with d in the above example.

5.2 Learning Copying Heads with Transformers

Here, we use our theory to examine a concrete learning problem that has attracted recent attention (Reddy, 2023; Edelman et al., 2024; Sanford et al., 2024; Nichani et al., 2024; Singh et al., 2024) - learning induction heads (Olsson et al., 2022). An induction head performs an elementary form of ICL, where the next token is predicted based on the frequency of the tokens that followed it in the context, as in the form

$$[A][B]...[A] \to [B]. \tag{18}$$

Olsson et al. (2022) observed that learning induction heads consistently requires $\approx 10^9$ tokens (training steps) across a range of model sizes, with constant vocabulary size and context size.

The large number of tokens required to learn this simple function might appear surprising. Motivated by this we give here a spectral bias perspective on the problem. We consider the simpler problem of a Transformer kernel learning a copying head (that is simply copying the token that came before), which is essential for the creation of induction heads and mesa-optimization algorithms (Olsson et al., 2022; von Oswald et al., 2023). Thus, such a simple example can be used as a lower bound on sample complexity for a wide variety of ICL mechanisms in the kernel regime.

We define the input as $[X]_i^a$ where $i = 1, ..., V + 1$ is the vocabulary index, and $a = 1, ..., L + 1$ is the position of the token in the sequence. In this notation the target is $[Y(X)]_i^a = [X]_i^{a-1}$. For simplicity, we introduce a vector notation where $\vec{x}^a$ is a slice of X at the position a, and a scalar notation $x_i^a = [X]_i^a$. We note that we do not consider causal masking in this example, as the target function, as defined in the simplified setup above, does not rely on a notion of causality.

We set out to establish a sample complexity lower bound for this task based on our bound in Eq. equation 8. As a "standard ruler" distribution q we choose a simple symmetric distribution where all tokens within sequences and samples are drawn uniformly i.i.d. from the vocabulary and are one-hot encoded, i.e. integers $v = 1, ..., V + 1$ such that $x_i^a = \delta_{i,v}$ where δ is the Kronecker delta. The training dataset D will be one-hot encoded, with the same vocabulary size $V + 1$, but with arbitrary dependencies between the tokens. In particular, *D can be a true natural language dataset, tokenized with vocabulary size $V + 1$.*

We consider a family of transformer models in the NTK and NNGP limits described in Hron et al. (2020). We require the use of learned positional encoding (to preserve statistical permutation symmetry over all possible initializations) and no causal masking (as mentioned above). We allow models of any depth and non-linearities, as our argument relies on symmetries and representation theory following the work of Lavie et al. (2024).

As mentioned in Subsection 3.1, the eigenvalues of a kernel can be bounded from above by the trace over the dimension of the irreducible representation (irrep). To get the strongest lower bound we should target the component that belongs to the largest irrep. Linear functions, such as copying heads, contain only two irreps: a *trivial* and a *standard* representation. The trivial representation is one-dimensional and is simply the sum over all coordinates, so we can find the standard irrep component by projecting out this part. Concretely, we use the projectors

$$P \coloneqq (I - P^V_{\text{trivial}})(I - P^L_{\text{trivial}}); \qquad P^V_{\text{trivial}} f_i^a \coloneqq \frac{1}{V}\sum_i f_i^a; \qquad P^L_{\text{trivial}} f_i^a \coloneqq \frac{1}{L}\sum_a f_i^a \ , \tag{19}$$

where P^V and the P^L project into the trivial subspaces. Applying this projector to the target and normalizing to the feature to unite norm gives

$$\vec{\phi}_t^a(X) = \frac{Py}{\sqrt{L\left(1-L^{-1}\right)\left(1-V^{-1}\right)}} = \frac{1}{\sqrt{L\left(1-L^{-1}\right)\left(1-V^{-1}\right)}}\left(\vec{x}^{\,a-1} - \frac{1}{L}\sum_{b=1}^{L}\vec{x}^{\,b}\right) \tag{20}$$

with an eigenvalue that can be bounded from above (see more details in Appendix H)

$$\lambda_t \leq \frac{\mathbb{E}_{X\sim q}[k(X,X)]}{(L-1)(V-1)}. \tag{21}$$

We next use this result together with our main result equation 8 to bound the sample complexity of copying heads.

Next, to apply our bound in Eq. 8 we need to calculate or bound the norm of the target on the training set. Under the assumption of one-hot encoded input $\mathbb{E}_{X\sim D}\left[\mathrm{Tr}[Y(X)Y^T(X)]\right] = \mathbb{E}_{X\sim D}\left[\sum_{a=1}^{L}\vec{y}^{\,a}(X)\cdot\vec{y}^{\,a}(X)\right] = L$ for all D. Finally

$$\begin{aligned}\mathbb{E}_{X\sim D}\left[\mathrm{Tr}[\Phi\left(X\right)\Phi^T\left(X\right)]\right] &= \mathbb{E}_{X\sim D}\left[\sum_{a=1}^{L}\vec{\phi}_t^a\left(X\right)\cdot\vec{\phi}_t^a\left(X\right)\right] \\ &= \frac{1-L^{-2}\sum_{a,b=1}^{L}\sum_{i=1}^{V}\mathbb{E}_{X\sim D}[x_i^a x_i^b]}{\left(1-L^{-1}\right)\left(1-V^{-1}\right)}\end{aligned} \tag{22}$$

depends on the choice of D, nevertheless, one can easily derive bounds on the quantity. A simple bound is $\mathbb{E}_{X\sim D}\left[\mathrm{Tr}[\Phi\left(X\right)\Phi^T\left(X\right)]\right] \leq \left(1-V^{-1}\right)^{-1}$ for any one-hot encoded input.

These results can be plugged in to equation 8 to find a general lower bound on the sample complexity, given in Eq. 47 in Appendix H. Assuming a long context $L \gg 1$, large vocabulary $V \gg 1$, and normalized kernel $\mathbb{E}_{X\sim q}[k(X,X)] \simeq 1$, the sample complexity bound simplifies to

$$P^* \gtrsim \sigma^2(1-\epsilon)LV. \tag{23}$$

In simple terms, the number of samples has to scale like the product of the context length and the vocabulary size to learn copying heads. This result can be seen as a lower bound for the number of samples required for models in the kernel limit to start performing ICL. We note Olsson et al. (2022) used $V = 2^{16}$ and $L = 8192$ (giving $L \cdot V \approx 0.5 \cdot 10^9$), showing that from a spectral bias perspective the sample complexity of such a task is indeed very large.

6 Outlook

In this work, we have introduced the x-learnability, a measure of learnability independent of the test distribution, allowing us to separate the information processing done by the model from the statistical properties of the data. We have shown that spectral bias measured by x-learnability has a universal part that can be largely understood by model symmetries, and a dataset-dependent part that accounts for the change in the norm of the target function and features.

Our results generalize symmetry-based spectral bias analyses, previously developed for the NTK/NNGP of fully connected networks (Basri et al., 2019; Yang & Salman, 2020; Bietti & Bach, 2021), CNNs (Bietti, 2021; Xiao, 2022; Cagnetta et al., 2023), and transformers (Lavie et al., 2024) under dataset symmetry assumptions, as bounds to real-world datasets when measured on a standard ruler distribution q through x-learnability. As a concrete application, we obtain a sample-complexity lower bound for learning copying heads in transformer kernels, highlighting a strong spectral-bias relevant for basic in-context learning mechanisms.

Many of the possibilities opened by our approach remain unexplored, like the application of such a bound to ridgeless regression. When analyzing the dataset D, it has been shown there is an effective ridge coming

from the unlearnable features (Cohen et al., 2021; Canatar et al., 2021b; Simon et al., 2023). It would be of interest to find a similar effective ridge for our setting, using the eigenvalues and vectors from the symmetric distribution q, without having to solve the eigenvalue problem on the underlying data distribution p (from which D is drawn). Finally, a recent line of work has shown that NNs trained in the rich regime can also be described as kernel machines, where the kernel adapts to the data Li & Sompolinsky (2021); Seroussi et al. (2023); Fischer et al. (2024); Lauditi et al. (2025). It would be of great interest to extend our results to this regime.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Martin Arjovsky. Out of Distribution Generalization in Machine Learning, March 2021. URL http://arxiv.org/abs/2103.02667. arXiv:2103.02667.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, number 731, pp. 8141–8150. Curran Associates Inc., Red Hook, NY, USA, December 2019. URL https://dl.acm.org/doi/10.5555/3454287.3455018.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference, May 2021. URL http://arxiv.org/abs/2003.10409. arXiv:2003.10409 [stat].

Ronen Basri, David Jacobs, Yoni Kasten, and Shira Kritchman. The Convergence Rate of Neural Networks for Learned Functions of Different Frequencies. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/5ac8bb8a7d745102a978c5f8ccdb61b8-Abstract.html.

Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. Frequency Bias in Neural Networks for Input of Non-Uniform Density. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 685–694. PMLR, November 2020. URL https://proceedings.mlr.press/v119/basri20a.html. ISSN: 2640-3498.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of Representations for Domain Adaptation. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://papers.nips.cc/paper_files/paper/2006/hash/b1b0432ceafb0ce714426e9114852ac7-Abstract.html.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, May 2010. ISSN 1573-0565. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

Alberto Bietti. Approximation and Learning with Deep Convolutional Models: a Kernel Perspective. October 2021. URL https://openreview.net/forum?id=lrocYB-0ST2.

Alberto Bietti and Francis Bach. Deep Equals Shallow for ReLU Networks in Kernel Regimes, August 2021. URL http://arxiv.org/abs/2009.14397. arXiv:2009.14397 [cs, stat].

Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-Tolerant Learning, the Parity Problem, and the Statistical Query Model. *arXiv e-prints*, art. cs/0010022, October 2000. doi: 10.48550/arXiv.cs/0010022.

Francesco Cagnetta, Alessandro Favero, and Matthieu Wyart. What Can Be Learnt With Wide Convolutional Neural Networks? In *Proceedings of the 40th International Conference on Machine Learning*, pp. 3347–3379. PMLR, July 2023. URL https://proceedings.mlr.press/v202/cagnetta23a.html. ISSN: 2640-3498.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Out-of-Distribution Generalization in Kernel Regression. In *Advances in Neural Information Processing Systems*, November 2021a. URL https://openreview.net/forum?id=-h6LdcOMO-.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1): 2914, May 2021b. ISSN 2041-1723. doi: 10.1038/s41467-021-23103-1. URL https://www.nature.com/articles/s41467-021-23103-1. Number: 1 Publisher: Nature Publishing Group.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/ae614c557843b1df326cb29c57225459-Abstract.html.

Omry Cohen, Or Malka, and Zohar Ringel. Learning curves for overparametrized deep neural networks: A field theory perspective. *Physical Review Research*, 3(2):023034, April 2021. doi: 10.1103/PhysRevResearch.3.023034. URL https://link.aps.org/doi/10.1103/PhysRevResearch.3.023034. Publisher: American Physical Society.

Alex Damian, Loucas Pillaud-Vivien, Jason D. Lee, and Joan Bruna. Computational-Statistical Gaps in Gaussian Single-Index Models, March 2024. URL http://arxiv.org/abs/2403.05529. arXiv:2403.05529 [cs].

Leonardo Defilippis, Yizhou Xu, Julius Girardin, Emanuele Troiani, Vittorio Erba, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Scaling Laws and Spectra of Shallow Neural Networks in the Feature Learning Regime, September 2025. URL http://arxiv.org/abs/2509.24882. arXiv:2509.24882 [cs].

Sofiia Dubova, Yue M. Lu, Benjamin McKenna, and Horng-Tzer Yau. Universality for the global spectrum of random inner-product kernel matrices in the polynomial regime, October 2023. URL http://arxiv.org/abs/2310.18280. arXiv:2310.18280 [math].

Benjamin L. Edelman, Ezra Edelman, Surbhi Goel, Eran Malach, and Nikolaos Tsilivis. The Evolution of Statistical Induction Heads: In-Context Learning Markov Chains, February 2024. URL http://arxiv.org/abs/2402.11004. arXiv:2402.11004 [cs].

Xingdong Feng, Xin He, Caixing Wang, Chao Wang, and Jingnan Zhang. Towards a Unified Analysis of Kernel-based Methods Under Covariate Shift. *Advances in Neural Information Processing Systems*, 36:73839–73851, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e9b0ae84d6879b30c78cb8537466a4e0-Abstract-Conference.html.

Kirsten Fischer, Javed Lindner, David Dahmen, Zohar Ringel, Michael Krämer, and Moritz Helias. Critical feature learning in deep neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 13660–13690. PMLR, July 2024. URL https://proceedings.mlr.press/v235/fischer24a.html. ISSN: 2640-3498.

Christopher Frye and Costas J. Efthimiou. Spherical Harmonics in p Dimensions. *arXiv e-prints*, art. arXiv:1205.3548, May 2012. doi: 10.48550/arXiv.1205.3548.

William Fulton and Joe Harris. *Representation Theory: A First Course*, volume 129 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2004. ISBN 978-3-540-00539-1 978-1-4612-0979-9. doi: 10.1007/978-1-4612-0979-9. URL http://link.springer.com/10.1007/978-1-4612-0979-9.

Amnon Geifman, Meirav Galun, David Jacobs, and Basri Ronen. On the Spectral Bias of Convolutional Neural Tangent and Gaussian Process Kernels. *Advances in Neural Information Processing Systems*, 35:11253–11265, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/48fd58527b29c5c0ef2cae43065636e6-Abstract-Conference.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL https://www.nature.com/articles/s42256-020-00257-z. Number: 11 Publisher: Nature Publishing Group.

Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: NNGP and NTK for deep attention networks, June 2020. URL http://arxiv.org/abs/2006.10540. arXiv:2006.10540 [cs, stat].

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

A. Krizhevsky. Learning Multiple Layers of Features from Tiny Images. 2009. URL https://www.semanticscholar.org/paper/Learning-Multiple-Layers-of-Features-from-Tiny-Krizhevsky/5d90f06bb70a0a3dced62413346235c02b1aa086.

Hermann König. *Eigenvalue Distribution of Compact Operators*, volume 16 of *Operator Theory: Advances and Applications*. Birkhäuser, Basel, 1986. ISBN 978-3-0348-6280-6 978-3-0348-6278-3. doi: 10.1007/978-3-0348-6278-3. URL http://link.springer.com/10.1007/978-3-0348-6278-3.

Clarissa Lauditi, Blake Bordelon, and Cengiz Pehlevan. Adaptive kernel predictors from feature-learning infinite limits of neural networks, September 2025. URL http://arxiv.org/abs/2502.07998. arXiv:2502.07998 [cs].

Itay Lavie, Guy Gur-Ari, and Zohar Ringel. Towards Understanding Inductive Bias in Transformers: A View From Infinity. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 26043–26069. PMLR, July 2024. URL https://proceedings.mlr.press/v235/lavie24a.html. ISSN: 2640-3498.

Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep Neural Networks as Gaussian Processes. February 2018. URL https://openreview.net/forum?id=B1EA-M-0Z.

Jaehoon Lee, Samuel S. Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. Finite Versus Infinite Neural Networks: an Empirical Study, September 2020. URL http://arxiv.org/abs/2007.15801. arXiv:2007.15801 [cs, stat].

Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Phys. Rev. X*, 11:031059, Sep 2021. doi: 10.1103/PhysRevX.11.031059. URL https://link.aps.org/doi/10.1103/PhysRevX.11.031059.

Cong Ma, Reese Pathak, and Martin J. Wainwright. Optimally tackling covariate shift in RKHS-based nonparametric regression, June 2023. URL http://arxiv.org/abs/2205.02986. arXiv:2205.02986.

Siyuan Ma and Mikhail Belkin. Kernel machines that adapt to GPUs for effective large batch training, March 2019. URL http://arxiv.org/abs/1806.06144. arXiv:1806.06144 [stat].

Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. Gaussian Process Behaviour in Wide Deep Neural Networks, August 2018. URL http://arxiv.org/abs/1804.11271. arXiv:1804.11271 [cs, stat].

Theodor Misiakiewicz. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression, April 2022. URL http://arxiv.org/abs/2204.10425. arXiv:2204.10425 [math].

Radford M. Neal. Priors for Infinite Networks. In Radford M. Neal (ed.), *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics, pp. 29–53. Springer, New York, NY, 1996. ISBN 978-1-4612-0745-0. doi: 10.1007/978-1-4612-0745-0_2. URL https://doi.org/10.1007/978-1-4612-0745-0_2.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 7. Granada, 2011. URL http://research.google.com/pubs/archive/37648.pdf. Issue: 5.

Eshaan Nichani, Alex Damian, and Jason D. Lee. How Transformers Learn Causal Structure with Gradient Descent. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 38018–38070. PMLR, July 2024. URL https://proceedings.mlr.press/v235/nichani24a.html. ISSN: 2640-3498.

Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Jiri Hron, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-dickstein. Bayesian Deep Convolutional Networks with Many Channels are Gaussian Processes. September 2018. URL https://openreview.net/forum?id=B1g30j0qF7.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context Learning and Induction Heads, September 2022. URL http://arxiv.org/abs/2209.11895. arXiv:2209.11895 [cs].

Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010. ISSN 1558-2191. doi: 10.1109/TKDE.2009.191. URL https://ieeexplore.ieee.org/document/5288526/?arnumber=5288526. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. October 2023. URL https://openreview.net/forum?id=aN4Jf6Cx69.

Noa Rubin, Inbar Seroussi, and Zohar Ringel. Grokking as a First Order Phase Transition in Two Layer Networks. October 2023. URL https://openreview.net/forum?id=3ROGsTX3IR.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. One-layer transformers fail to solve the induction heads task, August 2024. URL http://arxiv.org/abs/2408.14332. arXiv:2408.14332 [cs].

Meyer Scetbon and Zaid Harchaoui. A Spectral Analysis of Dot-product Kernels. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 3394–3402. PMLR, March 2021. URL https://proceedings.mlr.press/v130/scetbon21b.html. ISSN: 2640-3498.

Inbar Seroussi, Gadi Naveh, and Zohar Ringel. Separation of scales and a thermodynamic description of feature learning in some CNNs. *Nature Communications*, 14(1):908, February 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36361-y. URL https://www.nature.com/articles/s41467-023-36361-y. Number: 1 Publisher: Nature Publishing Group.

Shai Shalev-Shwartz, Ohad Shamir, and Shaked Shammah. Failures of Gradient-Based Deep Learning. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3067–3075. PMLR, July 2017. URL https://proceedings.mlr.press/v70/shalev-shwartz17a.html. ISSN: 2640-3498.

B. W. Silverman. Spline Smoothing: The Equivalent Variable Kernel Method. *The Annals of Statistics*, 12 (3):898–916, 1984. ISSN 0090-5364. URL https://www.jstor.org/stable/2240968. Publisher: Institute of Mathematical Statistics.

James B. Simon, Madeline Dickens, Dhruva Karkada, and Michael Deweese. The Eigenlearning Framework: A Conservation Law Perspective on Kernel Ridge Regression and Wide Neural Networks. *Transactions on Machine Learning Research*, February 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=FDbQGCAViI.

Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 36, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/58692a1701314e09cbd7a5f5f3871cc9-Abstract-Conference.html.

Peter Sollich and Christopher Williams. Using the Equivalent Kernel to Understand Gaussian Process Regression. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper/2004/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Wu-Ki Tung. *Group Theory in Physics: An Introduction to Symmetry Principles, Group Representations, and Special Functions in Classical and Quantum Physics*. WORLD SCIENTIFIC, August 1985. ISBN 978-9971-966-57-7 978-981-238-498-0. doi: 10.1142/0097. URL http://www.worldscientific.com/worldscibooks/10.1142/0097.

Johannes von Oswald, Eyvind Niklasson, Maximilian Schlegel, Seijin Kobayashi, Nicolas Zucchet, Nino Scherrer, Nolan Miller, Mark Sandler, Blaise Agüera y Arcas, Max Vladymyrov, Razvan Pascanu, and João Sacramento. Uncovering mesa-optimization algorithms in Transformers, September 2023. URL http://arxiv.org/abs/2309.05858. arXiv:2309.05858 [cs].

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747. arXiv: 1708.07747 tex.bibsource: dblp computer science bibliography, https://dblp.org tex.timestamp: Mon, 13 Aug 2018 16:47:27 +0200.

Lechao Xiao. Eigenspace Restructuring: A Principle of Space and Frequency in Neural Networks. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pp. 4888–4944. PMLR, June 2022. URL https://proceedings.mlr.press/v178/xiao22a.html. ISSN: 2640-3498.

Greg Yang. Wide Feedforward or Recurrent Neural Networks of Any Architecture are Gaussian Processes. September 2019. URL https://openreview.net/forum?id=BylLyHHlUS.

Greg Yang. Tensor Programs II: Neural Tangent Kernel for Any Architecture, November 2020. URL http://arxiv.org/abs/2006.14548. arXiv:2006.14548 [stat].

Greg Yang and Etai Littwin. Tensor Programs IIb: Architectural Universality Of Neural Tangent Kernel Training Dynamics. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 11762–11772. PMLR, July 2021. URL https://proceedings.mlr.press/v139/yang21f.html. ISSN: 2640-3498.

Greg Yang and Hadi Salman. A Fine-Grained Spectral Perspective on Neural Networks, April 2020. URL http://arxiv.org/abs/1907.10599. arXiv:1907.10599 [cs, stat].

Wenqian Ye, Luyang Jiang, Eric Xie, Guangtao Zheng, Yunsheng Ma, Xu Cao, Dongliang Guo, Daiqing Qi, Zeyu He, Yijun Tian, Megan Coffee, Zhe Zeng, Sheng Li, Ting-hao, Huang, Ziran Wang, James M. Rehg, Henry Kautz, and Aidong Zhang. The Clever Hans Mirage: A Comprehensive Survey on Spurious Correlations in Machine Learning, October 2025. URL http://arxiv.org/abs/2402.12715. arXiv:2402.12715 [cs] version: 4.

A Proof of The Main Theorem

The main idea of the proof builds upon the fact that while the eigendecomposition of the operator

$$\hat{K}_q \phi_i(x) = \int k(x, x')\phi_i(x')q(x')dx' = \lambda_i \phi_i(x) \tag{24}$$

depends on the measure q, the representation of the kernel found by Mercer decomposition holds for all datasets within the support on q. We then use a known Mercer decomposition on q to represent the kernel function that is sampled in the predictor $\hat{f}_D$ from D.

The usual GPR/KRR predictor on the dataset D is given by Eq.equation 1. Using Mercer's theorem (König, 1986) to decompose the kernel function to eigenfunctions *on* q, the predictor can be written as

$$\begin{aligned} \hat{f}_D(x) &= \sum_{\nu\rho=1}^{P} \sum_{i=1}^{\infty} \phi_i(x) \lambda_i \phi_i(x_\nu) \left[K + I\sigma^2\right]^{-1}_{\nu\rho} y(x_\rho); \\ \hat{K}_q \phi_i(x) &= \lambda_i \phi_i(x). \end{aligned} \tag{25}$$

We now project the predictor onto the target feature ϕ_t with the inner product $\langle \cdot, \cdot \rangle_q$ defined with $q(x)$ as a weighting function (such that $\langle f, g \rangle_q = \mathbb{E}_{x\sim q}[f(x)g(x)]$)

$$\begin{aligned} \left\langle \phi_t, \hat{f}_D \right\rangle_q &= \sum_{\nu\rho} \sum_{i=1}^{\infty} \langle \phi_t, \phi_i \rangle_q \lambda_i \phi_i(x_\nu) \left[K + I\sigma^2\right]^{-1}_{\nu\rho} y(x_\rho) \\ &= \lambda_t \sum_{\nu\rho} \phi_t(x_\nu) \left[K + I\sigma^2\right]^{-1}_{\nu\rho} y(x_\rho), \end{aligned} \tag{26}$$

where the inner product can be carried out immediately based on the orthonormality of $\{\phi_i\}_i$ w.r.t. the inner product $\langle \cdot, \cdot \rangle_q$. We may now use Cauchy-Schwartz inequality to bound the inner product on D, given by the summation on the index ν,

$$\begin{aligned} \left| \left\langle \phi_t, \hat{f}_D \right\rangle_q \right| \leq \frac{\lambda_t}{\sigma^2} &\sqrt{\sum_{\nu} \phi_t^2(x_\nu)} \\ &\cdot \sqrt{\sum_{\mu} \left(\sum_{\rho} \left[\sigma^{-2}K + I\right]^{-1}_{\mu\rho} y(x_\rho) \right)^2}. \end{aligned} \tag{27}$$

Lastly, since $\left[\left(\sigma^{-2}K + I\right)\right]^{-1}$ is weakly contracting we can bound the result from above by

$$\begin{aligned} \left| \left\langle \phi_t, \hat{f}_D \right\rangle_q \right| &\leq \sigma^{-2}\lambda_t \sqrt{\left(\sum_{\nu} \phi_t(x_\nu) \phi_t(x_\nu) \right) \left(\sum_{\mu} y_\mu y_\mu \right)} \\ &= \sigma^{-2}\lambda_t P \sqrt{\mathbb{E}_{x\sim D}\left[\phi_t^2(x)\right] \mathbb{E}_{x\sim D}\left[y^2(x)\right]}. \end{aligned} \tag{28}$$

plugging this result into the definition of x-learnability (Eq. equation 6) and noting we can choose the sign of $\langle \phi_t, y \rangle$ to be positive without loss of generality we arrive at the result Eq. equation 7 in Thm. 3.1. From here the corollary is a straightforward progress: To quantify whether the feature is learned or not we require that the learnability is ϵ close to perfect $\mathcal{L}_i^{D,q} \stackrel{!}{=} 1 - \epsilon$ yielding the result in Eq.8.

B Comparing Learnability Bound to Our Cross-Dataset Learnability Bound

Our bound equation 7 can be compared with a common learnability measure at the EK limit that can be found only given a solution to the generically intractable eigenvalue problem on the data distribution p.

Expanding the learnability in Eq. equation 4 away from 1 we find

$$\mathcal{L}_i = \frac{\eta_i}{\eta_i + \sigma^2/P} = \frac{\eta_i P}{\sigma^2} + O\left(\left(\frac{\eta_i P}{\sigma^2}\right)^2\right) \leq \frac{\eta_i P}{\sigma^2}, \tag{29}$$

where the last inequality arises because the learnability is a concave function of P. Like Eq. equation 7, Eq. equation 29 is a necessary condition for learnability but it uses a different pair of eigenvalue-eigenfunction η_t, ψ_t. ψ_t is guaranteed to be normalized as the eigendecomposition was carried out with the same probability measure as the dataset. From this perspective, the bound can be expected to be tight when $P \ll \eta_i^{-1}\sigma^2$ and less tight as P grows larger. This is coherent with the view of our bound as an *necessary* condition for learnability, rather than an exact prediction of the learnability. The power expansion approximates the learnability well when far from good learnability and bounds it from above throughout training.

C Additional Results

Here we repeat the same experiments as in the main text for SVHN, MNIST, and Fashion MNIST. We start with the experiments in Figure 2 and move to Figure 3

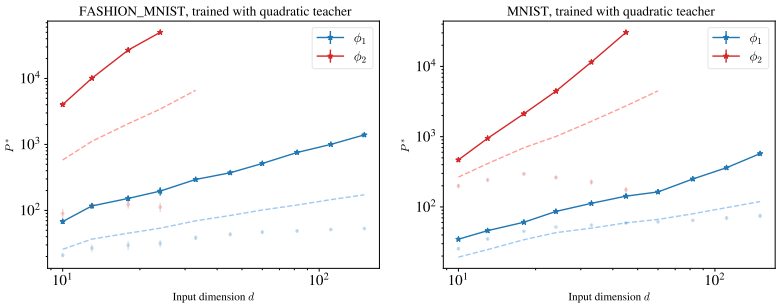

Figure 4: (same figure as Fig. 2, Fashion MNIST (left) and MNIST(right))

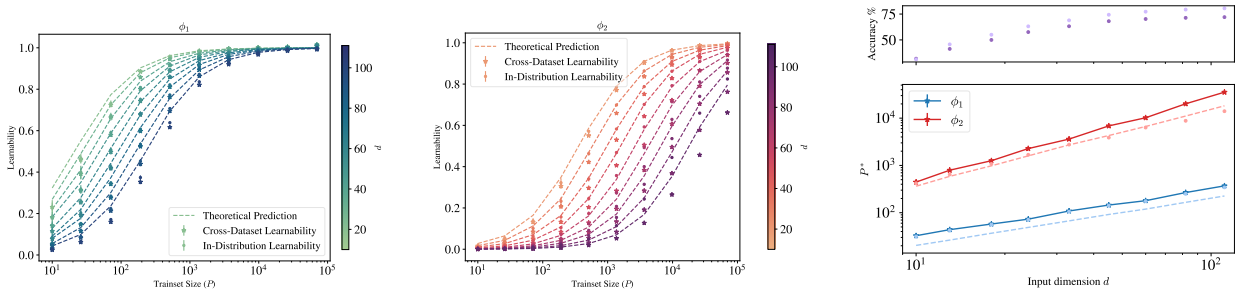

Figure 5: (same figure as Fig. 3, but for SVHN)

We also include here results for synthetic targets made out of random $\ell = 1, 2, 4$ eigenfunctions. See Figure 8 for sample complexity on non-whitened data and Figures 9, 10,11,12 for learning curves on whitened data for CIFAR-10, SVHN, Fashion MNIST, and MNIST, respectively.

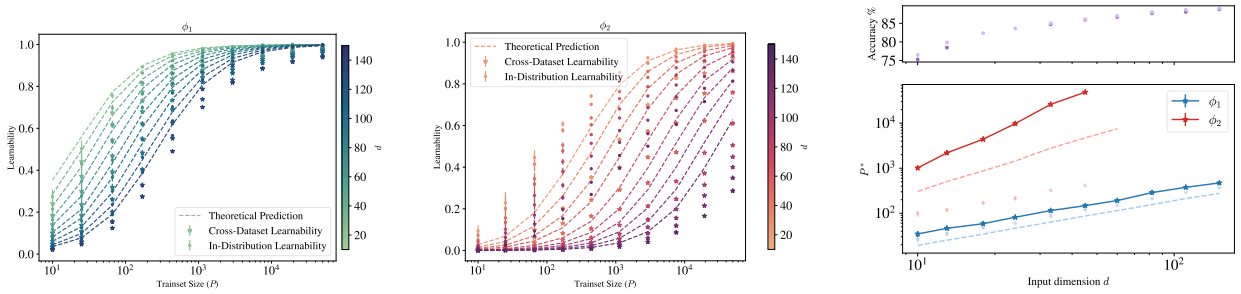

Figure 6: (same figure as Fig. 3, but for Fashion MNIST)

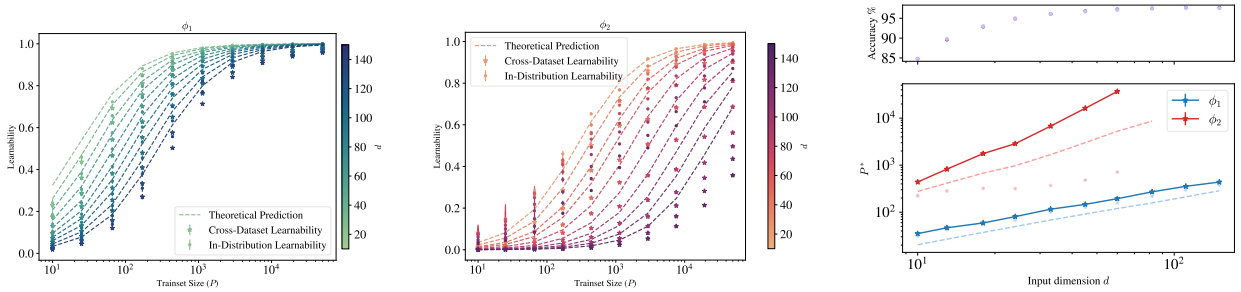

Figure 7: (same figure as Fig. 3, but for MNIST)

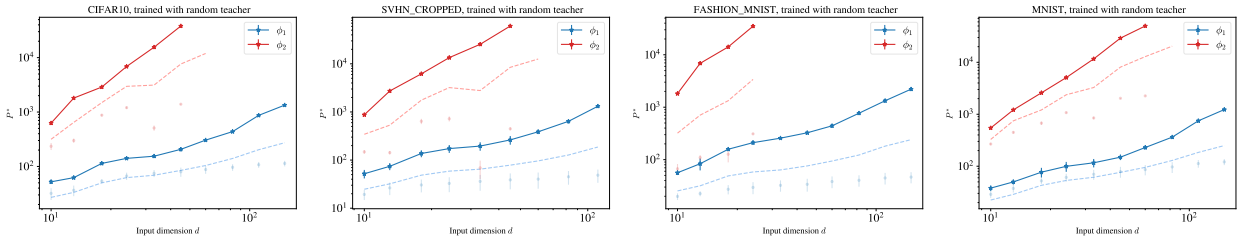

Figure 8: (same figure as Fig. 2, for CIFAR-10 (left), SVHN (middle-left) Fashion MNIST (middle-right), and MNIST (right) with a random teacher as the target.

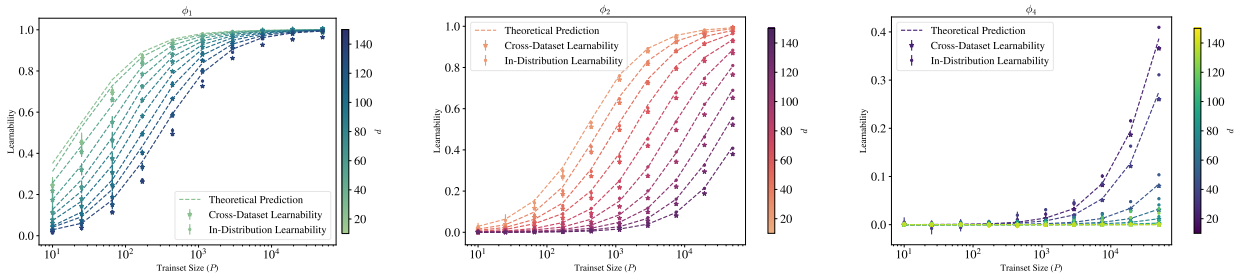

Figure 9: (same figure as Fig. 3, for CIFAR-10 with a random teacher as the target.

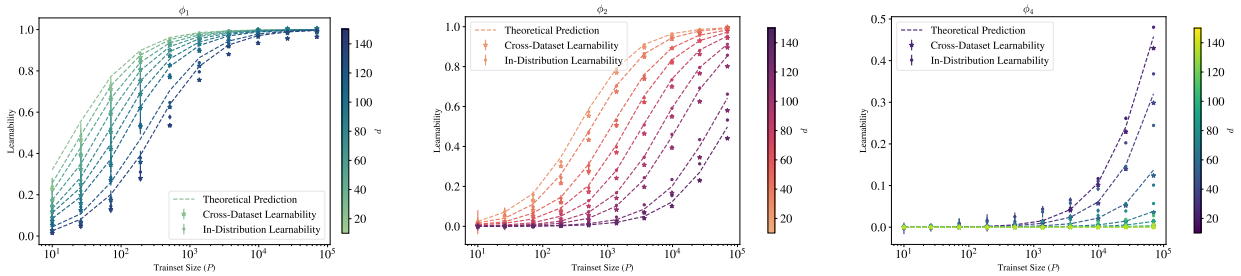

Figure 10: (same figure as Fig. 3, for SVHN with a random teacher as the target.

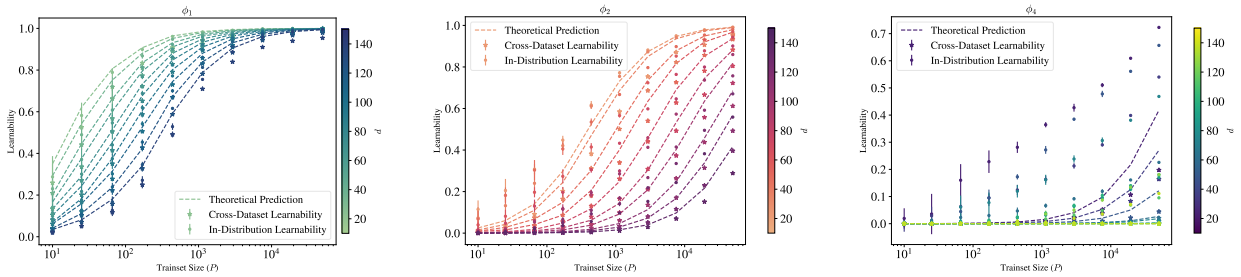

Figure 11: (same figure as Fig. 3, for Fashion MNIST with a random teacher as the target.

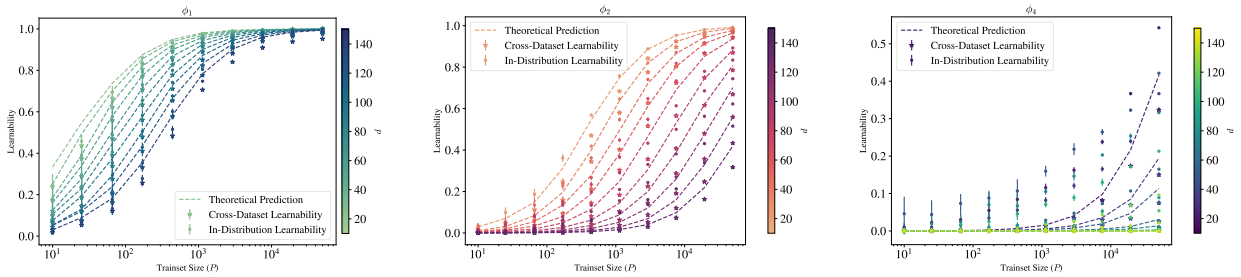

Figure 12: (same figure as Fig. 3, for MNIST with a random teacher as the target.

D PCA Reconstruction Samples

In Figure 13 we visualize the results of using PCA for dimensional reduction for CIFAR-10, Fashion MNIST and MNIST.

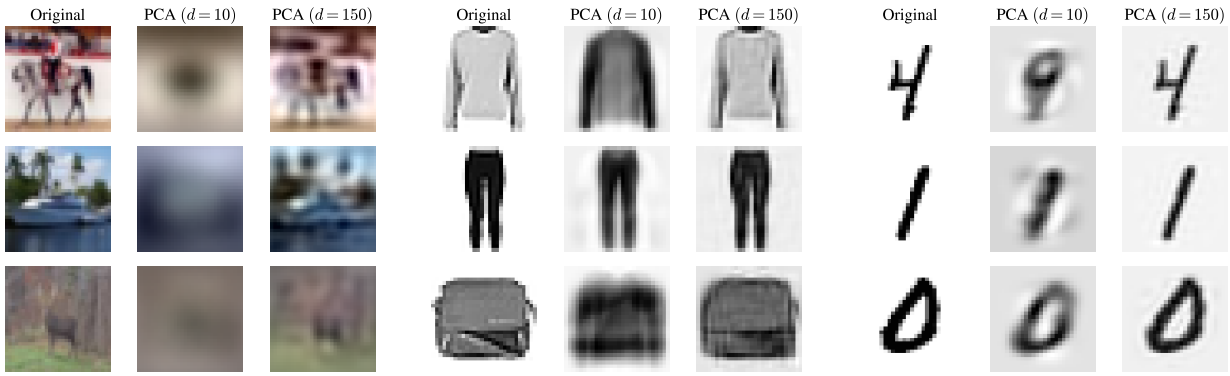

Figure 13: PCA reconstruction of 3 samples from CIFAR-10, Fashion MNIST, and MNIST for the lower and and upper end of the experiments ($d = 10, 150$).

E Multi-spectral extensions

In the main text, we bound the magnitude of the predictor's projection onto the kernel's eigenfunctions on q and compare it to those of the target. Here we consider the case where $y(x)$ is highly multispectral and receives contributions from a large, potentially infinite, number of ϕ_t modes. In such a scenario, $\langle \phi_t, y \rangle_q^2$ would scale inversely with the number of dominant modes whereas y^2 appearing in our bound would remain $O(1)$. As a consequence, the bound may become very loose as is the case where the data distribution and q match, where our bound would be $O\left(\langle \phi_t, y \rangle_q^{-2}\right)$ off the EK result.

Here we extend our bound to certain multispectral circumstances. Specifically, let us assume that $y(x)$ can be written as $y_<(x) + y_>(x)$ such that $y_>(x)$ is spanned by $\{\phi_i(x)\}_i$ having $\lambda_i \leq \lambda_>$. Furthermore let us assume $|y|^2 \propto |y_>(x)|^2 \propto O(1)$ and $|y_>(x)|_q^2 \geq O(1)$. As we argue next in such cases we essentially derive a similar bound with $\lambda_>$ playing the role of λ_t. Specifically we consider $|\langle \hat{f}_D(x), y_>(x) \rangle|_q$ given by

$$
\begin{aligned}
|\langle \hat{f}_D(x), y_>(x) \rangle|_q &= |\int dx q(x) y_>(x) \sum_{\mu\nu} K(x, x_\mu)[K + I\sigma^2]^{-1}_{\mu\nu} y_\nu| = |\sum_{t \geq t_>} \lambda_t \langle \phi_t, y \rangle_q \phi_t(x_\mu)[K + I\sigma^2]^{-1}_{\mu\nu} y_\nu| \\
&= |[\sum_{t \geq t_>} \lambda_t \langle \phi_t, y \rangle_q \vec{\phi}_t]^T [K + I\sigma^2]^{-1} \vec{y}| \leq \sqrt{||[\sum_{t \geq t_>} \lambda_t \langle \phi_t, y \rangle_q \vec{\phi}_t]||^2 ||[K + I\sigma^2]^{-1} \vec{y}||^2} \\
&\leq \sigma^{-2} \sqrt{||[\sum_{t \geq t_>} \lambda_t \langle \phi_t, y \rangle_q \vec{\phi}_t]||^2 ||\vec{y}||^2} = \lambda_> \sigma^{-2} \sqrt{\sum_\mu \left[\sum_{t \geq t_>} \frac{\lambda_t}{\lambda_>} \langle \phi_t, y \rangle_q \phi_t(x_\mu) \right]^2 ||\vec{y}||^2}
\end{aligned} \tag{30}
$$

Notably, if all λ_t's are degenerate, we retrieve our previous bound with the feature being $\sum_t \langle \phi_t, y \rangle_q \phi_t(x)$. More generally, we need to average the feature $\sum_t \frac{\lambda_t}{\lambda_>} \langle \phi_t, y \rangle_q \phi_t(x)$ squared over the training set.

Here we suggest several ways of treating this latter average. In some cases, where we have good control of all λ_t's and ϕ_t's, we may know how to bound this last quantity directly. Alternatively, we may write

$$\sum_{\mu}\left[\sum_{t\geq t_>}\frac{\lambda_t}{\lambda_>}\left\langle\phi_t,y\right\rangle_q\phi_t(x_\mu)\right]^2 = \sum_{t,p\geq t_>}\frac{\lambda_t}{\lambda_>}\frac{\lambda_p}{\lambda_>}\left\langle\phi_t,y\right\rangle_q\left\langle\phi_u,y\right\rangle_q\left[\vec{\phi}_t^T\vec{\phi}_u\right] \tag{31}$$

and argue that in a typical scenario, $\vec{\phi}_t^T\vec{\phi}_u$ with $u\neq t$ would be smaller and also sum up incoherently. The diagonal contributions would thus be the dominant ones. Further assurance may be obtained by sampling $\lambda_t\lambda_u\left\langle\phi_t,y\right\rangle_q\left\langle\phi_u,y\right\rangle_q\vec{\phi}_t^T\vec{\phi}_u$ and verifying that off-diagonal contributions are indeed smaller and incoherent. Considering the diagonal contribution alone we obtain

$$|\langle\hat{f}_D(x),y_>(x)\rangle|_q \tag{32}$$

$$\leq P\lambda_>(1-\epsilon)\sigma^{-2}\sqrt{\sum_{t\geq t_>}\left(\frac{\lambda_t}{\lambda_>}\right)^2\left\langle\phi_t,y\right\rangle_q^2\mathbb{E}_D[\phi_t^2]\mathbb{E}_D[y^2]} \tag{33}$$

which recalling $\lambda_t/\lambda_>\leq 1$ and $\sum_{t\geq t_>}[\left\langle\phi_t,y\right\rangle_q]^2=O(1)$ yields a similar result to before.

Otherwise, we can take a worst-case scenario in which all $\phi_t(x_\mu)$ contribute coherently to the sum [indeed think of ϕ_t as 1d Fourier modes, and the training distribution is a delta function at zero, and we wish to learn a delta function of the training set. Our $[K+\sigma^2]^{-1}y\approx\sigma^{-2}y$ estimate would be very poor however the ϕ_t's would all sum coherently around zero]. In this case, we may use Cauchy Schwarz again on the summation over t to obtain

$$|\langle\hat{f}_D(x),y_>(x)\rangle|_q \tag{34}$$

$$\leq\lambda_>\sigma^{-2}(1-\epsilon)\sqrt{\sum_{t\geq t_>}|\left\langle\phi_t,y\right\rangle_q|^2\sum_{\mu,t}\left|\frac{\lambda_t}{\lambda_>}\phi_t(x_\mu)\right|^2||\vec{y}||^2} \tag{35}$$

Consider this as a learnability namely $\langle f,y_>\rangle_q/\langle y_>,y_>\rangle_q$ bearing in mind that $y_>(x)$ has an $O(1)$ norm (or more) also under q. Scaling-wise, we may thus remove the $\sum_{t\geq t_>}|\left\langle\phi_t,y\right\rangle_q|^2$ factor. Doing so we retrieve our previous bound, only with $\sum_t\frac{\lambda_t^2}{\lambda_>^2}E_D(\phi_t^2)$ instead of just $E_D(\phi_t^2)$.

Notably, since in our normalization all $\lambda_t<1$, $\sum_t\lambda_t^2$ decays to zero quicker than λ_t and hence by the finiteness of the trace yields a finite number even if $y_>$ contains an infinite amount of features.

F Measures of Learnability & Learning Parity with a Correlated Dataset

We begin by recovering a known result - learning parity from a uniform distribution on the hyper-cube with an FCN-GP is hard (Yang & Salman, 2020; Simon et al., 2023) and extend it to general distributions on the hyper-cube. We then present an example where the data measure has very low entropy and probes only a small low dimensional space of the hyper-cube. In that example, a function that mostly agrees with parity on the data can be learned easily, but learning a predictor that generalizes out-of-distribution remains hard. We discuss what learning a function can mean when considering different distributions, potentially including out-of-distribution test points, and suggest maximal entropy distributions as a reasonable measure upon which learnability can be gauged. Finally, we compare the suggested measure of learnability to a familiar one.

Consider learning parity on the $\vec{x}\in\{-1,1\}^d$ hyper-cube using FCNs. Here $\vec{x}$ is drawn from an arbitrary, possibly correlated, measure on the hyper-cube, and the target function is parity $y=\Pi_{i=1}^d x_i$ with no added noise.

Learning parity with noise is believed to require P scaling exponentially with d (Blum et al., 2000). Parity without noise can be learned with $O(d)$ samples using Gaussian elimination and relations between boolean

operations and algebra in Z_2 fields (Blum et al., 2000). A GP, however, involves a larger hypothesis class including real rather than boolean variables. It also seems highly unlikely that it could learn from examples an $O(d^3)$-algorithm such as Gaussian elimination. In the case of a uniform measure, an FCN GP is known to require P^* which is exponential in d (Simon et al., 2023) to learn parity. It is reasonable to assume that a generic non-uniform measure would not reduce the complexity of this task, however, we are not aware of any existing bounds applicable to this broader case.

To this end, we take as an ideal distribution q a uniform distribution on the sphere containing the corners of the hyper-cube. We turn to calculate or bound the different elements in Eq. 8. First, we require the $\phi_t(x)$ associated with parity. Under q, any FCN kernel is diagonal in the basis of hyperspherical harmonics. The latter can be described as rank-n homogeneous harmonic polynomials (Frye & Efthimiou, 2012). Each rank constitutes an irreducible representation of the rotation group (Tung, 1985; Fulton & Harris, 2004). As a consequence, each rank-n polynomial is a kernel eigenfunction whose eigenvalue depends only on n. We may conclude our eigenfuction of interest ϕ_t is $\hat{n}^{-1/2}y$ where $\hat{n}$ is a normalization factor w.r.t. to the measure q. Having identified the eigenfunction we would like to estimate its eigenvalue. One can show that there are $N(n,d) = \frac{2n+d-2}{n}\binom{n+d-3}{n-1}$ polynomials at given n, d (Frye & Efthimiou, 2012). Considering parity, it is a harmonic homogeneous polynomial of rank d and consequently part of a $N(d,d)$-degenerate subspace of any FCN kernel with eigenvalue λ_d. Noting that $\mathbb{E}_{x\sim q}[k(x,x)]$ equal the sum of all eigenvalues and that eigenvalues are positive, we obtain $\lambda_d \leq \mathbb{E}_{x\sim q}[k(x,x)]/N(d,d)$.

Following its appearance in the numerator and denominator and the fact that parity squares to 1 on the hyper-cube we find

$$\begin{aligned} P^* &\geq \sigma^2 \lambda_t^{-1}(1-\epsilon)\frac{\mathbb{E}_{x\sim q(x)}[y(x)y(x)]}{\sqrt{\mathbb{E}_{x\sim D}\left[y^2\left(x\right)\right]\mathbb{E}_{z\sim D}\left[y^2(x)\right]}} \\ &\geq \sigma^2(1-\epsilon)\hat{n}N(d,d)/\mathbb{E}_{x\sim q}[k(x,x)] \end{aligned} \tag{36}$$

We calculate the normalization factor $\hat{n}$ in Appendix G, and quote the result here

$$\hat{n} = \frac{2^{-d}d^d\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{3d}{2}\right)}. \tag{37}$$

Next, we use Stirling's formula for an asymptotic expansion for $d \gg 1$

$$P^* \geq \frac{\sigma^2(1-\epsilon)}{\mathbb{E}_{x\sim q}[k(x,x)]}\sqrt{\frac{3^3}{2^6\pi d}}\left(\frac{4e}{3^{3/2}}\right)^d \tag{38}$$

We find that in high-dimension $d \gg 1$, given any training dataset on the hyper-cube, the sample complexity of parity for a FCN GP is at least exponential in d.

An extreme yet illustrative case to consider is a predominantly correlated measure on the hyper-cube which forces all x_i's to be equal $p_1(x)$ plus a uniform measure namely $p(x) = (1-\alpha)p_1(x) + \alpha q(x)$ with $\alpha \ll 1$. For even d on such measure, y would be well approximated by a constant on D. While a constant function is easily learnable and may appear to yield a low test loss on $p(x)$, it grossly differs from the true target on q. Thus, in terms of generalization, the bound is useful for gauging the generalization properties on the analytically tractable measure q, rather than on the empirical measure ($p(x)$) on which the GP would seem to perform very well. In scenarios where a complex feature (such as parity) on the ideal measure q is well approximated by much simpler ones (e.g. a constant) on the training measure one should be wary of associating an unlearnable target, in the sense of our bound, with poor generalization performance within the training distribution.

Measuring learnability on q. As the example above manifests, it is crucial to note our bound essentially bounds the learnability from the perspective of q as can be seen in eq.equation 6. That means even for $P < P^*$ the model can perform well on the training dataset and even test datasets. Prominently, this can happen when training (resp. testing) on a low-entropy distribution that can collapse complicated functions onto simpler ones. In this sense, symmetric measures can be seen as maximal entropy distributions under certain constraints; they thus suggest themselves as a ground upon which OOD generalization can be predicted. For more complicated domains it remains an open question which dataset "truly" reflects the feature $\phi_t(x)$ though we expect these differences to be small in practice.

G Calculating the normalization factor for parity & FCN

To calculate this normalization factor we first extend it to be a function of the sphere radius namely

$$\begin{aligned}\hat{n}[r] &= \left(\frac{r^{d-1}2\pi^{d/2}}{\Gamma(d/2)}\right)^{-1}\int_{R^d} d^d x\delta(|x|-r)\prod_{i=1}^{d} x_i^2 \\ &=: \left(\frac{r^{d-1}2\pi^{d/2}}{\Gamma(d/2)}\right)^{-1} n[r],\end{aligned} \tag{39}$$

where the first factor is the hypersphere surface area in d dimensions under the assumption of even d.

While we are interested in $n[r=\sqrt{d}]$ we instead first look at

$$N[s] = \int_0^\infty dr e^{-sr^2} n[r] = \int_0^\infty d(r^2) e^{-sr^2}\frac{n[r]}{2r} \tag{40}$$

Notably $N[s]$ is then the Laplace transform of $n[r]/(2r)$ viewed as a function of r^2. Calculating it based on the second expression amounts to independent Gaussian integrations and yields

$$N[s] = \int_{R^d} d^d x e^{-s|x|^2}\prod_i x_i^2 = \left(\frac{\pi}{2^2 s^3}\right)^{d/2} \tag{41}$$

Inverting this Laplace transform we obtain

$$\begin{aligned}\frac{n[r]}{2r} &= \frac{2^{-d}\pi^{d/2}r^{3d-2}}{\Gamma\left(\frac{3d}{2}\right)} \Rightarrow n[r] = \frac{2^{-(d-1)}\pi^{d/2}r^{3d-1}}{\Gamma\left(\frac{3d}{2}\right)} \\ \hat{n}[r] &= \left(\frac{r^{d-1}2\pi^{d/2}}{\Gamma(d/2)}\right)^{-1} n[r] = \frac{2^{-d}r^{2d}\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{3d}{2}\right)}\end{aligned} \tag{42}$$

H Eigenvalues for copying heads

For the data distribution $q(X)$ we may use the results of (Lavie et al., 2024) to characterize the eigenvalues and eigenvectors of the NNGP/NTK kernel of a transformer. The approach relies on the permutation symmetry between tokens in the same sample and uses representation theory to upper-bound eigenvalues based on their degeneracy. In the case of q from the main text, we have a permutation symmetry in sequence space and an additional permutation symmetry in vocabulary space, since it is uniformly distributed by assumption. This makes the transformer kernel *operator* symmetric under the action of the symmetric group $S_L \times S_V$ in both sequence and vocab. space.

Eigenvalues that belong to the same irreducible representation (irrep) R and degenerate subspace V_R^i can be bounded by the kernel's trace over the dimension of the irrep

$$\lambda_{V_R^i} \leq \frac{\mathbb{E}_{x\sim q}[k(x,x)]}{\dim_R}, \tag{43}$$

where $\lambda_{V_R^i}$ is the $\dim_R$-fold degenerate eigenvalue of the subspace V_R^i, $E_{x\sim q}[k(x,x)]$ is the kernel's trace and $\dim_R$ is the dimension of the irrep. We are thus interested in decomposing the target function into irreps and finding a component that lives in a high-dimensional irrep in order to find a strong lower bound.

As shown in (Lavie et al., 2024), linear functions (such as a copying head) decompose into two irreps of the symmetric group, "trivial" and "standard". The trivial irrep is just a sum over all entries and is an irrep of dimension 1. Since linear functions contain only two irreps, the standard irrep contains all the other functions, so it is of dimension $L-1$ (or $V-1$ for vocab.). Following the same logic, to find the component that lives in the standard irrep, we can simply project out the trivial component - subtract the sum over

all coordinates. Subtracting the sums over the sequence and vocabulary indices from the definition of the copying head function, we get

$$
\begin{aligned}
P &\coloneqq (I - P^V_{\text{trivial}})(I - P^L_{\text{trivial}}); \quad P^V_{\text{trivial}} f^a_i \coloneqq \frac{1}{V}\sum_i f^a_i; \quad P^L_{\text{trivial}} f^a_i \coloneqq \frac{1}{L}\sum_a f^a_i \ ; \\
\vec{\phi}^a_t(X) &= \frac{1}{z}\left(\vec{x}^{\,a-1} - \frac{1}{L}\sum_{b=1}^{L}\vec{x}^{\,b}\right); \\
z &\coloneqq \sqrt{L-1}; \\
\mathbb{E}_{X\sim q}[\vec{\phi}^a_t(X)\vec{y}^a(X)] &= z.
\end{aligned}
\tag{44}
$$

It is part of a $(L-1)(V-1)$ degenerate space of the standard irrep of both the vocabulary and sequence permutation symmetry

$$\dim_t = (L-1)(V-1) \tag{45}$$

hence the eigenvalues can be bounded by

$$\lambda_t \leq \frac{\mathbb{E}_{x\sim q}[k(x,x)]}{(L-1)(V-1)}. \tag{46}$$

Combining this result with the results in the main text and plugging them into our main result in Eq. 8 we find

$$P^* \geq \sigma^2(1-\epsilon)\frac{(L-1)(V-1)}{\mathbb{E}_{X\sim q}[k(X,X)]}\sqrt{(1-L^{-1})\,(1-V^{-1})}. \tag{47}$$