

PREDICTING KERNEL REGRESSION LEARNING CURVES FROM ONLY RAW DATA STATISTICS

Anonymous authors

Paper under double-blind review

ABSTRACT

We study kernel regression with common rotation-invariant kernels on real datasets including CIFAR-5m, SVHN, and ImageNet. We give a theoretical framework that predicts learning curves (test risk vs. sample size) from only two measurements: the empirical data covariance matrix and an empirical polynomial decomposition of the target function f_* . The key new idea is an analytical approximation of a kernel’s eigenvalues and eigenfunctions with respect to an anisotropic data distribution. The eigenfunctions resemble Hermite polynomials of the data, so we call this approximation the *Hermite eigenstructure ansatz* (HEA). We prove the HEA for Gaussian data, but we find that real image data is often “Gaussian enough” for the HEA to hold well in practice, enabling us to predict learning curves by applying prior results relating kernel eigenstructure to test risk. Extending beyond kernel regression, we empirically find that MLPs in the feature-learning regime learn Hermite polynomials in the order predicted by the HEA. Our HEA framework is a proof of concept that an end-to-end theory of learning which maps dataset structure all the way to model performance is possible for nontrivial learning algorithms on real datasets.

1 INTRODUCTION

The quest to understand machine learning is largely motivated by a desire to predict and explain learning behavior in realistic settings. This means that, sooner or later, scientists of machine learning must develop theory that works for real datasets, somehow incorporating task structure into predictions of model performance, optimal hyperparameters, and other objects of interest. This necessity has been the elephant in the room of much of deep learning theory for some time: despite much progress in the study of neural network training and generalization, it has proven difficult to move beyond simplistic models of data and make analytical predictions applicable to real data distributions.

The central difficulty is of course the complexity of real data. There can be no full analytical description of any real data distribution, so it is difficult to see how we might develop mathematical theory that describes how such a dataset is learned. How might we hope to proceed?

One way forward may be to identify a comparatively succinct “reduced description” of a data distribution that characterizes its structure, at least insofar as a particular class of learner is concerned. We would like this reduced description to be sufficient to predict quantities of interest yet minimal enough to be a significant reduction in complexity. Ideally, we would like the theory that makes predictions from this reduced description to be mathematically simple, and we would like the description itself to give some insight into how the class of learner in question sees the data.

In this paper, we present such a reduced description of high-dimensional datasets that is suitable for describing their learning by kernel ridge regression (KRR) with rotation-invariant kernels. We find that just the data covariance matrix $\Sigma := \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$, together with a Hermite decomposition of the target function,² is sufficient to characterize learning by rotation-invariant kernels. We obtain this reduced description, which we term “Hermite eigenstructure,” from a study of Gaussian data, but we nonetheless find it predictive for complex image datasets including CIFAR-5m, SVHN, and ImageNet. From just the covariance matrix, we can predict kernel eigenstructure and learning curves

Code to reproduce all experiments available at [\[url_redacted_for_review\]](#).

²See Section 3.2 and Appendix A for a review of Hermite polynomials.

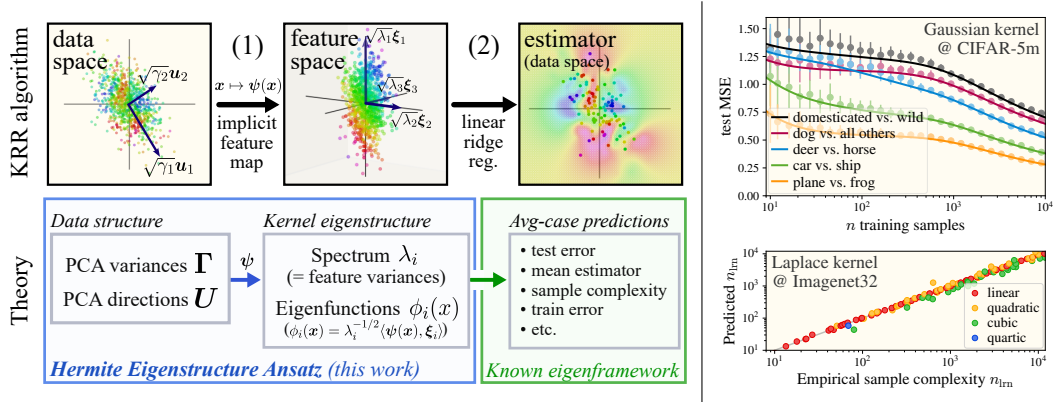


Figure 1: **We provide an end-to-end theory of learning for kernel ridge regression (KRR) that maps minimal statistics of the data distribution to test-time performance.** (Top left.) KRR implicitly consists of two steps: (1) the kernel maps the data to high-dimensional nonlinear features $x \mapsto \psi(x)$, then (2) it fits a linear estimator to these features. Let the PCA covariance in data space be $\mathbb{E}[xx^\top] = \mathbf{U}\mathbf{T}\mathbf{U}^\top$ and let the feature covariance in latent space be $\mathbb{E}[\psi(x)\psi(x)^\top] = \mathbf{\Xi}\mathbf{\Lambda}\mathbf{\Xi}^\top$. (Lower left.) We introduce an ansatz (blue box) that predicts the *feature* covariance statistics $(\mathbf{\Xi}, \mathbf{\Lambda})$ from only the *data* covariance $(\mathbf{U}, \mathbf{T})$ and the functional form of $\psi(\cdot)$. This is sufficient to predict average-case test error using known results (green box). (Top right.) We are able to predict learning curves for KRR on image tasks without requiring omniscient knowledge of the feature statistics (i.e., without ever constructing or diagonalizing a kernel matrix). (Bottom right.) We are able to accurately predict the KRR sample complexity, including constant prefactors, for learning polynomials of the ImageNet dataset. See Figure 3 for additional plots and Appendix D for experimental details.

for synthetic functions. Using the true labels to estimate the target function’s Hermite decomposition, we are additionally able to predict KRR learning curves on real tasks. Unlike previous approaches for predicting learning curves, this method does not require numerically constructing or diagonalizing a kernel matrix to find the kernel eigensystem.

Our approach relies on recent results in the theory of KRR which assert that knowledge of the kernel’s eigenstructure with respect to a data measure is sufficient to predict learning behavior (Sollich (2001); Bordelon et al. (2020); Jacot et al. (2020); Simon et al. (2021)). These works provide a set of equations that map this eigenstructure to predictions of test-time error. Our central observation is that, despite the complexity of high-dimensional datasets and the great variety of rotation-invariant kernels, this kernel eigenstructure is often very close to a simple analytical form expressible in terms of Hermite polynomials in the original data space. We term this claim the “Hermite eigenstructure ansatz,” and we identify a (broad) set of conditions under which it empirically holds.

Concretely, our contributions are as follows:

- We propose the *Hermite eigenstructure ansatz* (Section 4), a closed-form expression for the eigensystem of rotation-invariant kernels on real datasets. We find empirically that it holds to an excellent approximation for real image datasets (Figure 2).
- We prove that the HEA holds in the case of Gaussian data for two limiting cases of the kernel function (Theorems 1 and 2).
- We use the HEA to predict KRR learning curves on CIFAR-5m, SVHN, and ImageNet from only data covariance statistics and a Hermite decomposition of the target function (Figures 1 and 3).
- We empirically find that MLPs in the feature-learning regime learn Hermite polynomials of CIFAR-5m in the same order as the HEA predicts for KRR (Figure 4).

2 RESEARCH CONTEXT AND RELATED WORKS

Kernel models as proxies for neural networks. Our motivation for studying KRR comes from the “neural tangent kernel” (NTK) line of work, which finds suitably parametrized infinite-width networks are equivalent to KRR, and that for MLPs, the kernel function is rotation-invariant (Neal, 1996; Lee et al., 2018; Jacot et al., 2018). Kernel methods have proven useful as models of network dynamics and optimization (Chizat et al., 2019; Du et al., 2019; Bordelon & Pehlevan, 2021).

Learning curves for KRR. Motivated by the NTK, many recent works have converged on a set of equations which predict KRR’s test-time error from the *kernel and task eigenstructure* (Sollich, 2001; Bordelon et al., 2020; Jacot et al., 2020; Simon et al., 2021). This “KRR eigenframework” depends on the kernel’s eigenvalues and eigenfunctions with respect to the data distribution. Our main result is an approximate *analytical* expression for these eigenvalues and eigenfunctions, permitting the inductive bias of KRR to be studied directly in the data space. Appendix C reviews this eigenframework.

Exactly-solved cases of kernel eigenstructure. Exact kernel diagonalizations with machine-learning-relevant kernels are known in many highly-symmetric settings, including stationary kernels on the torus $\mathbb{T}^d$ and rotation-invariant kernels on the sphere $\mathbb{S}^d$ (Mei & Montanari, 2019). Moving to anisotropic domains, Ghorbani et al. (2020) gave the eigenstructure of rotation-invariant kernels when the measure is a “product of spheres” of different radii. The case of a Gaussian kernel on a Gaussian measure was solved exactly by Zhu et al. (1997). Our Hermite eigenstructure ansatz is consistent with all these results and unifies them in a limiting case.

Modeling data with a Gaussian measure. A developing body of literature argues that MLPs’ learning of complex data distributions is similar to the behavior one would see if the data were Gaussian with the same covariance (Goldt et al., 2020; Refinetti et al., 2023). We broadly adopt this lens in the study of KRR and find that, indeed, the data is well-modeled as Gaussian.

Single- and multi-index models. Much recent literature has studied MLPs’ learning of *single- and multi-index functions* which depend only on a rank-one or rank- k projection of the input $\mathbf{x}$ (Dudeja & Hsu, 2018; Bietti et al., 2022; Dandi et al., 2023; Lee et al., 2024; Mousavi-Hosseini et al., 2024). This work partially motivated our study, and the multidimensional Hermite basis we use in this work is a basis of multi-index functions. Prior work in this vein has found that higher-order Hermite polynomials require more samples or gradient steps to learn. Two ways in which we depart from this body of work are that (a) we seek to predict the *value* of the test error (including constant prefactors), not just asymptotics or scaling laws, and (b) we study *anisotropic* data, which allows application of our results to real datasets.

Analytical models of data. Several recent works have proposed theoretical models for the hierarchical structure in image and text data with the aim of understanding neural network performance on such datasets (Cagnetta et al., 2024; Sclocchi et al., 2025; Cagnetta & Wyart, 2024). Our work in this paper is undertaken in a similar spirit.

3 PRELIMINARIES

We will work in a standard supervised setting: our dataset consists of n samples $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ drawn i.i.d. from a measure μ over $\mathbb{R}^d$, and we wish to learn a target function f_* from noisy training labels $\mathbf{y} = \{y_i\}_{i=1}^n$ where $y_i = f_*(\mathbf{x}_i) + \mathcal{N}(0, \epsilon^2)$ with noise level $\epsilon \geq 0$. We will assume with minimal loss of generality that μ has mean zero: $\mathbb{E}_{\mathbf{x} \sim \mu}[\mathbf{x}] = \mathbf{0}$. Once a learning rule returns a predicted function $\hat{f}$, we evaluate its test mean-squared error $\text{MSE}_{\text{te}} = \mathbb{E}_{\mathbf{x} \sim \mu}[(f_*(\mathbf{x}) - \hat{f}(\mathbf{x}))^2] + \epsilon^2$.

We write $\langle g, h \rangle_\mu := \mathbb{E}_{\mathbf{x} \sim \mu}[g(\mathbf{x})h(\mathbf{x})]$ and $\|g\|_\mu^2 := \langle g, g \rangle_\mu$ for the L^2 inner product and norm with respect to μ . We write $(a_i)_{i \in \mathcal{I}}$ to denote an ordered sequence with index set $\mathcal{I}$, and we write only (a_i) when the index set is clear from context.

3.1 KERNEL REGRESSION AND KERNEL EIGENSYSTEMS

KRR is a learning rule specified by a positive-semidefinite “kernel function” $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ and a ridge parameter $\delta \geq 0$. Given a dataset $(\mathcal{X}, \mathbf{y})$, KRR returns the predicted function

$$\hat{f}(\mathbf{x}) = \mathbf{k}_{\mathbf{x}\mathcal{X}}(\mathbf{K}_{\mathcal{X}\mathcal{X}} + \delta \mathbf{I}_n)^{-1} \mathbf{y}, \quad (1)$$

where the vector $[\mathbf{k}_{\mathcal{X}\mathcal{X}}]_i = K(\mathbf{x}, \mathbf{x}_i)$ and matrix $[\mathbf{K}_{\mathcal{X}\mathcal{X}}]_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ contain evaluations of the kernel function. In this paper, we will restrict our attention to two special classes of kernel:

Definition 1 (Rotation-invariant kernel). A kernel function is *rotation-invariant* if it takes the form $K(\mathbf{x}, \mathbf{x}') = K(\|\mathbf{x}\|, \|\mathbf{x}'\|, \mathbf{x}^\top \mathbf{x}')$.

Such a kernel K is called “rotation-invariant” because $K(\mathbf{U}\mathbf{x}, \mathbf{U}\mathbf{x}') = K(\mathbf{x}, \mathbf{x}')$ for any orthonormal matrix $\mathbf{U}$. Many widely-used kernels are rotation-invariant, including the Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = e^{-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{x}'\|^2}$, the Laplace kernel $K(\mathbf{x}, \mathbf{x}') = e^{-\frac{1}{\sigma}\|\mathbf{x}-\mathbf{x}'\|}$, and the Neural Network Gaussian Process (NNGP) kernels and NTKs of infinite-width MLPs. We will be particularly interested in a subset of rotation-invariant kernels which discard the explicit radial dependence:

Definition 2 (Dot-product kernel). A kernel function is a *dot-product kernel* if it takes the form $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}^\top \mathbf{x}')$.

For a dot-product kernel to be positive-semidefinite on all domains, it must admit a Taylor series $K(\mathbf{x}^\top \mathbf{x}') = \sum_{\ell \geq 0} \frac{c_\ell}{\ell!} (\mathbf{x}^\top \mathbf{x}')^\ell$ with nonnegative *level coefficients* $c_\ell \geq 0$ (Schoenberg, 1942). We will find it useful to describe dot-product kernels in terms of their level coefficients $(c_\ell)_{\ell \geq 0}$.

We would like to study arbitrary rotation-invariant kernels, but it is easier to study dot-product kernels, which admit the above series expansion. Fortunately, a rotation-invariant kernel is a dot product kernel when the domain is restricted to a sphere, and if we know that our data has typical norm r , we may approximate the rotation-invariant kernel as the dot-product kernel which matches on $r\mathbb{S}^{d-1}$:

Definition 3 (On-sphere level coefficients). The *on-sphere level coefficients* of a rotation-invariant kernel K at a radius $r > 0$ are the nonnegative sequence coeffs $(K, r) := (c_\ell)_{\ell \geq 0}$ such that

$$K(\mathbf{x}, \mathbf{x}') = \sum_{\ell \geq 0} \frac{c_\ell}{\ell!} (\mathbf{x}^\top \mathbf{x}')^\ell \quad \text{for all } \mathbf{x}, \mathbf{x}' \text{ such that } \|\mathbf{x}\| = \|\mathbf{x}'\| = r. \quad (2)$$

We give the on-sphere level coefficients for various kernels in [Appendix B](#).

An *eigenfunction* of a kernel K with respect to a measure μ is a function ϕ such that $\mathbb{E}_{\mathbf{x}' \sim \mu}[K(\mathbf{x}, \mathbf{x}')\phi(\mathbf{x}')] = \lambda\phi(\mathbf{x})$ for some $\lambda \geq 0$. By Mercer’s Theorem (Mohri et al., 2018, Theorem 6.2), any compact kernel admits a complete basis of orthonormal eigenfunctions with $\langle \phi_i, \phi_j \rangle_\mu = \delta_{ij}$ and may be spectrally decomposed³ as $K(\mathbf{x}, \mathbf{x}') = \sum_i \lambda_i \phi_i(\mathbf{x})\phi_i(\mathbf{x}')$. We will write eigensystem $(\mu, K) = (\lambda_i, \phi_i)_{i=1}^\infty$ to denote the sequence of all eigenpairs, indexed in decreasing eigenvalue order ($\lambda_i \geq \lambda_{i+1}$) unless otherwise specified. It will prove useful to decompose the target function in the kernel eigenbasis as $f_*(\mathbf{x}) = \sum_i v_i \phi_i(\mathbf{x})$, where (v_i) are eigencoeficients.

3.2 HERMITE POLYNOMIALS AS A NATURAL BASIS FOR GAUSSIAN DATA

Throughout, we write $(h_k)_{k \geq 0}$ for the normalized probabilist’s Hermite polynomials. These are the orthogonal polynomials for the standard Gaussian measure, satisfying $\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})}[h_k(\mathbf{x})h_m(\mathbf{x})] = \delta_{km}$. The first few such polynomials are $h_0(x) = 1$, $h_1(x) = x$, $h_2(x) = \frac{1}{\sqrt{2}}(x^2 - 1)$. See [Appendix A](#) for a review of Hermite polynomials.

We can use these 1D Hermite polynomials to construct an orthonormal basis for a multivariate Gaussian measure $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$ with positive-definite covariance $\Sigma \succ \mathbf{0}$. First we diagonalize the covariance as $\Sigma = \mathbf{U}\mathbf{\Gamma}\mathbf{U}^\top$ with orthogonal matrix $\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_d]$ and diagonal matrix $\mathbf{\Gamma} = \text{diag}(\gamma_1, \dots, \gamma_d)$. Then for any multi-index $\alpha \in \mathbb{N}_0^d$, we define

$$h_\alpha^{(\Sigma)}(\mathbf{x}) := \prod_{i=1}^d h_{\alpha_i}(z_i), \quad \text{where} \quad \mathbf{z} = \mathbf{\Gamma}^{-1/2}\mathbf{U}^\top \mathbf{x}. \quad (3)$$

In [Equation \(3\)](#), the elements of α specify the order of the Hermite polynomial along each principal direction of data covariance. These multidimensional Hermite polynomials are orthonormal, satisfying $\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, \Sigma)}[h_\alpha^{(\Sigma)}(\mathbf{x})h_{\alpha'}^{(\Sigma)}(\mathbf{x})] = \delta_{\alpha\alpha'}$.

³Here is an intuitive description of a kernel eigensystem which is shown visually in [Figure 1](#). Any kernel function K may be viewed as an inner product in a high-dimensional feature space: $K(\mathbf{x}, \mathbf{x}') = \langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle$. Consider mapping the dataset into this high-dimensional space and then computing the principal components of $\Sigma_\psi := \mathbb{E}_{\mathbf{x}}[\psi(\mathbf{x})\psi^\top(\mathbf{x})] = \Xi\Lambda\Xi^\top$. Each eigenvalue λ_i is a kernel eigenvalue. The corresponding eigenfunction is a projection onto the i -th principal direction: $\phi_i(\mathbf{x}) = \lambda_i^{-1/2}\langle \psi(\mathbf{x}), \xi_i \rangle$.

In the next section, we assert that this naïve guess of basis is in fact often close to the *true basis of kernel eigenfunctions* for synthetic and real datasets.

4 THE HERMITE EIGENSTRUCTURE ANSATZ: THEORY AND EXPERIMENT

Prior work has shown that predicting kernel regression learning curves boils down to understanding the kernel eigensystem. With this as motivation, we are ready to introduce our primary mathematical object: an *explicit functional form* for the kernel eigensystem, suitable for rotation-invariant kernels and high-dimensional datasets.

Plan of attack. First we will write down this explicit functional form (Definition 4). Then we will state our assertion that this functional form approximates the true kernel eigensystem (HEA). Next, we will demonstrate that the HEA is an excellent approximation for several rotation-invariant kernels on several real image datasets (Figure 2). We will then give an intuitive justification for the HEA and give two formal theorems which state that the HEA holds for Gaussian data as kernel width grows (Theorems 1 and 2). Afterwards, we will characterize the factors that make the HEA work better or worse (Section 4.2). Finally, we will use the HEA to predict KRR learning curves in Section 5.

We begin by explicitly defining our Hermite eigensystem:

Definition 4 (Hermite eigensystem). Given a data covariance matrix $\Sigma = U\Gamma U^\top$ and a sequence of level coefficients (c_ℓ) , we define the $(\Sigma, (c_\ell))$ -Hermite eigensystem to be the set of (scalar, function) pairs

$$\mathcal{HE}(\Sigma, (c_\ell)) = \{(\lambda_\alpha, \phi_\alpha) \text{ for all } \alpha \in \mathbb{N}_0^d\} \quad (4)$$

where for each multi-index α the proposed eigenvalue and eigenfunction are constructed as

$$\lambda_\alpha = c_{|\alpha|} \cdot \prod_{i=1}^d \gamma_i^{\alpha_i} \quad \text{and} \quad \phi_\alpha = h_\alpha^{(\Sigma)}, \quad (5)$$

where $|\alpha| = \sum_i \alpha_i$ and $h_\alpha^{(\Sigma)}$ is the multivariate Hermite polynomial given in Equation (3).

The $(\Sigma, (c_\ell))$ -Hermite eigensystem is a set of Hermite polynomials ϕ_α and associated positive scalars λ_α , one for each multi-index $\alpha \in \mathbb{N}_0^d$. The eigenvalues (λ_α) are monomials in the data covariance eigenvalues (γ_i) , rescaled by the appropriate level coefficient $c_{|\alpha|}$.

We now present the Hermite eigenstructure ansatz, which attests that this set of (scalar, function) pairs is in fact a close match to the true kernel eigensystem.

Let K be a rotation-invariant kernel and let μ be a measure over $\mathbb{R}^d$ with zero mean. Then let:

- $\Sigma = \mathbb{E}_{x \sim \mu}[xx^\top]$ be the data covariance matrix,
- $r = \text{Tr}[\Sigma]^{1/2}$ be the root-mean-squared data norm, and
- $(c_\ell) = \text{coeffs}(K, r)$ be the level coefficients of K restricted to the sphere $r\mathbb{S}^{d-1}$.

The *Hermite eigenstructure ansatz* asserts that

$$\text{eigensystem}(\mu, K) \approx \mathcal{HE}(\Sigma, (c_\ell)). \quad (\text{HEA})$$

That is, the true kernel eigensystem is approximately equal to the $(\Sigma, (c_\ell))$ -Hermite eigensystem given in Definition 4.

The HEA is a strong claim: it asserts that the kernel eigensystem, to a good approximation, has a simple analytical form which depends only on the second-order statistics of μ and no higher moments and that the kernel eigenfunctions are multivariate Hermite polynomials *independent* of the kernel chosen (so long as it is rotation-invariant).

Rather than a provable fact, the HEA should be treated as a falsifiable claim that may hold well or poorly in any given setting. A natural first question to ask is: does the HEA hold in settings of interest? This is a question best answered empirically – it suffices to check that both the predicted spectrum and predicted Hermite eigenbasis match the empirical kernel spectrum and eigenbasis (up to finite-dataset effects). Surprisingly, we find that the answer is *yes*: in Figure 2, we give four

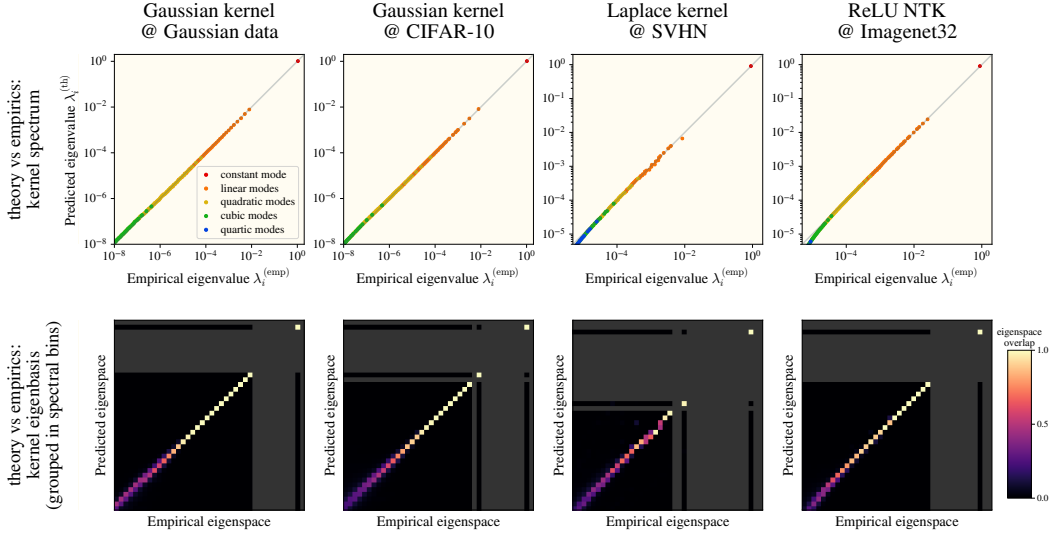


Figure 2: **The Hermite eigenstructure ansatz (HEA) accurately predicts the eigenvalues (top) and eigenfunctions (bottom) of various kernel/dataset combinations.** For four kernel/dataset settings (columns), we compute the empirical kernel eigensystem $\{(\lambda_i^{(\text{emp})}, \phi_i^{(\text{emp})})\}$ and compare to the theoretical eigenpairs $\{(\lambda_i^{(\text{th})}, \phi_i^{(\text{th})})\}$ obtained from Definition 4 and indexed in order of decreasing $\lambda_i^{(\text{th})}$. In the top plot in each column, the i -th point from the top right has coordinates $(\lambda_i^{(\text{emp})}, \lambda_i^{(\text{th})})$ and its color indicates the polynomial degree of $\phi_i^{(\text{th})}$. In the bottom plot in each column, we bin both the predicted and empirical eigenfunctions into logarithmic spectral bins and visualize the pairwise subspace overlap (Equation (43)), with axes matching the top plot. Grey pixels indicate bins with no eigenvalues. In all plots, concentration along the diagonal indicates theory-experiment match. See Appendix D.2 for further explanation and experimental details.

examples across various kernels and image datasets. In the following subsections, we provide intuitive derivations and rigorous theorem statements that explain when and why the HEA approximates the true kernel eigensystem well in natural settings.

4.1 THE HEA FOR GAUSSIAN DATA: SOME INTUITION AND TWO THEOREMS

When might we expect the HEA to hold? To gain the central intuition, it is sufficient to consider a simple case of univariate Gaussian data $x \sim \mu = \mathcal{N}(0, \gamma)$ and the Gaussian kernel $K_\sigma(x, x') = e^{-\frac{1}{2\sigma^2}(x-x')^2}$. This kernel admits the feature map $K_\sigma(x, x') = \langle \psi_\sigma(x), \psi_\sigma(x') \rangle$ where

$$\psi_\sigma(x) = e^{-\frac{x^2}{2\sigma^2}} \cdot \begin{bmatrix} 1 & \frac{x}{\sigma} & \frac{x^2}{\sqrt{2}\sigma^2} & \cdots & \frac{x^\ell}{\sqrt{\ell!}\sigma^\ell} & \cdots \end{bmatrix}, \quad (6)$$

We would like to find the directions of principal covariance of $\psi_\sigma(x)$. Let us suppose that $\sigma^2 \gg \gamma$: the kernel width dominates the width of the data distribution. Examining Equation (6), we can make two observations. First, the exponential prefactor will be close to one, and we may approximate ψ_σ componentwise as $[\psi_\sigma(x)]_\ell \approx \frac{x^\ell}{\sqrt{\ell!}\sigma^\ell}$. This amounts to approximating our kernel as $K_\sigma(x, x') = \sum_\ell \frac{(xx')^\ell}{\sigma^{2\ell} \cdot \ell!}$ — that is, as a dot-product kernel with coefficients $c_\ell = \sigma^{-2\ell}$. Second, each component of ψ_σ will dominate all subsequent components:

$$\mathbb{E}_x [[\psi_\sigma(x)]_\ell^2] \propto \sigma^{-2\ell} \gamma^\ell \gg \mathbb{E}_x [[\psi_\sigma(x)]_{\ell+1}^2] \propto \sigma^{-2(\ell+1)} \gamma^{\ell+1}. \quad (7)$$

Since the first element of ψ_σ is by far the largest (and since we do not center ψ_σ before computing eigendirections), the first direction of principal variation will correspond to $\phi_0(x) \approx 1$, with variance $\lambda_0 \approx 1$.⁴ The next direction will correspond to $\phi_1(x) \approx \gamma^{-1/2}x$ with variance $\lambda_1 \approx \sigma^{-2}$. The next eigenfunction must incorporate the x^2 direction, but we have a problem: x^2 is not orthogonal to

⁴We index eigenmodes from 0 instead of 1 here to match the polynomial order ℓ .

$\phi_0(x) = 1$. We must therefore orthogonalize it with respect to ϕ_0 against our measure μ in the usual Gram-Schmidt fashion. This yields $\phi_2(x) \approx \frac{1}{\sqrt{2}}(\gamma^{-1}x^2 - 1)$, which using standard formulas for Gaussian integrals gives an eigenvalue

$$\lambda_2 = \int K(x, x') \phi_2(x) \phi_2(x') d\mu(x) d\mu(x') \approx \sigma^{-4} \gamma^2. \quad (8)$$

Continuing this process to higher orders, we find that the kernel eigensystem matches that predicted by the HEA: ϕ_ℓ is the ℓ -th orthogonal polynomial with respect to our Gaussian measure — that is, the Hermite polynomial $h_\ell(\gamma^{-1/2}x)$ — and the ℓ -th eigenvalue is $\lambda_\ell \approx \sigma^{-2\ell} \gamma^\ell = c_\ell \gamma^\ell$. The corrections hidden by every “ $\approx$ ” in this derivation are of relative size $O(\sigma^{-2}\gamma)$ and thus vanish as σ grows.⁵

This same analysis holds for any dot-product kernel $K(x, x') = \sum_\ell \frac{c_\ell}{\ell!} (xx')^\ell$ with level coefficients (c_ℓ) such that $\frac{c_{\ell+1}}{c_\ell} \ll 1$. It can, with some difficulty, be further extended to apply to multivariate Gaussian data $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$. But what if the kernel is not a dot-product kernel, such as the Laplace kernel $K(\mathbf{x}, \mathbf{x}') = e^{\frac{1}{\sigma} \|\mathbf{x} - \mathbf{x}'\|}$? Unlike the Gaussian kernel, the Laplace kernel is not well-approximated by a dot-product kernel even at large width because of its nonanalyticity at zero. However, like all rotation-invariant kernels, the Laplace kernel *is* a dot-product kernel when restricted to a sphere (and will be close to a dot-product kernel when restricted to a spherical shell whose thickness is not too big). In such a case, we will require the data to be high-dimensional: for data with a high effective dimension $d_{\text{eff}} := \frac{\text{Tr}[\Sigma]^2}{\text{Tr}[\Sigma^2]} \gg 1$, samples will tend to concentrate in norm.⁶ We may thus safely approximate *any* rotation-invariant K as a dot-product kernel.

Having given an intuitive derivation of the HEA for Gaussian data, we now move to formal statements. Our first theorem states that the HEA holds for the Gaussian kernel at large width.

Theorem 1 (The HEA holds for a wide Gaussian kernel on a Gaussian measure).

Let $\mu = \mathcal{N}(0, \Sigma)$ be a multivariate Gaussian measure and let $K_\sigma(\mathbf{x}, \mathbf{x}') = e^{-\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{x}'\|^2}$ be the Gaussian kernel with width σ . Let $r = \text{Tr}[\Sigma]^{1/2}$ and let $(c_\ell) = \text{coeffs}(K_\sigma, r)$, which yields $c_\ell = \sigma^{-2\ell} e^{-\frac{r^2}{2\sigma^2}}$. Then:

$$\text{as } \sigma \rightarrow \infty, \text{eigensystem}(\mu, K_\sigma) \rightarrow \mathcal{HE}(\Sigma, (c_\ell)).$$

Proof sketch (full proof in [Appendix H](#)). Mehler’s formula can give the Gaussian kernel’s eigensystem exactly ([Mehler, 1866](#)). Taking $\sigma \rightarrow \infty$ in the resulting expressions yields agreement with the HEA.

Our second theorem applies to dot-product kernels with fast-decaying level coefficients.

Theorem 2 (The HEA holds for a fast-decaying dot-product kernel on a Gaussian measure).

Let $\mu = \mathcal{N}(0, \Sigma)$ be a multivariate Gaussian measure with variance $\Sigma \succ \mathbf{0}$ and let

$$K_{(c_\ell)}(\mathbf{x}, \mathbf{x}') = \sum_{\ell=0}^{\infty} \frac{c_\ell}{\ell!} (\mathbf{x}^\top \mathbf{x}')^\ell$$

be a dot-product kernel with coefficients $c_\ell \geq 0$ such that $c_{\ell+1} \leq \epsilon \cdot c_\ell$ for some $\epsilon > 0$. Then:

$$\text{as } \epsilon \rightarrow 0, \text{eigensystem}(\mu, K_{(c_\ell)}) \rightarrow \mathcal{HE}(\Sigma, (c_\ell)) \text{ linearly in } \epsilon.$$

Proof sketch: Our proof formalizes the intuitive “Gram-Schmidt” derivation of the HEA given above. We use perturbation theory to show that the kernel eigenstructure splits into exponentially-separated segments, with the ℓ -th segment eigenstructure determined almost fully by the ℓ -th order term of

⁵Were we to repeat this calculation with a different measure μ for $\mathbf{x}$, we would obtain the orthogonal polynomials with respect to μ as eigenfunctions. For example, if $\mu = U[-1, 1]$, we get the Legendre polynomials.

⁶For Gaussian data $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$, the relative variance of the norm is $\text{Var}[\|\mathbf{x}\|^2] / \mathbb{E}[\|\mathbf{x}\|^2]^2 = 2/d_{\text{eff}}$, which falls to zero as d_{eff} grows.

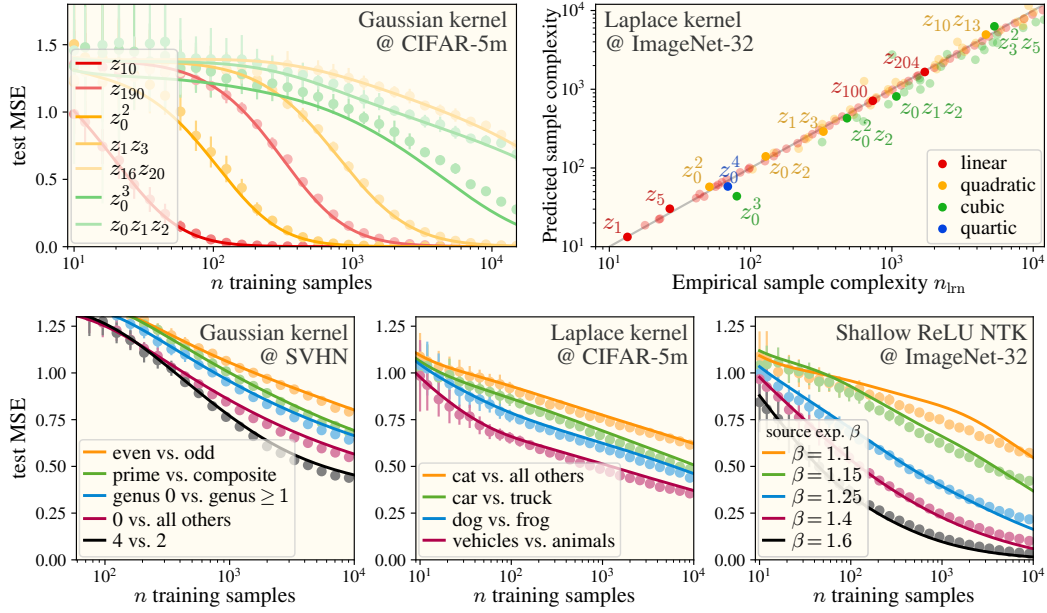


Figure 3: **Using only the empirical data covariance and a polynomial decomposition of the target function, we predict learning curves across a variety of kernels, datasets, and targets. (Top row.)** We predict test error on synthetic target functions of real datasets. Each target is the multi-Hermite polynomial (Equation (3)) whose leading term is indicated in the plots. (Recall that $z_i = \mathbf{u}_i^T \mathbf{x} / \sqrt{\gamma_i}$ are the rescaled PCA coordinates.) For each of these targets, our ansatz predicts both learning curves (top left) and the sample complexity required to achieve $\text{MSE} \leq 0.5$ (top right). **(Bottom left and center.)** We train on binarized true target functions. See Appendix D.1 for details. **(Bottom right.)** We construct synthetic targets by drawing from a Gaussian process. The source exponent controls the difficulty of the target. See Appendix D.1 for details. For all learning curve predictions, we estimate the coefficients of the target function in the predicted eigenbasis using the Gram-Schmidt process described in Appendix D.3. See Appendix D.4 for full experimental details.

$K_{(c_\ell)}$. Due to the complexity of the proof, we break it up into stages: we rigorously state and prove the one-dimensional case in Appendix I, then state and prove the general case in Appendix J.

4.2 CONDITIONS FOR SUCCESS: FAST DECAY OF c_ℓ , HIGH DATA DIMENSION, AND A “GAUSSIAN ENOUGH” DATA DISTRIBUTION

Informed by these theoretical results, we empirically identify (Appendix E) three conditions under which the HEA holds reasonably well:

- 1) **Fast decay of level coefficients.** As discussed in Section 4.1, we need $c_\ell \gg \gamma_1 c_{\ell+1}$ for the Gram-Schmidt process underlying the HEA to work. In Figure 12, we show that as we decrease the Gaussian kernel’s width (and thus increase $\frac{c_{\ell+1}}{c_\ell}$) on a fixed dataset, the HEA eventually breaks.
- 2) **High data dimension (for some kernels).** As previously discussed, concentration of norm (via high d_{eff}) is required if we are to approximate an arbitrary rotation-invariant kernel as a dot-product kernel. In Figure 13, we show that for the Laplace kernel and ReLU NTK, agreement with the HEA worsens as d_{eff} decreases. However, since the Gaussian kernel is smooth at $\mathbf{x} = \mathbf{x}'$, it does not require concentration of norm, and low d_{eff} is fine (Figure 14).
- 3) **“Gaussian enough” data distribution.** Common image datasets are complex enough to roughly satisfy simple tests of Gaussianity, such as coordinatewise Gaussian marginals. As we make the dataset simpler (CIFAR $\rightarrow$ SVHN $\rightarrow$ MNIST $\rightarrow$ tabular), these marginals become less Gaussian, and HEA agreement degrades (Figures 15 and 16). It is noteworthy that our theory empirically works better on more complex datasets thanks to the blessings of dimensionality.

5 THE HEA ALLOWS PREDICTION OF KRR LEARNING CURVES

Under the conditions outlined in the previous section, we expect the HEA to accurately predict kernel eigenstructure. We aim to plug these results directly into the aforementioned KRR eigenframework

(of e.g. [Simon et al. \(2021\)](#)) to predict the final test risk of KRR. However, a key challenge remains: using the eigenframework requires knowing the coefficients of the target function in the kernel eigenbasis, $f_*(\mathbf{x}) = \sum_i v_i \phi_i(\mathbf{x})$. We must estimate these coefficients v_i from finitely many samples of the target function.

Were the data perfectly Gaussian, the multi-Hermite polynomials would be an orthonormal basis with respect to the measure. We could then estimate the coefficients by simply taking inner products between the target vector and generated Hermite polynomials and expect the estimation error to decay as $\mathcal{O}(N^{-1/2})$ with the total number of samples N . However, small amounts of non-Gaussianity in the data introduce cascading non-orthogonality in the Hermite basis. As a result, the naive method overestimates the power in the overlapping modes. To rectify this effect, we modify our measurement technique by re-orthogonalizing the sampled Hermite polynomials via the Gram-Schmidt process:⁷

$$\text{iterate over increasing } \alpha : \quad h_{\alpha}^{(\text{GS})} = \text{unitnorm} \left(h_{\alpha} - \sum_{\alpha' < \alpha} \langle h_{\alpha'}^{(\text{GS})}, h_{\alpha} \rangle h_{\alpha'}^{(\text{GS})} \right) \quad (9)$$

where $h_{\alpha} := h_{\alpha}(\mathcal{X})$ is the sampled multi-Hermite polynomial and the α are sorted by their HEA eigenvalues. We proceed to estimate the coefficients as $\hat{v}_i = \langle h_i^{(\text{GS})}, \mathbf{y} \rangle$. As we show in [Figure 3](#), with the resulting coefficients we can reliably predict learning curves on a variety of tasks and kernels.

6 MLPs LEARN HERMITE POLYNOMIALS IN THE ORDER PREDICTED BY THE HEA

One consequence of our theory is that there exists a canonical learning order in which KRR learns Hermite polynomials as sample size increases: each polynomial’s learning priority is given by its associated HEA eigenvalue. Here, we check whether this order also predicts the *training time* learning order of *feature-learning MLPs* ([Yang & Hu, 2021](#)). We train MLPs online on multi-Hermite polynomial target functions of Gaussian data and CIFAR-5m. We find that the effective optimization time required to reach $\text{MSE} \leq 0.1$ is well predicted by the HEA eigenvalue ([Figure 4](#)).

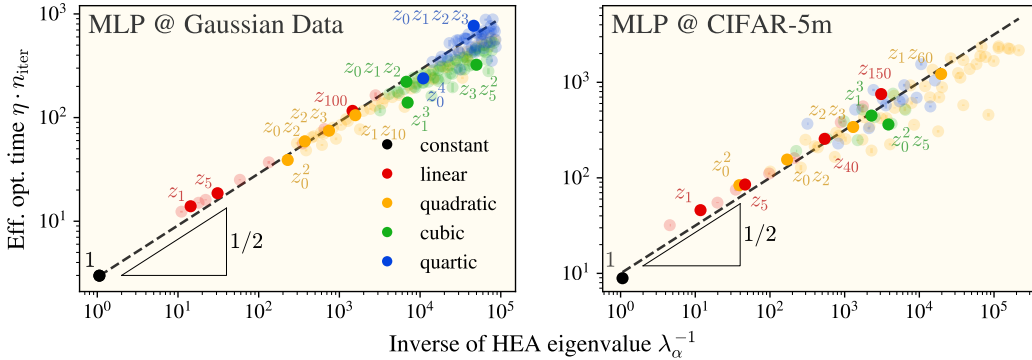


Figure 4: The HEA accurately predicts polynomial learning order in feature-learning MLPs. We measure the amount of time it takes to train feature-learning MLPs, each dot being a trained MLP on one multi-Hermite polynomial target once a test error of $\text{MSE} \leq 0.1$ is reached. We observe find that the optimization time is well-predicted by the *inverse square root* of the HEA eigenvalue $\lambda_{\alpha}^{-1/2}$.

Details of our exact MLP setup experiments can be found in [Appendix F](#), detailing validation of the feature-learning regime, insight into hyperparameter choices, and model performance when taken into the NTK/lazy and ultra-rich regimes.

7 DISCUSSION

We have presented a theoretical framework which describes how KRR “sees” complex natural datasets: namely, as a nearly-Gaussian measure with Hermite eigenstructure as per [Definition 4](#). This is a proof of concept that end-to-end theories of learning — mapping dataset structure all the way to model performance — are possible for a nontrivial learning algorithm on real datasets. Theories of this sort applicable to more general algorithms may be a good end goal for learning theory.

⁷For a full discussion of this method, see [Appendix D.3](#).

REFERENCES

- Alexander Atanasov, Alexandru Meterez, James B Simon, and Cengiz Pehlevan. The optimization landscape of sgd across the feature learning strength. *arXiv preprint arXiv:2410.04642*, 2024. Cited on page 31.
- Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *arXiv preprint arXiv:2303.01372*, 2023. Cited on page 18.
- Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *Advances in neural information processing systems*, 35:9768–9783, 2022. Cited on page 3.
- Blake Bordelon and Cengiz Pehlevan. Learning curves for sgd on structured features. *arXiv preprint arXiv:2106.02713*, 2021. Cited on page 3.
- Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pp. 1024–1034. PMLR, 2020. Cited on pages 2, 3, and 18.
- Francesco Cagnetta and Matthieu Wyart. Towards a theory of how the structure of language is acquired by deep neural networks. *Advances in Neural Information Processing Systems*, 37: 83119–83163, 2024. Cited on page 3.
- Francesco Cagnetta, Leonardo Petrini, Umberto M Tomasini, Alessandro Favero, and Matthieu Wyart. How deep neural networks learn compositional data: The random hierarchy model. *Physical Review X*, 14(3):031001, 2024. Cited on page 3.
- 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):1–12, 2021. Cited on page 18.
- Yang Chen and Nigel Lawrence. Small eigenvalues of large hankel matrices. *Journal of Physics A: Mathematical and General*, 32(42):7305, 1999. Cited on page 46.
- Chen Cheng and Andrea Montanari. Dimension free ridge regression. *arXiv preprint arXiv:2210.08571*, 2022. Cited on page 18.
- Lenaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019. Cited on page 3.
- Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*, 2023. Cited on page 3.
- Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. Cited on pages 45 and 55.
- J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. Cited on page 19.
- Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of predictions: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018. Cited on page 18.
- Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685. PMLR, 2019. Cited on page 3.
- Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL . Cited on pages 19 and 27.
- Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In *Conference On Learning Theory*, pp. 1887–1930. PMLR, 2018. Cited on page 3.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33: 14820–14830, 2020. Cited on page 3.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020. Cited on page 3.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022. Cited on page 18.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. Cited on page 3.

Arthur Jacot, Berfin Şimşek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Kernel alignment risk estimator: Risk prediction from training data. *Advances in neural information processing systems*, 33:15568–15578, 2020. Cited on pages 2, 3, and 18.

Tosio Kato. Variation of discrete spectra. *Communications in Mathematical Physics*, 111(3):501–504, 1987. Cited on page 61.

Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. The MNIST database of handwritten digits. , 1998. Dataset. Cited on page 19.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations (ICLR)*, 2018. Cited on page 3.

Jason D Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *Advances in Neural Information Processing Systems*, 37:58716–58756, 2024. Cited on page 3.

Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021. Cited on page 18.

F. Gustav Mehler. Ueber die entwicklung einer function von beliebig vielen variablen nach laplaceschen functionen höherer ordnung. *Journal für die reine und angewandte Mathematik*, 66: 161–176, 1866. Cited on pages 7 and 39.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019. Cited on page 3.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018. ISBN 978-0-262-03940-6. Cited on page 4.

Alireza Mousavi-Hosseini, Denny Wu, and Murat A Erdogdu. Learning multi-index models with neural networks via mean-field langevin dynamics. *arXiv preprint arXiv:2408.07254*, 2024. Cited on page 3.

Preetum Nakkiran, Behnam Neyshabur, and Hanie Sedghi. The deep bootstrap framework: Good online learners are good offline generalizers. *arXiv preprint arXiv:2010.08127*, 2020. Cited on page 19.

Radford M Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pp. 29–53. Springer, 1996. Cited on page 3.

- 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*, pp. 7, 2011. Workshop paper. Cited on page 19.
- Maria Refinetti, Alessandro Ingrosso, and Sebastian Goldt. Neural networks trained with sgd learn distributions of increasing complexity. In *International Conference on Machine Learning*, pp. 28843–28863. PMLR, 2023. Cited on page 3.
- Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, pp. 3889–3897. PMLR, 2021. Cited on page 18.
- I. J. Schoenberg. Positive definite functions on spheres. *Duke Mathematical Journal*, 9(1):96 – 108, 1942. doi: 10.1215/S0012-7094-42-00908-6. URL . Cited on page 4.
- Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1): e2408799121, 2025. Cited on page 3.
- James B Simon, Madeline Dickens, Dhruva Karkada, and Michael R DeWeese. The eigenlearning framework: A conservation law perspective on kernel ridge regression and wide neural networks. *arXiv preprint arXiv:2110.03922*, 2021. Cited on pages 2, 3, 9, and 18.
- Peter Sollich. Gaussian process regression with mismatched models. In *Advances in Neural Information Processing Systems*, pp. 519–526. MIT Press, 2001. Cited on pages 2, 3, and 18.
- Terence Tao. *Topics in Random Matrix Theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012. ISBN 978-0-8218-7430-1. Cited on page 44.
- Gerald Teschl. *Mathematical Methods in Quantum Mechanics: With Applications to Schrödinger Operators*. Number Vol. 99 in Graduate Studies in Mathematics. American mathematical society, Providence, R.I, 2009. ISBN 978-0-8218-4660-5. Cited on page 44.
- Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2022. Cited on page 18.
- Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020. Cited on page 18.
- Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pp. 11727–11737. PMLR, 2021. Cited on page 9.
- Greg Yang, James B Simon, and Jeremy Bernstein. A spectral condition for feature learning. *arXiv preprint arXiv:2310.17813*, 2023. Cited on page 32.
- Huaiyu Zhu, Christopher K. I. Williams, Richard Rohwer, and Michal Morciniec. Gaussian regression and optimal finite dimensional linear models. Technical Report NCRG/97/011, Neural Computing Research Group, Aston University, Birmingham, UK, 1997. URL . Cited on pages 3 and 39.

STATEMENT ON THE USE OF LARGE LANGUAGE MODELS

We used large language models (LLMs) for analytical computations, writing code, detailed literature search on narrow topics, and the Taylor expansions of Laplace and ReLU kernels appearing in Appendix B. We found that LLMs performed certain tasks faster than us but with a propensity for miscommunication or overconfidence, especially when fashioning proofs, so we sought or performed independent verification of everything useful we got from an LLM.

A REVIEW OF HERMITE POLYNOMIALS

Our ansatz for kernel eigenstructure is constructed from Hermite polynomials, and we use several key properties throughout the paper. In this appendix, we give a brief review of Hermite polynomials.

Let us write He_ℓ for the *probabilist's* Hermite polynomials.⁸ These are the unique set of polynomials satisfying the following properties:

- (i) **Degree.** He_ℓ is a polynomial of degree ℓ .
- (ii) **Monic.** The leading coefficient is 1, so $\text{He}_\ell(x) = x^\ell + \dots$ (in particular $\text{He}_0(x) = 1$).
- (iii) **Orthogonality w.r.t. $\mathcal{N}(0, 1)$.** For $\ell \neq m$, it holds that $\mathbb{E}_{x \sim \mathcal{N}(0,1)}[\text{He}_\ell(x)\text{He}_m(x)] = 0$.

The probabilist's Hermite polynomials have squared norm $\mathbb{E}_{x \sim \mathcal{N}(0,1)}[\text{He}_\ell^2(x)] = \ell!$. Since we will use the Hermite polynomials as a basis in which to express other functions, we will prefer them to have unit norm, so we use the *normalized* probabilist's Hermite polynomials $h_\ell(x) = \frac{1}{\sqrt{\ell!}}\text{He}_\ell(x)$, which satisfy $\mathbb{E}_{x \sim \mathcal{N}(0,1)}[h_k(x)h_\ell(x)] = \delta_{k\ell}$. The first several such polynomials are given by

$$\begin{aligned} h_0(x) &= 1, \\ h_1(x) &= x, \\ h_2(x) &= \frac{1}{\sqrt{2}}(x^2 - 1), \\ h_3(x) &= \frac{1}{\sqrt{6}}(x^3 - 3x), \\ h_4(x) &= \frac{1}{\sqrt{24}}(x^4 - 6x^2 + 3), \\ h_5(x) &= \frac{1}{\sqrt{120}}(x^5 - 10x^3 + 15x). \end{aligned}$$

The Hermite polynomials obey a remarkable number of useful mathematical relations. In particular, single- and multidimensional integrals of Hermite polynomials against exponentials and other polynomials are often computable in closed form. The Wikipedia page (https://en.wikipedia.org/wiki/Hermite_polynomials) is a good first reference. Here we state only one such integral which is useful for understanding the intuitive derivation of the HEA in Section 4.1:

$$\mathbb{E}_{x \sim \mathcal{N}(0,1)}[h_\ell(x) x^m] = \begin{cases} \frac{m!}{\sqrt{\ell!} 2^{(m-\ell)/2} ((m-\ell)/2)!}, & m \geq \ell \text{ and } m \equiv \ell \pmod{2}, \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

That is, the ℓ -th Hermite polynomial is orthogonal to all monomials x^m whose order is less than ℓ (or whose order is simply of a different parity from ℓ). In particular, this implies that $\mathbb{E}_{x' \sim \mathcal{N}(0,1)}[(xx')^m h_\ell(x')] = 0$ when $m < \ell$: the function h_ℓ lies in the nullspace of the rank-one kernel term $K(x, x') = (xx')^m$.

At one point, we need the multiplication and differentiation formulas:

$$\text{He}_n \text{He}_m = \sum_{j=0}^{\min(m,n)} \binom{m}{j} \binom{n}{j} j! \text{He}_{n+m-2j} \quad (11)$$

$$\text{He}'_n = n \text{He}_{n-1} \quad (12)$$

⁸Some references use the *physicist's* Hermite polynomials H_ℓ . These are related to the probabilist's Hermite by He_ℓ by $\text{H}_\ell(x) = 2^{\ell/2} \text{He}_\ell(\sqrt{2}x)$. When using Hermite polynomials, be sure you know which ones you're talking about!

B ON-SPHERE LEVEL COEFFICIENTS FOR COMMON ROTATION-INVARIANT KERNELS

In this appendix, we tabulate the on-sphere level coefficients for various kernels. We begin by recalling [Definition 3](#).

Definition 3 (On-sphere level coefficients). The *on-sphere level coefficients* of a rotation-invariant kernel K at a radius $r > 0$ are the nonnegative sequence $\text{coeffs}(K, r) := (c_\ell)_{\ell \geq 0}$ such that

$$K(\mathbf{x}, \mathbf{x}') = \sum_{\ell \geq 0} \frac{c_\ell}{\ell!} (\mathbf{x}^\top \mathbf{x}')^\ell \quad \text{for all } \mathbf{x}, \mathbf{x}' \text{ such that } \|\mathbf{x}\| = \|\mathbf{x}'\| = r. \quad (13)$$

B.1 GAUSSIAN KERNEL

For the Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = e^{-\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{x}'\|^2}$, the level coefficients $(c_\ell) = \text{coeffs}(K, r)$ may be found by noting that, when $\|\mathbf{x}\| = \|\mathbf{x}'\| = r$, then

$$K(\mathbf{x}, \mathbf{x}') = e^{-\frac{r^2}{\sigma^2}} \cdot e^{\frac{\mathbf{x}^\top \mathbf{x}'}{\sigma^2}} = e^{-\frac{r^2}{\sigma^2}} \cdot \sum_{\ell \geq 0} \frac{1}{\sigma^{2\ell} \ell!} (\mathbf{x}^\top \mathbf{x}')^\ell. \quad (14)$$

Pattern-matching to [Definition 3](#), we may then observe that

$$\begin{aligned} c_0 &= e^{-\frac{r^2}{\sigma^2}}, \\ c_1 &= e^{-\frac{r^2}{\sigma^2}} \cdot \sigma^{-2}, \\ c_2 &= e^{-\frac{r^2}{\sigma^2}} \cdot \sigma^{-4}, \\ &\vdots \\ c_\ell &= e^{-\frac{r^2}{\sigma^2}} \cdot \sigma^{-2\ell}. \end{aligned}$$

B.2 LAPLACE KERNEL

Here we obtain the on-sphere level coefficients for the Laplace kernel $K(\mathbf{x}, \mathbf{x}') = e^{-\frac{1}{\sigma} \|\mathbf{x} - \mathbf{x}'\|}$. Let

$$s := \frac{\mathbf{x}^\top \mathbf{x}'}{r^2} \in [-1, 1], \quad \beta := \frac{\sqrt{2} r}{\sigma}. \quad (15)$$

Since $\|\mathbf{x} - \mathbf{x}'\| = \sqrt{2} r \sqrt{1 - s}$ on the sphere,

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\beta \sqrt{1 - s}). \quad (16)$$

Closed form for on-sphere level coefficients. Let $y_n(x)$ denote the reverse Bessel polynomials, with exponential generating function

$$\sum_{\ell \geq 0} y_{\ell-1}(x) \frac{t^\ell}{\ell!} = \exp\left(\frac{1}{2x}(1 - \sqrt{1 - 2xt})\right). \quad (17)$$

Applying (17) with $x = \frac{1}{2\beta}$ and $t = \frac{\beta}{2} s$ gives

$$\exp(\beta(1 - \sqrt{1 - s})) = \sum_{\ell \geq 0} y_{\ell-1}\left(\frac{1}{2\beta}\right) \frac{(\beta s/2)^\ell}{\ell!}. \quad (18)$$

Multiplying by $e^{-\beta}$ and substituting $s = (\mathbf{x}^\top \mathbf{x}')/r^2$, the definition $K(\mathbf{x}, \mathbf{x}') = \sum_{\ell \geq 0} \frac{c_\ell}{\ell!} (\mathbf{x}^\top \mathbf{x}')^\ell$ implies

$$c_\ell = \frac{e^{-\beta}}{r^{2\ell}} y_{\ell-1}\left(\frac{1}{2\beta}\right) \left(\frac{\beta}{2}\right)^\ell, \quad \beta = \frac{\sqrt{2} r}{\sigma}. \quad (19)$$

The first few coefficients. With the convention $y_{-1} \equiv 1$, $y_0(x) = 1$, $y_1(x) = 1 + x$, $y_2(x) = 3 + 3x + x^2$, $y_3(x) = 15 + 15x + 6x^2 + x^3$, we obtain (for $\beta = \sqrt{2}r/\sigma$)

$$\begin{aligned} c_0 &= e^{-\beta}, \\ c_1 &= \frac{e^{-\beta}}{r^2} \frac{\beta}{2}, \\ c_2 &= \frac{e^{-\beta}}{r^4} \left(\frac{\beta^2}{4} + \frac{\beta}{8} \right), \\ c_3 &= \frac{e^{-\beta}}{r^6} \left(\frac{\beta^3}{8} + \frac{3\beta^2}{16} + \frac{\beta}{16} \right), \\ c_4 &= \frac{e^{-\beta}}{r^8} \left(\frac{\beta^4}{16} + \frac{3\beta^3}{16} + \frac{5\beta^2}{32} + \frac{5\beta}{128} \right). \end{aligned}$$

Large- ℓ asymptotics. The dominant singularity of $F(s) = e^{-\beta\sqrt{1-s}}$ is at $s = 1$, with $F(s) = 1 - \beta\sqrt{1-s} + O(1-s)$, yielding $[s^\ell]F(s) \sim \frac{\beta}{2\sqrt{\pi}} \ell^{-3/2}$, where the coefficient extraction operator $[s^\ell]F(s)$ returns the ℓ -th coefficient in the power series of $F(s)$. Since $c_\ell = r^{-2\ell} \ell! [s^\ell]F(s)$,

$$c_\ell \sim \frac{\beta}{2\sqrt{\pi}} \frac{\ell!}{r^{2\ell} \ell^{3/2}} \quad (\ell \rightarrow \infty), \quad \beta = \frac{\sqrt{2}r}{\sigma}. \quad (20)$$

B.3 THE RELU NNGP KERNEL

For a one-hidden-layer ReLU network with first-layer weight variance σ_w^2 and bias variance σ_b^2 (and no output bias), the first-layer preactivation kernel is

$$K_1(\mathbf{x}, \mathbf{x}') = \sigma_w^2 \mathbf{x}^\top \mathbf{x}' + \sigma_b^2. \quad (21)$$

On the sphere $\|\mathbf{x}\| = \|\mathbf{x}'\| = r$, set

$$q := K_1(\mathbf{x}, \mathbf{x}) = \sigma_w^2 r^2 + \sigma_b^2, \quad s := \frac{\mathbf{x}^\top \mathbf{x}'}{r^2}, \quad \rho := \frac{K_1(\mathbf{x}, \mathbf{x}')}{\sqrt{K_1(\mathbf{x}, \mathbf{x})K_1(\mathbf{x}', \mathbf{x}')}} = \frac{\sigma_w^2 r^2 s + \sigma_b^2}{q}. \quad (22)$$

The ReLU NNGP kernel is

$$K_2(\mathbf{x}, \mathbf{x}') = \frac{\sigma_w^2}{2\pi} q \left(\sqrt{1 - \rho^2} + (\pi - \arccos \rho) \rho \right) =: \frac{\sigma_w^2}{2\pi} q H(\rho). \quad (23)$$

We Taylor-expand K_2 in powers of $\mathbf{x}^\top \mathbf{x}'$ and write

$$K_2(\mathbf{x}, \mathbf{x}') = \sum_{\ell \geq 0} \frac{c_\ell}{\ell!} (\mathbf{x}^\top \mathbf{x}')^\ell. \quad (24)$$

A change of variables gives the coefficients

$$c_\ell = \frac{\sigma_w^2}{2\pi} q \left(\frac{\sigma_w^2}{q} \right)^\ell H^{(\ell)}(a) \quad \text{with} \quad a := \frac{\sigma_b^2}{q}, \quad q = \sigma_w^2 r^2 + \sigma_b^2, \quad (25)$$

where $H(\rho) = \sqrt{1 - \rho^2} + (\pi - \arccos \rho) \rho$ and $H^{(\ell)}$ denotes the ℓ -th derivative.

The first few coefficients follow from

$$\begin{aligned} H(a) &= \sqrt{1 - a^2} + (\pi - \arccos a) a, \\ H'(a) &= \pi - \arccos a, \\ H''(a) &= \frac{1}{\sqrt{1 - a^2}}, \\ H^{(3)}(a) &= \frac{a}{(1 - a^2)^{3/2}}, \\ H^{(4)}(a) &= \frac{2a^2 + 1}{(1 - a^2)^{5/2}}, \end{aligned}$$

yielding

$$\begin{aligned}
c_0 &= \frac{\sigma_w^2}{2\pi} q \left[\sqrt{1-a^2} + (\pi - \arccos a) a \right], \\
c_1 &= \frac{\sigma_w^2}{2\pi} q \left(\frac{\sigma_w^2}{q} \right) (\pi - \arccos a), \\
c_2 &= \frac{\sigma_w^2}{2\pi} q \left(\frac{\sigma_w^2}{q} \right)^2 \frac{1}{\sqrt{1-a^2}}, \\
c_3 &= \frac{\sigma_w^2}{2\pi} q \left(\frac{\sigma_w^2}{q} \right)^3 \frac{a}{(1-a^2)^{3/2}}, \\
c_4 &= \frac{\sigma_w^2}{2\pi} q \left(\frac{\sigma_w^2}{q} \right)^4 \frac{2a^2+1}{(1-a^2)^{5/2}}.
\end{aligned} \tag{26}$$

Asymptotics. As ℓ grows, the coefficient c_ℓ grows as

$$c_\ell = \Theta \left(\frac{\sigma_w^2}{2\pi} q \left(\frac{\sigma_w^2}{q} \right)^\ell \frac{\ell!}{\ell^{3/2}} \right), \quad \ell \rightarrow \infty. \tag{27}$$

B.4 THE RELU NTK

As in the previous subsection, we treat a shallow ReLU network. The ReLU NNGP kernel remains

$$K_2(\mathbf{x}, \mathbf{x}') = \frac{\sigma_w^2}{2\pi} q \left(\sqrt{1-\rho^2} + (\pi - \arccos \rho) \rho \right) =: \frac{\sigma_w^2}{2\pi} q H(\rho). \tag{28}$$

The corresponding two-layer ReLU NTK is

$$\Theta_2(\mathbf{x}, \mathbf{x}') = \underbrace{\sigma_w^2 K_1(\mathbf{x}, \mathbf{x}') \cdot \frac{1}{2\pi} (\pi - \arccos \rho)}_{\text{training first layer}} + \underbrace{K_2(\mathbf{x}, \mathbf{x}')}_{\text{training second layer}} \tag{29}$$

$$= \frac{\sigma_w^2}{2\pi} q \left(\sqrt{1-\rho^2} + 2\rho(\pi - \arccos \rho) \right) =: \frac{\sigma_w^2}{2\pi} q J(\rho). \tag{30}$$

We Taylor-expand Θ_2 in powers of $\mathbf{x}^\top \mathbf{x}'$ and write

$$\Theta_2(\mathbf{x}, \mathbf{x}') = \sum_{\ell \geq 0} \frac{c_\ell}{\ell!} (\mathbf{x}^\top \mathbf{x}')^\ell. \tag{31}$$

A change of variables gives the coefficients

$$c_\ell = \frac{\sigma_w^2}{2\pi} q \left(\frac{\sigma_w^2}{q} \right)^\ell J^{(\ell)}(a) \quad \text{with} \quad a := \frac{\sigma_b^2}{q}, \quad q = \sigma_w^2 r^2 + \sigma_b^2, \tag{32}$$

where $J(\rho) = \sqrt{1-\rho^2} + 2\rho(\pi - \arccos \rho)$ and $J^{(\ell)}$ denotes the ℓ -th derivative.

The first few coefficients follow from the identities

$$\begin{aligned}
J(a) &= \sqrt{1-a^2} + 2a(\pi - \arccos a), \\
J'(a) &= 2(\pi - \arccos a) + \frac{a}{\sqrt{1-a^2}}, \\
J''(a) &= \frac{3-2a^2}{(1-a^2)^{3/2}}, \\
J^{(3)}(a) &= \frac{a(5-2a^2)}{(1-a^2)^{5/2}}, \\
J^{(4)}(a) &= \frac{5+14a^2-4a^4}{(1-a^2)^{7/2}},
\end{aligned}$$

yielding

$$\begin{aligned}c_0 &= \frac{\sigma_w^2}{2\pi} q \left[\sqrt{1-a^2} + 2a (\pi - \arccos a) \right], \\c_1 &= \frac{\sigma_w^2}{2\pi} q \left(\frac{\sigma_w^2}{q} \right) \left[2(\pi - \arccos a) + \frac{a}{\sqrt{1-a^2}} \right], \\c_2 &= \frac{\sigma_w^2}{2\pi} q \left(\frac{\sigma_w^2}{q} \right)^2 \frac{3-2a^2}{(1-a^2)^{3/2}}, \\c_3 &= \frac{\sigma_w^2}{2\pi} q \left(\frac{\sigma_w^2}{q} \right)^3 \frac{a(5-2a^2)}{(1-a^2)^{5/2}}, \\c_4 &= \frac{\sigma_w^2}{2\pi} q \left(\frac{\sigma_w^2}{q} \right)^4 \frac{5+14a^2-4a^4}{(1-a^2)^{7/2}}.\end{aligned}$$

Asymptotics. As ℓ grows, the coefficient c_ℓ grows as

$$c_\ell = \Theta \left(\frac{\sigma_w^2}{2\pi} q \left(\frac{\sigma_w^2}{q} \right)^\ell \frac{\ell!}{\ell^{3/2}} \right), \quad \ell \rightarrow \infty. \quad (33)$$

C REVIEW OF THE KRR EIGENFRAMEWORK FOR PREDICTING TEST PERFORMANCE FROM TASK EIGENSTRUCTURE

The central piece of existing theory which we use in this paper is a set of equations which give the average-case test MSE of KRR in terms of the task eigenstructure. In this appendix, we will review this KRR eigenframework.

This eigenframework has been derived many times by different means in both the statistical physics community and the classical statistics community (which usually phrases the result as applying to linear ridge regression). References studying KRR include [Sollich \(2001\)](#); [Bordelon et al. \(2020\)](#); [Jacot et al. \(2020\)](#); [Simon et al. \(2021\)](#); [Loureiro et al. \(2021\)](#); [Wei et al. \(2022\)](#); references studying linear ridge regression include [Dobriban & Wager \(2018\)](#); [Wu & Xu \(2020\)](#); [Hastie et al. \(2022\)](#); [Richards et al. \(2021\)](#); [Cheng & Montanari \(2022\)](#). The result is essentially the same in all cases. Here we will adopt the terminology and notation of [Simon et al. \(2021\)](#).

Recall that we are studying KRR with a kernel function K with data sampled $\mathbf{x}_i \sim \mu$, targets generated as $y_i = f_*(\mathbf{x}_i) + \eta$, and noise $\eta \sim \mathcal{N}(0, \epsilon^2)$ with variance $\epsilon^2 \geq 0$. The kernel admits an eigendecomposition $K(\mathbf{x}, \mathbf{x}') = \sum_i \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}')$ with orthonormal eigenfunctions (ϕ_i) . Let us decompose the target function in the eigenbasis as $f_*(\mathbf{x}) = \sum_i v_i \phi_i(\mathbf{x})$. Suppose we run KRR with n samples and a ridge parameter $\delta \geq 0$, and we compute the population (i.e. test) and train MSEs as

$$\text{MSE}_{\text{te}} = \mathbb{E}_{\mathbf{x} \sim \mu} \left[(f_*(\mathbf{x}) - \hat{f}(\mathbf{x}))^2 \right] + \epsilon^2, \quad (34)$$

$$\text{MSE}_{\text{tr}} = \frac{1}{n} \sum_i (y_i - \hat{f}(\mathbf{x}_i))^2. \quad (35)$$

C.1 STATEMENT OF THE EIGENFRAMEWORK

We are now ready to state the eigenframework. Let $\kappa \geq 0$ be the unique nonnegative solution to

$$\sum_i \frac{\lambda_i}{\lambda_i + \kappa} + \frac{\delta}{\kappa} = n. \quad (36)$$

Then test risk is given approximately by

$$\text{MSE}_{\text{te}} \approx \mathcal{E}_{\text{te}} := \mathcal{E}_0 \mathcal{B}, \quad (37)$$

where the *overfitting coefficient* $\mathcal{E}_0$ is given by

$$\mathcal{E}_0 := \frac{n}{n - \sum_i \left(\frac{\lambda_i}{\lambda_i + \kappa} \right)^2} \quad (38)$$

and the *bias* is given by

$$\mathcal{B} = \sum_i \left(\frac{\kappa}{\lambda_i + \kappa} \right)^2 v_i^2 + \sigma^2. \quad (39)$$

Train risk is given by

$$\text{MSE}_{\text{tr}} \approx \mathcal{E}_{\text{tr}} \equiv \frac{\delta^2}{n^2 \kappa^2} \mathcal{E}_{\text{te}}, \quad (40)$$

What is meant by the “ $\approx$ ” in [Equations \(37\) and \(40\)](#)? This result only becomes exact in certain stringent cases; it is formally derived under an assumption that the eigenfunctions are independent Gaussian (or sub-Gaussian) variables when $\mathbf{x}$ is sampled from μ , and it is exact only in an asymptotic limit in which n and the number of eigenmodes in a given eigenvalue range (or the number of duplicate copies of any given eigenmode) both grow large at a proportional rate ([Hastie et al., 2022](#); [Bach, 2023](#)). These conditions emphatically do not apply to any realistic instance of KRR. Nonetheless, numerical experiments find that [Equations \(37\) and \(40\)](#) hold with small error even at modest n ([Canatar et al., 2021](#); [Simon et al., 2021](#); [Wei et al., 2022](#)): though derived in an idealized setting and exact only in a limit, this eigenframework holds reliably in practical cases.

In this paper, we use this eigenframework as a tool to map predictions of task eigenstructure to predictions of learning curves. Since we are here using it in settings very similar to those tested by previous works ([Bordelon et al., 2020](#); [Jacot et al., 2020](#); [Simon et al., 2021](#); [Wei et al., 2022](#)), we expect it to work well. Its use introduces some small error (as it is not perfect at finite n), but this is usually not the dominant source of error.

D EXPERIMENTS CHECKING THE HEA: DETAILS AND FURTHER DISCUSSION

This appendix contains descriptions of the experimental stack used to verify the HEA, as well as a discussion of practical considerations for applying the HEA to real datasets. It is organized as follows:

- In [Appendix D.1](#) we catalog the kernels, datasets, and target functions we use throughout our experiments.
- In [Appendix D.2](#) we explain the experiments that directly check whether the kernel eigenstructure matches the HEA prediction ([Figure 2](#)).
- In [Appendix D.3](#) we describe our method for estimating the decomposition of the target function in the Hermite eigenbasis. Unlike previous work, our method does not require constructing or diagonalizing an empirical kernel matrix.
- In [Appendix D.4](#) we detail the experimental setups for each of the learning curve and sample complexity plots ([Figures 1](#) and [3](#)).
- Finally, in [Appendix D.5](#) we show the results of various additional experiments.

D.1 KERNELS, DATASETS, AND TARGET FUNCTIONS

Kernels. We use the Gaussian kernel, Laplace kernel, ReLU NNGP kernel, and ReLU NTK in our experiments. A detailed review of these kernels can be found in [Appendix B](#).⁹

Datasets. We use the following datasets for the main experiments:

- **Mean-zero anisotropic Gaussian data.** This synthetic dataset is fully specified by its (diagonal) covariance. Different experiments set the data dimension and covariance decay rate differently; see experiment-specific details in [Appendices D.2, E](#) and [F](#).
- **CIFAR-5m** ([Nakkiran et al., 2020](#)). This dataset consists of more than 5 million samples of synthetic images akin to CIFAR-10. These images were sampled using a deep generative model trained on CIFAR-10. Though the distributions of CIFAR-5m and CIFAR-10 may not be identical, they are typically considered close enough for research purposes.
- **SVHN** ([Netzer et al., 2011](#)). This dataset contains over 600,000 images of numerals, taken from house numbers found on Google Street View.
- **ImageNet-32** ([Deng et al., 2009](#)). This dataset contains downsampled ImageNet images (32×32 pixels).
- **MNIST** ([LeCun et al., 1998](#)) and **Mushroom dataset** ([Dua & Graff, 2017](#)). We use MNIST and the UCI Mushroom tabular dataset in [Appendix E](#) as examples of insufficiently Gaussian datasets.

We sometimes employ *regularized ZCA whitening* to increase the effective dimension of the data. This is a data preprocessing technique parameterized by a ZCA strength ω^2 which maps

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top \mapsto \mathbf{U}\mathbf{S} \left(\omega^2 \overline{\mathbf{S}^2} + \mathbf{I}_d \right)^{-1/2} \mathbf{V}^\top \quad (41)$$

where $\mathbf{X} \in \mathbb{R}^{d \times N}$ is the data matrix, $\mathbf{U}\mathbf{S}\mathbf{V}^\top$ is its SVD, and we use the normalization notation $\overline{\mathbf{A}} := \mathbf{A} / (\|\mathbf{A}\|_F^2 / d)$. As the ZCA strength $\omega^2 \rightarrow 0$, we get no spectral transformation apart from a scalar normalization $\mathbf{X} \rightarrow \mathbf{U}\overline{\mathbf{S}}\mathbf{V}^\top$. Conversely, when $\omega^2 \rightarrow \infty$, we get full whitening $\mathbf{X} \rightarrow \mathbf{U}\mathbf{V}^\top$. The crossover point of this behavior occurs at $\omega^2 \sim 1$. Note that although partially-whitened Gaussian data are slightly less anisotropic, they are still distributed as a multivariate Gaussian.

We sometimes employ sample normalization, $\mathbf{x} \rightarrow \mathbf{x} / \|\mathbf{x}\|$. Note that although the normalized data lie on the hypersphere, their distribution is still anisotropic.

⁹A note for users of our codebase: in this paper, we define the Laplace kernel to be $K(\mathbf{x}, \mathbf{x}') = e^{-\frac{1}{\sigma} \|\mathbf{x}\mathbf{x}'\|}$, but because we initially used a different convention, in code it is $K(\mathbf{x}, \mathbf{x}') = e^{-\frac{1}{\sigma\sqrt{2}} \|\mathbf{x}\mathbf{x}'\|}$. When we report a kernel width in this paper, this is the width in the parameterization we use *in the paper*, not in code.

Both sample normalization and ZCA whitening are preprocessing techniques that, on aggregate, shift high-dimensional data samples closer to the hypersphere. Since the HEA relies on an expansion of kernel functions in terms of on-sphere coefficients (see [Appendix B](#)), these methods move any experimental setup closer to the regime well-described by the HEA. See [Appendix E](#) for further discussion of this point.

Targets. We use a variety of synthetic and real target functions in our experiments. All targets are scalar; the synthetic targets take continuous values, whereas the real targets are binarized ($y_i \in \{+1, -1\}$). All targets are mean-zero; for real targets, this means that the binary (super)classes are always balanced (even if the binary superclasses contain a differing number of base classes).

We use the following targets for the main experiments:

- **(Synthetic.) Multi-Hermite targets.** A single (normalized) multi-Hermite polynomials of the PCA components ([Equation \(3\)](#)).
- **(Synthetic.) Powerlaw targets.** We draw a random sample of the Gaussian process

$$f_{\star}(\mathbf{x}) = \sum_i^P c_i h_i(\mathbf{x}) + \epsilon \cdot (\text{white noise}), \quad (42)$$

where $h_i(\mathbf{x})$ is shorthand for the i^{th} multi-Hermite polynomial $h_{\alpha_i}(\mathbf{z})$ and c_i is a mean-zero Gaussian random variable with variance $(i+6)^{-\beta}$. The so-called *source exponent* β satisfies $\beta > 1$ and controls the Sobolev smoothness of the target: the larger β is, the smoother and easier-to-learn the target. We choose a numerical truncation threshold $P = 30,000$ for convenience, choosing the target noise level ϵ to ensure that the target is unit norm $\mathbb{E}[y_i] = 1$. Our results are empirically insensitive to the randomness in the target.

- **(Real.) class vs. class.** A binarization in which samples are only drawn from two classes.
- **(Real.) class vs. all others.** A binarization similar to a single output element of practical neural networks with one-hot label encodings. A key difference here is that samples are drawn from each binary superclass in equal proportion so that $\mathbb{E}[y_i] = 0$.
- **(Real.) Domesticated vs. wild animals.** CIFAR-5m binarization: [cat, dog, horse] vs. [bird, deer, frog].
- **(Real.) Vehicles vs. animals.** CIFAR-5m binarization: [plane, car, ship, truck] vs. [bird, cat, deer, dog, frog, horse]. Samples are drawn from each superclass in equal proportion so that $\mathbb{E}[y_i] = 0$.
- **(Real.) Even vs. odd.** SVHN binarization based on parity: [0, 2, 4, 6, 8] vs. [1, 3, 5, 7, 9].
- **(Real.) Prime vs. composite.** SVHN binarization based on primality: [2, 3, 5, 7] vs. [4, 6, 8, 9]. We leave out [0] and [1], numerals whose primality is undefined.
- **(Real.) Genus 0 vs. genus ≥ 1 .** SVHN binarization based on the numeral’s topological genus: [1, 3, 5, 7] vs. [0, 6, 8, 9]. We leave out [2] and [4], numerals whose topological genus is font-dependent.

D.2 DIRECT EIGENSTRUCTURE CHECKS

What is the appropriate way to numerically compare two eigensystems?

The spectra are easy to compare – one can simply check whether $\frac{|\lambda_i - \hat{\lambda}_i|}{\lambda_i}$ is small for all i . An easy visual check is to simply scatter one spectrum against the other on a log-log plot; if the points remain close to the $y = x$ line, then the spectra agree.

Comparing the eigenbases, on the other hand, is a subtler matter. One must be careful when the eigensystems have small eigenvalue gaps. This issue is most easily understood by considering the limit: what happens when comparing two diagonalizations of a degenerate matrix? In this case, numerical eigendecomposition is undefined since the true eigenvectors are not unique; the computed eigenvectors are thus arbitrary. Simply comparing $\hat{\phi}_i$ with ϕ_i for all i is therefore insufficient.

Clearly, any good eigenbasis comparison must be spectrum-aware. In particular, differences between eigenvectors belonging to (near-)degenerate subspaces should not be strongly penalized.

A coarse but simple way to make this comparison is with spectral binning. We divide $\mathbb{R}^+$ into logarithmically-spaced bins; then, for each eigenbasis, we treat the modes whose eigenvalues fall within the same bin as a single near-degenerate eigenspace. Applying this procedure to the HEA¹⁰ yields a set of disjoint Hermite eigenspaces $\{\Phi_i^{(\text{th})}\}_{i=1}^{\text{nbins}}$, and likewise for the empirical basis. Let $d_i^{(\text{th})} = \dim(\Phi_i^{(\text{th})})$ and likewise for $d_i^{(\text{emp})}$. Note that in general we do not expect $d_i^{(\cdot)}$ to equal $d_j^{(\cdot)}$ for $i \neq j$; indeed, some bins may contain no modes at all. However, we *do* expect $d_i^{(\text{th})} = d_i^{(\text{emp})}$ for all i (if the theory is accurate).

Having handled any issues of spectral near-degeneracy, we may directly compare the two eigenbases by computing the pairwise overlaps between the binned eigenspaces:

$$\text{Overlap}(i, j) = \begin{cases} \frac{1}{d_j^{(\text{emp})}} \|\Phi_i^{(\text{th})\top} \Phi_j^{(\text{emp})}\|_{\text{F}}^2, & d_i^{(\text{th})} \neq 0 \text{ and } d_j^{(\text{emp})} \neq 0, \\ \text{undefined}, & \text{otherwise.} \end{cases} \quad (43)$$

where again $1 \leq i, j \leq \text{nbins}$ enumerate the bins.

To visualize this in Figure 2, we plot the overlaps in a heatmap, graying out pixels whose spectral bins do not contain any eigenmodes (and thus have undefined overlap). We note that discrepancies in the tails are primarily caused by distortions in the *empirical* eigenbasis caused by finite-size effects, rather than genuine disagreement between the theory and the true eigensystem (see Figure 5).

The experiments that generated Figure 2 used the following hyperparameters:

- **Gaussian kernel, gaussian data.** Kernel width $\sigma = 6$. Data $\mathbf{x} \in \mathbb{R}^{200}$ drawn i.i.d. Gaussian with diagonal covariance $\mathbb{E}[x_i^2] = (i + 6)^{-3.0}$. This results in $d_{\text{eff}} \approx 7$. We choose a steep covariance decay exponent to ensure that cubic modes are clearly present in the plot.
- **Gaussian kernel, CIFAR-5m.** Kernel width $\sigma = 6$. No ZCA nor sample normalization. This results in $d_{\text{eff}} \approx 9$.
- **Laplace kernel, SVHN.** Kernel width $\sigma = 8\sqrt{2}$. Whiten data with ZCA strength $\omega^2 = 5 \times 10^{-3}$ and then unit-normalize. This results in $d_{\text{eff}} \approx 21$. We generally observed that $d_{\text{eff}} \geq 20$ is a reliable rule of thumb for obtaining good agreement with the HEA using the Laplace kernel.
- **ReLU NTK, ImageNet-32.** Bias variance $\sigma_b^2 = 1.68$ and weight variance $\sigma_w^2 = 0.56$. Whiten data with ZCA strength $\omega^2 = 5 \times 10^{-3}$ and then unit-normalize. This results in $d_{\text{eff}} \approx 40$. Examining the on-sphere coefficients, we see that $\sigma_b^2/\sigma_w^2 \gg 1$ is the ReLU NTK equivalent for the wide kernel condition for Gaussian kernels.

D.3 HOW TO DECOMPOSE THE TARGET FUNCTION

We are interested in recovering the coefficients of the target f_* in the kernel eigenbasis:

$$\text{Recover } v_i \text{ from samples of } f_*(\mathbf{x}) = \sum_i v_i \phi_i(\mathbf{x}), \quad (44)$$

where i is the mode index. According to the HEA, this amounts to expanding f_* in the multi-Hermite basis:

$$f_*(\mathbf{x}) = \sum_i \tilde{v}_i h_i(\mathbf{x}). \quad (45)$$

where $h_i(\mathbf{x})$ is shorthand for the i^{th} multi-Hermite polynomial $h_{\alpha_i}(\mathbf{z})$. We use different notation for the Hermite coefficients since the HEA will not hold exactly in practical settings.

Of course, obtaining any full expansion of f_* from N samples is exactly as hard as the original learning problem. However, we *can* hope to obtain an expansion that is good enough to predict the behavior of KRR trained with up to N samples.

¹⁰Here, we abuse terminology by referring to the proposed Hermite basis as an eigenbasis; evaluated on finitely many samples of real data, these basis vectors may not be truly orthonormal.

Let us define $\mathbf{y} \in \mathbb{R}^N$ as the vector of target samples and $\mathbf{h}_i \in \mathbb{R}^N$ as the i^{th} multi-Hermite polynomial evaluated on the samples. Let us stack the top $P < N$ modes $[\mathbf{h}_1 \ \mathbf{h}_2 \ \dots \ \mathbf{h}_P]^\top$ and call the resulting matrix $\mathbf{H} \in \mathbb{R}^{P \times N}$.¹¹ We would like to recover the top P coefficients $\hat{\mathbf{v}} \in \mathbb{R}^P$ from $\mathbf{y}$ and $\mathbf{H}$.

Naïve first try. As a first pass, let’s simply run linear regression: $\hat{\mathbf{v}} = \mathbf{H}^\dagger \mathbf{y}$, where $\mathbf{H}^\dagger$ is the pseudo-inverse. Empirically, the recovered coefficients are not very good. There are two main reasons.

- The true target contains Hermite modes which are not represented in our top P list. This is an essential model misspecification which appears to linear regression as target noise. This is not a problem in itself; the learning curve eigenframework can handle this kind of noise, *if* it knows how much noise power there is. However, it is unclear how to measure the magnitude of this misspecification noise, since it gets mixed in with the sampling “noise.”
- Linear regression has a flat inductive bias: it tends to fit the samples using each of its regressors with equal enthusiasm. Here, the regressors are functions, arranged (roughly) in order of decreasing smoothness. On the other hand, natural targets are relatively smooth; their power tends to concentrate on the early modes. This leads to large estimation errors in the top coefficients.

These two challenges, left unresolved, result in poorly predicted learning curves.

Second pass. Let us instead consider the following greedy iterative algorithm:

Algorithm 1 Greedy Residual Fitting (GRF)

Require: Target vector $\mathbf{y} \in \mathbb{R}^N$, Hermite feature matrix $\mathbf{H} \in \mathbb{R}^{P \times N}$.

```

1:  $\mathbf{y}^{(0)} \leftarrow \mathbf{y}$ 
2: for  $p = 1$  to  $P$  do
3:    $\mathbf{h}_p \leftarrow \mathbf{H}_{p,:}$ 
4:    $\hat{v}_p \leftarrow \frac{\langle \mathbf{h}_p, \mathbf{y}^{(p-1)} \rangle}{\|\mathbf{h}_p\|_2^2}$  ▷ find the coefficient of  $\mathbf{h}_p$  in  $\mathbf{y}^{(p-1)}$ 
5:    $\mathbf{y}^{(p)} \leftarrow \mathbf{y}^{(p-1)} - \hat{v}_p \mathbf{h}_p$  ▷ subtract off the projection onto  $\mathbf{h}_p$ 
6: end for
7: return  $\hat{\mathbf{v}} = (\hat{v}_1, \dots, \hat{v}_P)$  and  $\hat{\epsilon}^2 = \|\mathbf{y}^{(P)}\|_2^2$ 
```

This target recovery algorithm has a few notable virtues. First, it is *direct*; it does not require matrix inverses, so it runs in quadratic rather than cubic time. Second, since it is iterative and stateful, it automatically prioritizes correctly estimating the top coefficients, which is where the target power tends to lie. Third, it naturally estimates the noise power (it is simply the norm of the final residual $\|\mathbf{y}^{(P)}\|^2$). Fourth, it estimates the correct amount of total power: by the Pythagorean theorem, $\sum_p \hat{v}_p^2 + \hat{\epsilon}^2 = \|\mathbf{y}\|^2$ if $\|\mathbf{h}_p\| = \sqrt{N}$, that is, if the Hermite polynomials are indeed unit norm.¹²

We empirically find that this algorithm works well for synthetic targets on synthetic Gaussian data. However, real data is not perfectly Gaussian. Even small deviations from Gaussianity introduce *systematic* correlations between the Hermite polynomials $\{\mathbf{h}_i\}$; these overlaps cause the greedy algorithm to systematically overestimate the target power in the affected modes.

We empirically found that if the target is sufficiently “dense” in the Hermite modes (i.e., not dominated by very few coefficients), then these errors “average out” (loosely speaking) and the predicted learning curves remain accurate. Unfortunately, this is rarely the case in real targets. Real targets often contain much power in a few early modes, many of which suffer from this overlap problem. As a result, the predicted learning curves are often far off the mark. See [Figures 8 and 10](#) for empirics.

¹¹In our experiments we choose $P = 30,000$ and $N = 80,000$ since manipulating this tensor maximizes our GPU VRAM utilization.

¹²If they are not unit norm, then for the purposes of the KRR eigenframework, one usually wants to take $\hat{v}_p \mapsto N^{-1/2} \|\mathbf{h}_p\| \hat{v}_p$ to estimate the total amount of power in Hermite direction p . After adopting this scaling, GRF again conserves total power by the Pythagorean theorem: $\sum_p \hat{v}_p^2 + \hat{\epsilon}^2 = \|\mathbf{y}\|^2$.

Third pass. To rectify these modes, we simply squeeze out the non-orthogonality, starting from the top Hermite modes (which we expect to be the most accurate matches to the true eigenmodes). A standard technique for iteratively eliminating overlaps is to use the Gram-Schmidt process. In practice, this simply amounts to performing a QR decomposition, $QR = H$, and estimating the coefficients from the orthogonal component as $\hat{v} = Qy$.

The major strength of this coefficient recovery technique is that it accurately predicts learning curves. Another strength is that since the regressors are orthogonal, the noise power is easily obtained as $1 - \hat{v}^\top \hat{v}$. The main drawback is that we must once again resort to a cubic-time algorithm. A natural question is then: if we are going to run a cubic-time algorithm anyways, why not simply diagonalize the empirical kernel matrix? There are three reasons:

- *Finite-size effects.* Diagonalizing the finite-sample empirical kernel matrix typically introduces distortions in the tail modes. The HEA avoids this issue by constructing the basis directly.
- *Numerical conditioning.* The (non-orthogonalized) Hermite polynomial matrix H is very well-conditioned on natural datasets. Performing Gram-Schmidt orthogonalization tends to be numerically robust, even at 32-bit floating-point precision. On the other hand, the numerical conditioning of kernel diagonalization worsens as the number of samples (and retrievable modes) increases; in practice, we typically need 64-bit precision to obtain reliable eigenvectors. As a consequence, the prefactors for GPU VRAM space complexity are friendlier for Gram-Schmidt.
- *Theoretical insight.* Kernel diagonalization is an opaque numerical computation which does not easily expose the functional form of the eigenfunctions. The perturbed closed-form expression offered by HEA + Gram-Schmidt reveals the analytical structure of learning in kernels.

Takeaway. The HEA conceptually holds; the kernel eigenfunctions are small perturbations of the multi-Hermite polynomials, even for complex real data. If the target function is sufficiently dense, the perturbations are not important to model, and the target coefficients can be estimated using a direct greedy method. However, for some real targets with prominent coefficients, we must be careful to account for their perturbations.

D.4 LEARNING CURVES AND SAMPLE COMPLEXITIES

In this section, we detail the parameters used in each of our main experiments. All kernel regression experiments use a ridge of $\delta = 10^{-3}$ and run up to 50 trials, each with a new test set of size 5,000. All target function estimates use 80,000 samples.

- **Figure 1, top right.** Gaussian kernel, width $\sigma = 4$. No ZCA nor normalization.
- **Figure 1, bottom right.** Laplace kernel, width $\sigma = 4\sqrt{2}$. ZCA strength $\omega^2 = 10^{-3}$ with data sample normalization. Sample complexities are obtained by computing learning curves (empirical curves averaged over 20 trials), performing logarithmic interpolation, and finding where the test risk falls below 0.5.
- **Figure 3, top left.** Gaussian kernel, width $\sigma = 8$. No ZCA nor normalization. We use a wide kernel for better agreement at targets of high degree; for comparison with a narrower kernel, see [Figure 6](#) below.
- **Figure 3, top right.** Same as [Figure 1](#), bottom right.
- **Figure 3, bottom left.** Gaussian kernel, width $\sigma = 10$. No ZCA, but the samples are normalized. For comparison with a narrow kernel and no sample normalization, see [Figure 7](#) below.
- **Figure 3, bottom center.** Laplace kernel, width $\sigma = 4\sqrt{2}$. ZCA strength $\omega^2 = 5 \times 10^{-3}$ with data sample normalization. For comparison with no sample normalization, see [Figure 9](#).
- **Figure 3, bottom right.** ReLU NTK, bias and weight variances $\sigma_b^2 = 1.96$ and $\sigma_w^2 = 0.49$. ZCA strength $\omega^2 = 10^{-2}$ with data sample normalization. We found that the HEA for the ReLU NTK is particularly sensitive to insufficient effective dimension.

D.5 ADDITIONAL EXPERIMENTS

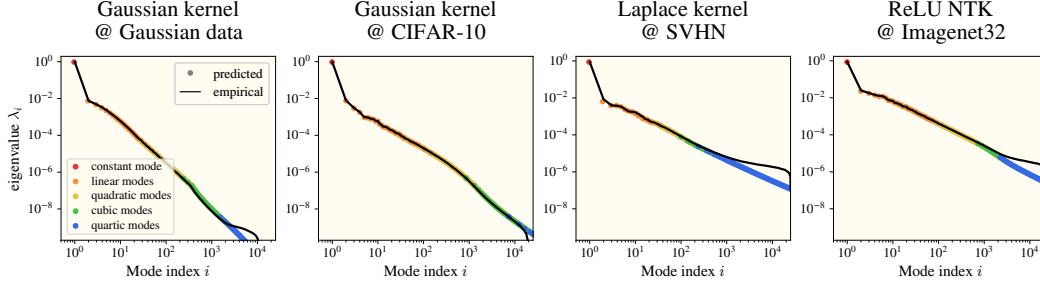


Figure 5: We show an alternative visualization of the top row of Figure 2, comparing the predicted and empirical spectra of various kernel/dataset combinations. We see that the HEA accurately predicts the minute details of the kernel spectrum. Furthermore, tail deviations are indeed caused by finite-kernel effects in the empirical spectrum rather than a failure of the HEA.

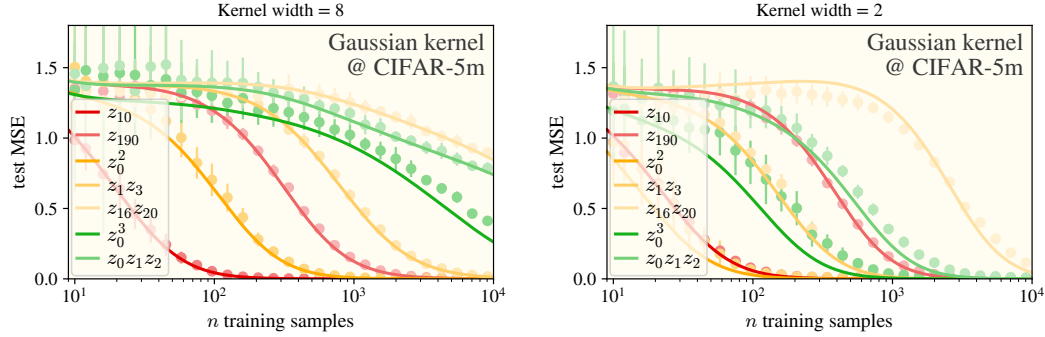


Figure 6: We compare the original plot from Figure 3 with kernel width 8 (left) to the same experimental setup except for a kernel width 2 (right). The true eigenfunctions of the narrower kernel deviate from the HEA prediction, especially for the high-degree modes. For a similar comparison, see Figure 12.

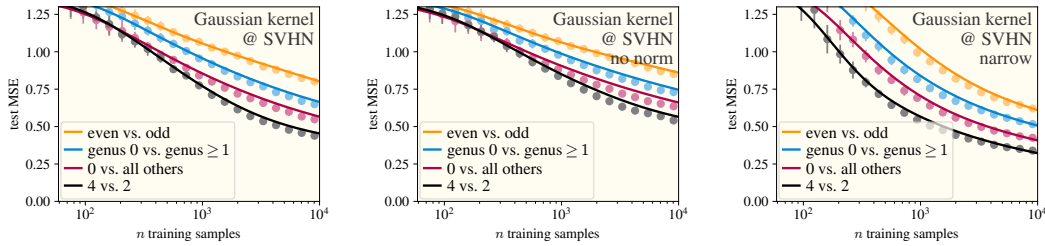


Figure 7: We compare the original experimental setup from Figure 3 (sample normalization, kernel width 10) on the left, to two similar setups: no sample normalization (center) and a narrower kernel width of 4 (right). Interestingly, the narrow kernel width does not change the agreement much; removing sample normalization worsens agreement slightly.

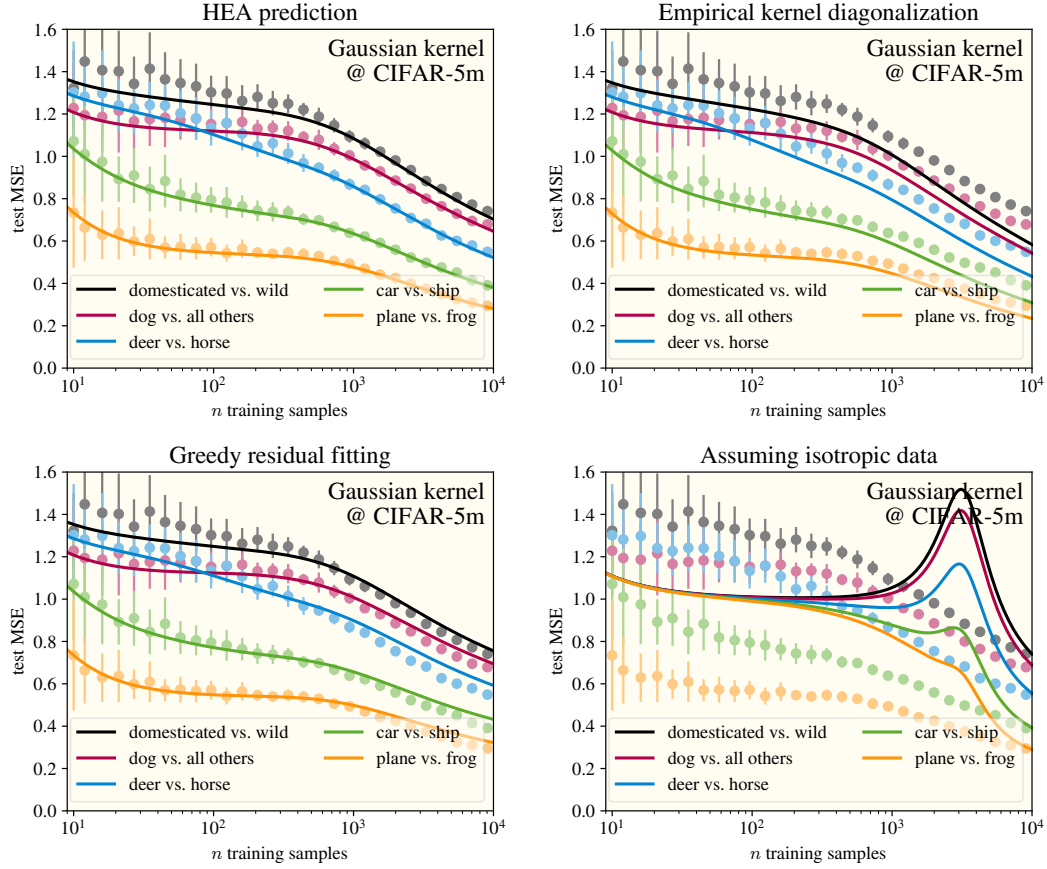


Figure 8: Here we show four plots in which the experimental setup does not change; instead, we compare different techniques for estimating the required theoretical quantities. **(Upper left.)** HEA prediction, identical to Figure 1. **(Upper right.)** We numerically diagonalize the kernel matrix (size 25000×25000) and use the obtained eigenstructure to make predictions. The accuracy of the prediction degrades as the number of training samples approaches the size of the kernel matrix used to estimate the true eigenstructure. **(Lower left.)** We use the greedy algorithm described in Algorithm 1 to estimate the target coefficients. We see good initial agreement, but it degrades due to accumulating non-orthogonality. **(Lower right.)** For comparison, we include the predictions one would obtain if one modeled the data distribution as an isotropic Gaussian. Clearly, a major contribution of our work is to handle the anisotropy in natural data, since it strongly affects the resulting learning curves.

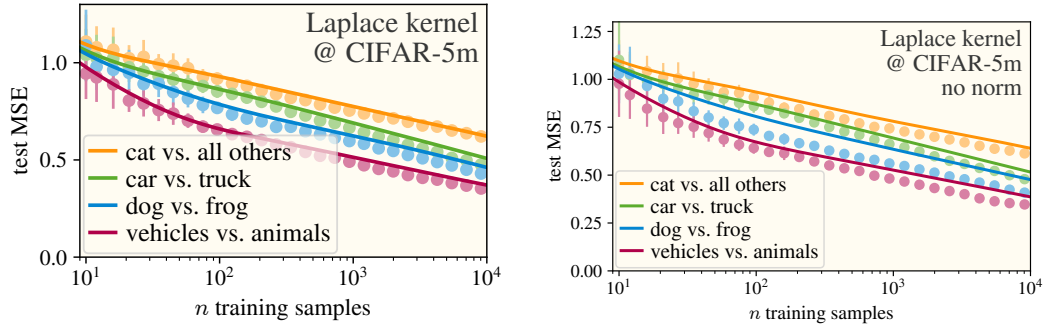


Figure 9: We compare the plot from Figure 3 on the left with an identical experimental setup, except without sample normalization. We see that normalization is necessary for obtaining agreement, since the HEA is derived using the kernel’s on-sphere level coefficients.

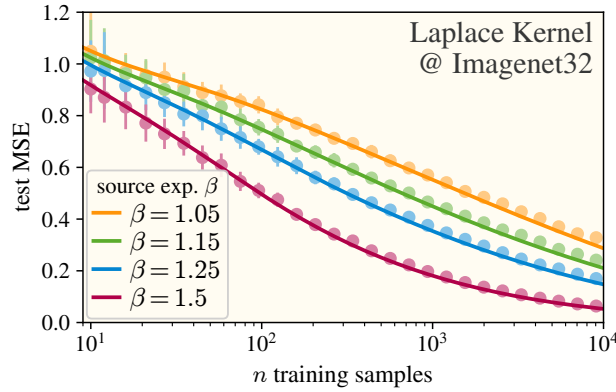


Figure 10: For “dense” synthetic targets (i.e., targets whose power is split among many eigenfunctions), the HEA predicts the performance of the Laplace kernel very well. Contrasting this with the predictions for Laplace kernel on real targets (Figure 9), we conclude that dense targets are more forgiving of errors in target coefficient recovery. For this experiment, we use ZCA strength $\omega^2 = 10^{-2}$ and sample normalization.

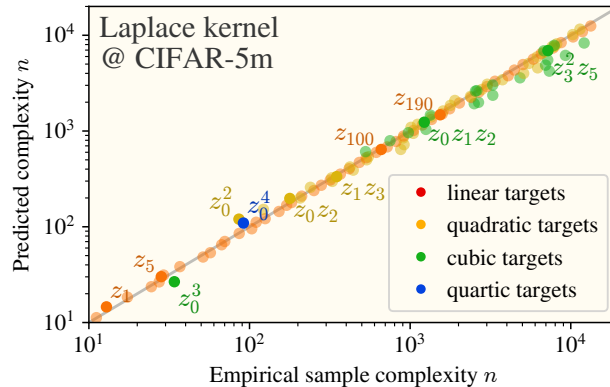


Figure 11: Additional sample complexity plot, this time for Laplace kernel on CIFAR-5m. We use sample normalization and a ZCA strength of 3×10^{-3} .

E WHAT BREAKS THE HEA? CASE STUDIES OF THREE CAUSES OF FAILURE

In the main text, we have primarily shown results in settings in which the HEA works quite well in predicting kernel eigenstructure, learning curves, and monomial learning rates. Here we get our hands dirty and show where it breaks.

First cause of failure: narrow kernel width. The discussion of [Section 4.1](#) suggests that if we make kernel width too narrow — or, in kernels without a notion of width, change the kernel so that $\frac{c_{\ell+1}\gamma_1}{c_\ell}$ is not much less than one — we should start to see the HEA break. In particular, since for a Gaussian kernel $c_\ell \propto \sigma^{-2\ell}$, narrow width means non-small level coefficient ratios and vice versa. [Figure 12](#) shows an empirical test: on the same synthetic data distribution, the kernel is made progressively narrower. As expected, the HEA breaks.

Second cause of failure: low data dimension leading to non-concentration of norm. The HEA treats all rotation-invariant kernels as if they were dot-product kernels. In fact, it throws away all information about the kernel function that cannot be obtained by its value on a sphere. For this to work in practice, we should expect our data to lie fairly close to a sphere. For Gaussian data $\mathbf{x} \sim \Sigma$, this is assured if the data has high effective dimension $d_{\text{eff}} := \frac{\text{Tr}[\Sigma]^2}{\text{Tr}[\Sigma^2]}$. [Figure 13](#) shows that with a Laplace kernel on a Gaussian dataset, the HEA breaks as d_{eff} decreases.

Interestingly, not all rotation-invariant kernels require high effective dimension. For example, our heuristic derivation of the HEA for Gaussian data in [Section 4.1](#) took place in just one dimension! As it turns out, this is because the Gaussian kernel is given by a polynomial feature map (of infinite dimension), but the Laplace kernel has a cusp at $\mathbf{x} = \mathbf{x}'$ and is not. (Equivalently, the Gaussian kernel is analytic, but the Laplace kernel is not.) For a stationary kernel $K(\mathbf{x}, \mathbf{x}') = k\left(\frac{1}{\sigma^2} \|\mathbf{x} - \mathbf{x}'\|^2\right)$ which depends only on the distance between points with scaling factor σ , the $\sigma \rightarrow 0$ limit will give the HEA on Gaussian data if the function $k(\cdot)$ is real-analytic at zero. The Gaussian kernel satisfies this real-analytic criterion, but the Laplace kernel does not. As evidence of this, [Figure 14](#) repeats the above experiment with a Gaussian kernel and finds that low d_{eff} is no problem.

Third cause of failure: the data not being “Gaussian enough.”

The HEA is theoretically derived for Gaussian data, but nonetheless works for some real datasets. This suggests that the datasets for which it works are “Gaussian enough” in some sense. Here we support this intuition with experiments. First, in [Figure 15](#), we plot the marginal distributions of the first few normalized principal coordinates for four datasets of decreasing complexity: CIFAR-10, SVHN, MNIST, and the tabular UCI Mushrooms dataset ([Dua & Graff, 2017](#)). Plotted this way, these datasets appear increasingly non-Gaussian. Then, in [Figure 16](#), we test the HEA on these four datasets, finding that the HEA indeed seems to work worse as the dataset becomes “less Gaussian.” For good measure and out of interest, we plot the first few PCA coordinates for the three image datasets in [Figure 17](#).

Of course, we put “Gaussian enough” in quotes because this is a nonrigorous and ill-defined notion: we have not invoked any precise method of determining which of two datasets is “closer” to Gaussian! We do not attempt to give a precise definition here. This seems like a worthwhile direction for future exploration, especially from the statistics community.

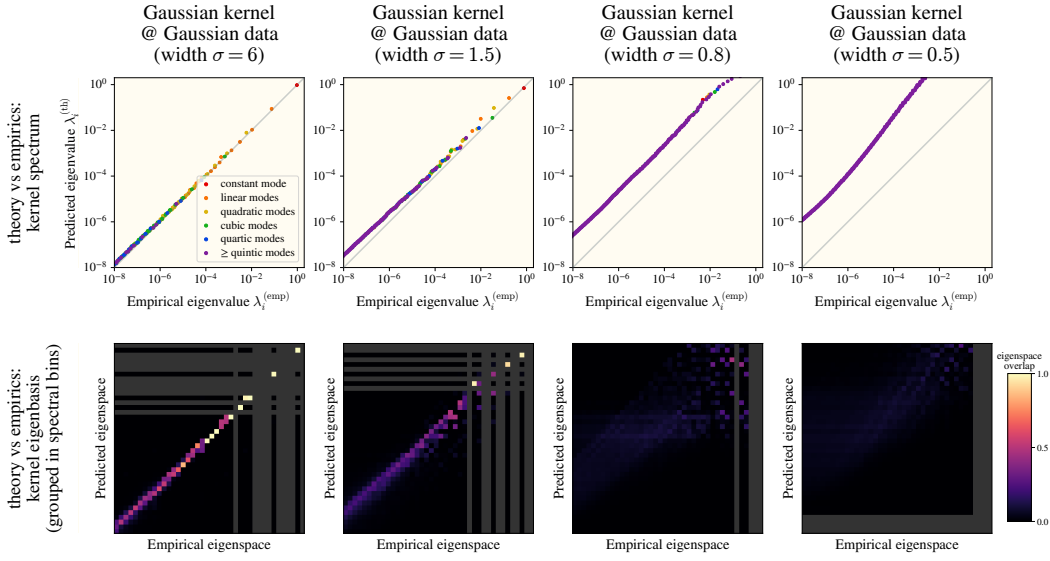


Figure 12: **The HEA breaks when kernel width is too narrow.** We repeat the experiment of Figure 2 with the Gaussian kernel function $K(x, x') = e^{-\frac{1}{2\sigma^2} \|x - x'\|^2}$ and synthetic Gaussian data. The data is generated as $x \sim \mathcal{N}(0, \Gamma)$, where Γ has eigenvalues $\gamma_i \propto i^{-3}$ for $i = 1 \dots 30$, with $\gamma_1 \approx 0.83$ after normalization. As the kernel width σ shrinks to $\sigma \lesssim 1$, the HEA breaks: eigenvalue prediction fails with higher-order modes dominating predictions, and the clear eigenspace overlap present at large width evaporates.

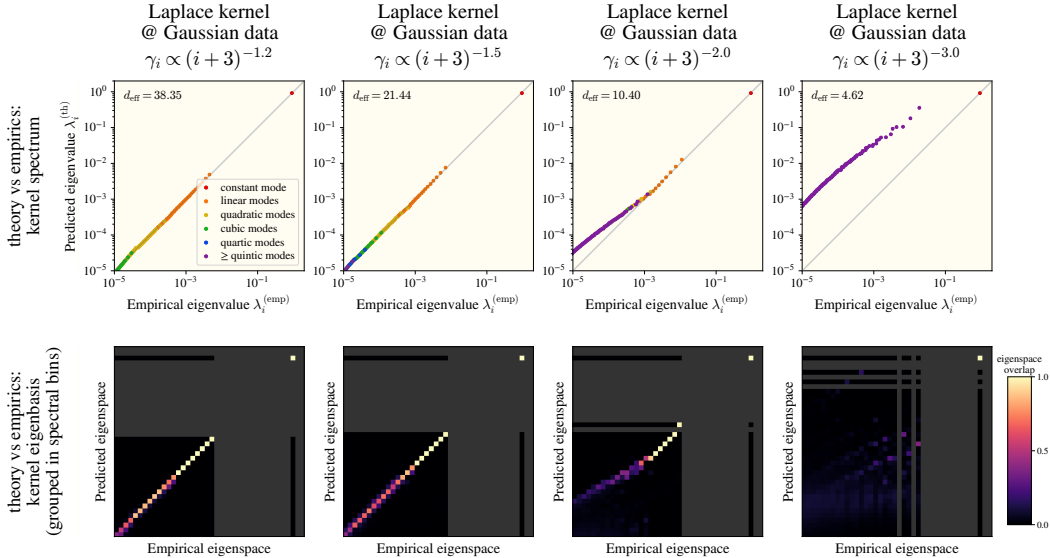


Figure 13: **With the Laplace kernel, the HEA requires high effective data dimension...** We repeat the experiment of Figure 12 with two modifications: first, the kernel function is Laplace $K(x, x') = e^{-\frac{1}{\sigma} \|x - x'\|}$ with width $\sigma = 8\sqrt{2}$, and second, the eigenvalues are $\gamma_i \propto (i + 3)^{-\alpha}$ with variable exponent α . As α increases (moving left to right), the effective dimension $d_{eff} = \frac{(\sum_i \gamma_i)^2}{\sum_i \gamma_i^2}$ decreases, and agreement with the HEA degrades. Empirically, we find that $d_{eff} \approx 20$ is usually high enough to see good agreement.

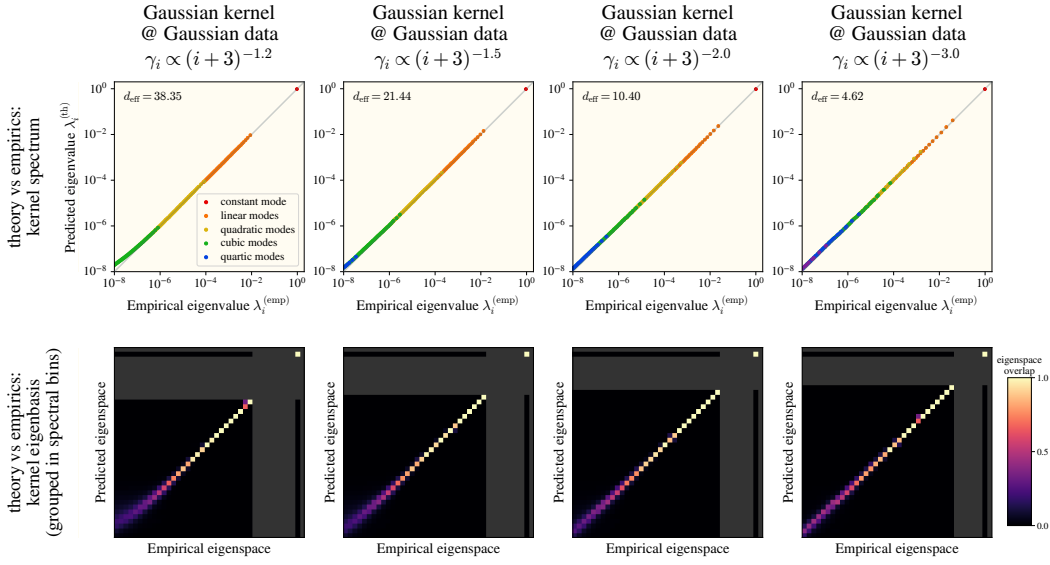


Figure 14: ...but with the Gaussian kernel, it does not need high effective dimension. We repeat the experiment of Figure 13 exactly, but use a Gaussian kernel with width $\sigma = 3$. Thanks to the smoothness of the Gaussian kernel, we see little degradation of the theory-experiment match of the HEA.

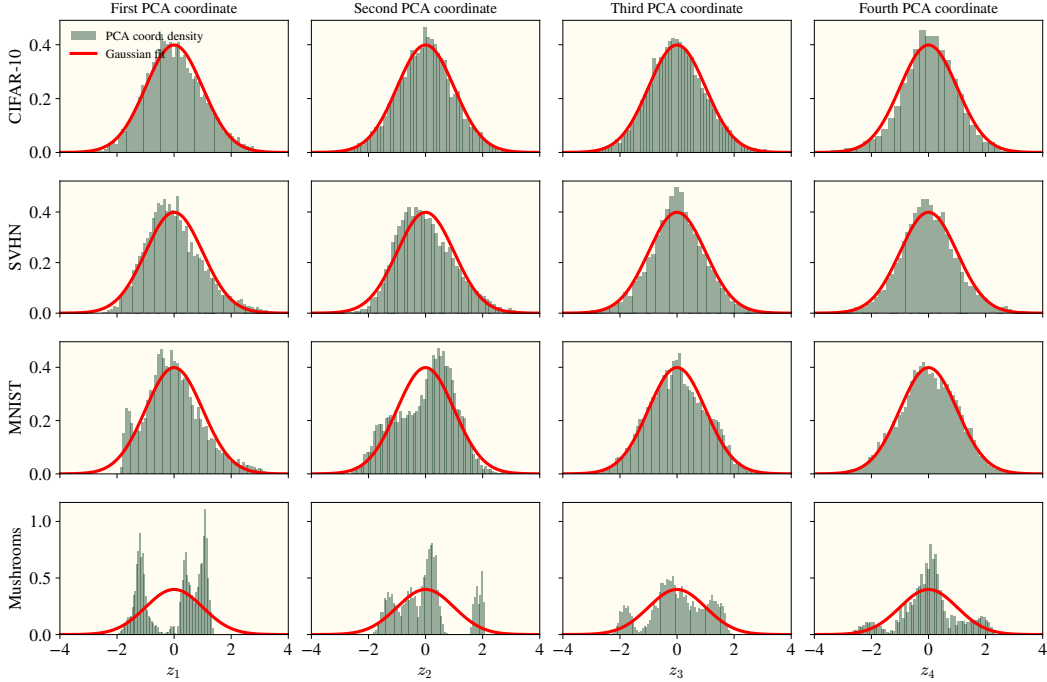


Figure 15: **Different datasets have different degrees of Gaussianity in their early principal components.** For four natural datasets — CIFAR-10, SVHN, MNIST, and the UCI Mushrooms tabular dataset — we compute the first four normalized principal coordinates $(z_i)_{i=1}^4$ and plot their distributions as histograms. We find that the first few principal coordinates for CIFAR-10 are fairly close to Gaussian, this is less true for SVHN, yet less true for MNIST, and not at all true for the Mushrooms dataset.

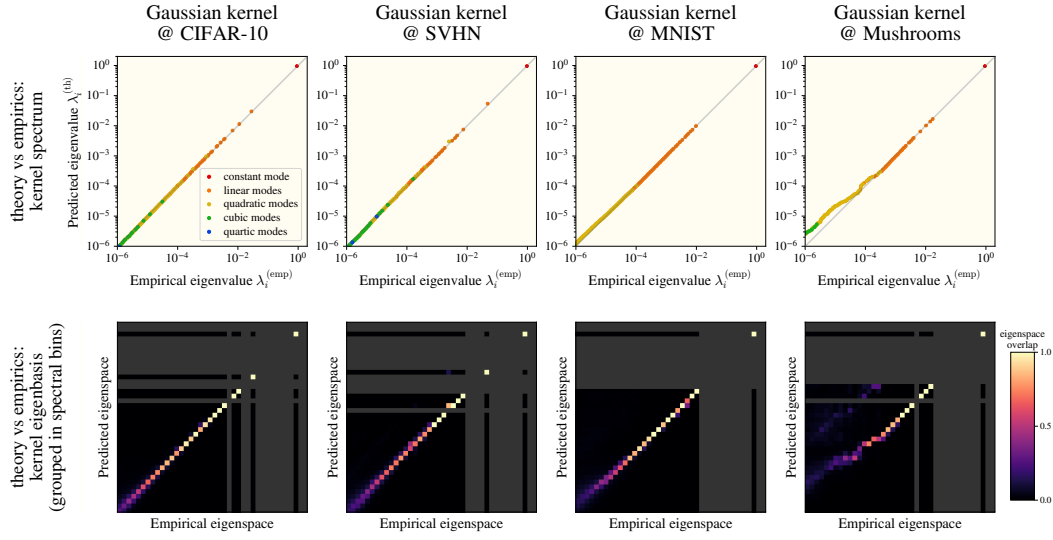


Figure 16: **The HEA works better with complex high-dimensional datasets and worse with simpler datasets.** We repeat the experiment of Figure 2 with a Gaussian kernel of width $\sigma = 3$ on 8000 samples from four datasets of increasing non-Gaussianity: CIFAR-10, SVHN, MNIST, and the UCI Mushrooms tabular dataset. As the dataset becomes simpler, the HEA works worse and worse, though the predicted trend is still roughly present, if quantitatively wrong, even for the tabular dataset.

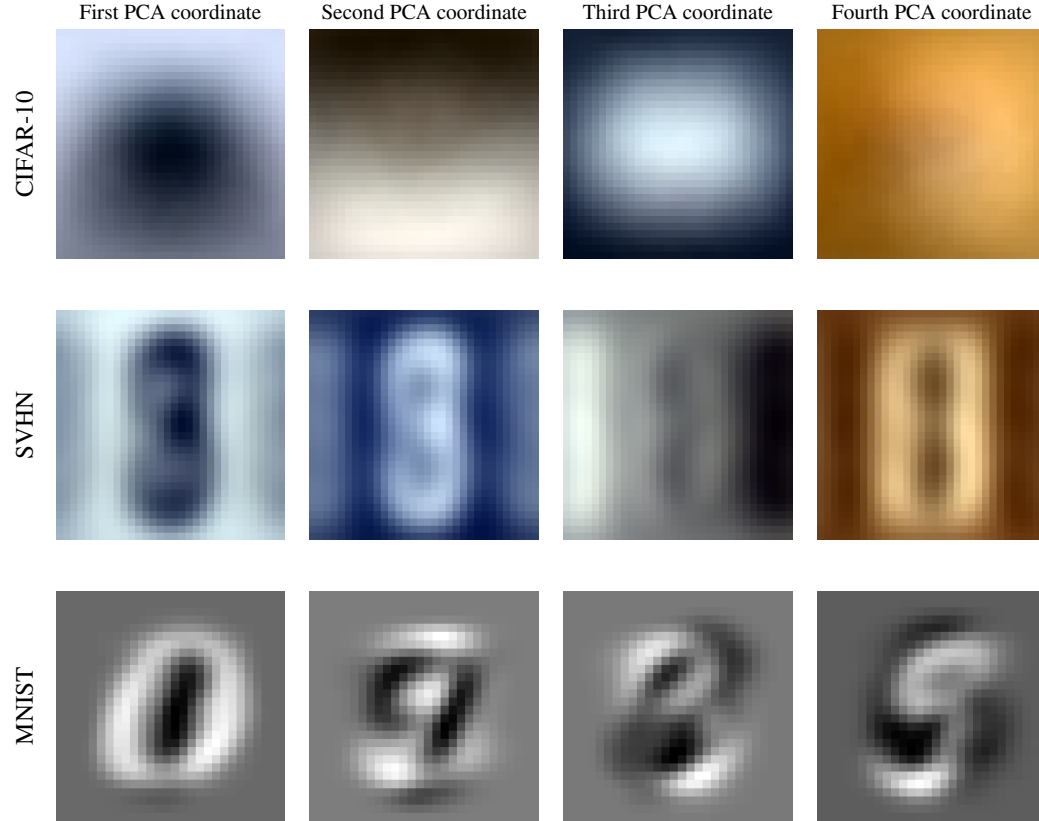


Figure 17: Visualization of the first four PCA directions for the CIFAR-10, SVHN, and MNIST datasets.

F MLP EXPERIMENTS: DETAILS AND FURTHER DISCUSSION

As the HEA requires level coefficients to predict λ_α , level coefficients for MLP experiments are computed with the ReLU NTK with both bias and weight variances of 1 which match the empirical variances of our MLP networks at initialization.

Unless otherwise stated, all experiments were performed using a feature-learning (μ P) MLP with the following data hyperparameters:

$$\begin{aligned}\gamma_i &= (i + 6)^{-\alpha} \\ \alpha &= 1.7\end{aligned}$$

and MLP hyperparameters,

$$\begin{aligned}\text{width } w &= 8192 \\ \text{depth (hidden layers+1)} L &= 3 \\ \text{learning rate } \eta &= 10^{-2} \\ \text{batch size } \text{bsz} &= 1024 \\ \zeta &= 1,\end{aligned}$$

where ζ is the richness parameter (defined by a rescaling of the network's output $f \mapsto f/\zeta$ and either learning rate $\eta_{\max} \mapsto \eta_{\max} \cdot \zeta^{2/L}$ for $\zeta > 1$ or $\eta_{\max} \mapsto \eta_{\max} \cdot \zeta^2$ for $\zeta < 1$) allowing a network to scale between an ultra-rich and a lazy/NTK regime (Atanasov et al., 2024). In order to avoid network outputs exploding for $\gamma \gg 1$, we have all MLPs output the difference $f(\mathbf{x}; \theta_t) - f(\mathbf{x}; \theta_0)$ for network parameters at gradient step t being θ_t .

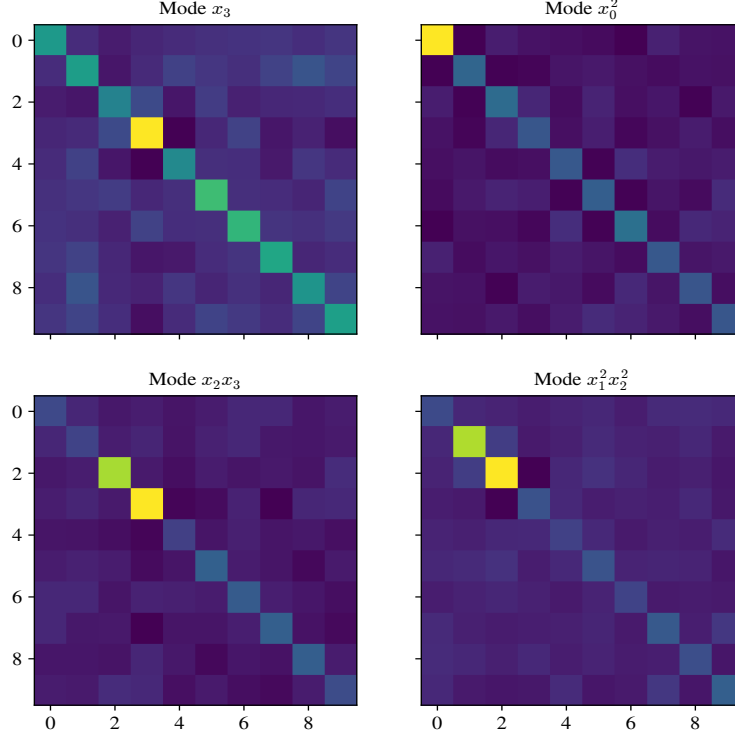


Figure 18: **Validation of MLPs being feature-learning.** We investigate the Gram matrix of the first weight matrix, $W_1^T W_1$. We train MLPs with $\gamma = 1$ on $d = 10$ Gaussian synthetic data for 1000 GD steps. Yellow denotes higher entries, while blue denotes near-0 entries. The Gram matrices are heavily weighted to modes present in the target function, confirming the feature learning regime.

All experiments were carried out in an online setting, with n_{iter} being the number of SGD steps it took to reach a train error $\text{MSE}_{\text{tr}} \leq 0.1$. Random samplings of $n=\text{bsz}$ Gaussian datapoints produce variance in high-order Hermite polynomials, so an exponential moving average (EMA) for MSE_{tr} is used with decay constant 0.9: $(\text{EMA}_{\text{tr}, t} = 0.9\text{MSE}_{\text{tr}, t} + 0.1\text{EMA}_{\text{tr}, t-1})$. This corresponds with a half-life for the train error of 6.58 steps: $\text{MSE}_{\text{tr}, t} \sim 0.5\text{MSE}_{\text{tr}, t-6.58}$. This ensures high-variance random samplings do not cause the effective optimization time $\eta \cdot n_{\text{iter}}$ to be skewed in favor of higher order modes. We varied the EMA constant and found values less than 0.9 to underestimate high order modes' optimization time, and values greater than 0.9 to overestimate lower order modes' optimization time. One could subtract the half-life from an empirically found n_{iter} with any EMA constant to get a more EMA-unbiased estimate for when the test error goes below the set cutoff, although we do not perform such a procedure. We additionally varied what the cutoff in $\text{MSE}_{\text{tr}} \leq \text{cutoff}$ was, although we found no large changes in findings for reasonable cutoff values.

This network is empirically validated to be in the feature learning regime by looking at the Gram matrix $W_1^\top W_1$, with $W_1 \in \mathbb{R}^{\text{width} \times \text{dim}}$ initialized with Gaussian entries. A lazy network has this Gram matrix constant throughout training, whereas feature learning allows for $\Theta(1)$ changes to the spectral norm of the W_1 , producing $\Theta(1)$ changes to $W_1^\top W_1$ (Yang et al., 2023). We train $\gamma = 1$ networks on simple PCA-mode target functions, and find significantly more power in the Gram matrix entries corresponding to the target function, validating the feature-learning regime Figure 18.

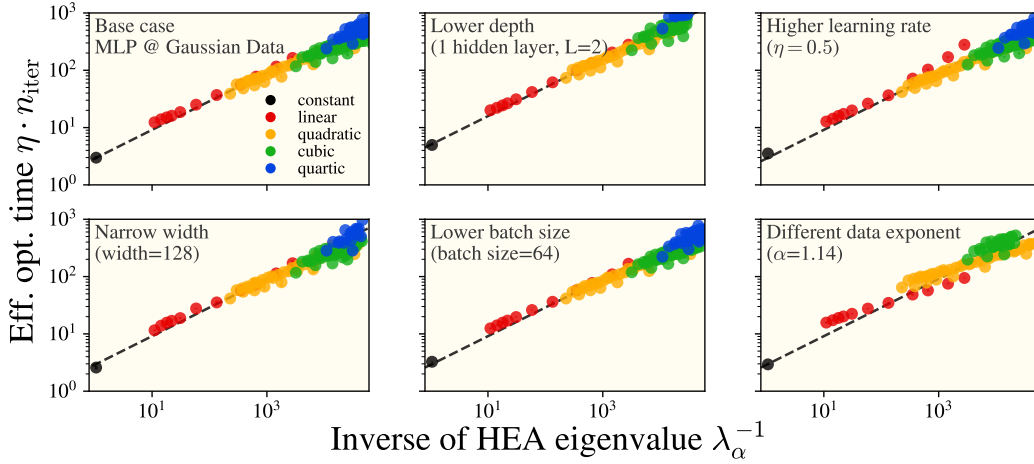


Figure 19: **Hyperparameter choices are largely arbitrary.** We train MLPs on synthetic Gaussian data, with one hyperparameter being modified. The base case is shown in the (top left) for reference. Apart from the (top center), all modifications produce no notable changes to our expected $\lambda_\alpha^{-1/2} \sim \eta \cdot n_{\text{iter}}$ scaling. (Top center). The 1 hidden layer case is only different than the rest by a change in eigenvalue-independent prefactor, the scaling law still holding.

We check if our hyperparameter choices were simply lucky, or if our $\lambda_\alpha^{-1/2} \sim \eta \cdot n_{\text{iter}}$ scaling is real and reproducible. We take our base case, and vary all hyperparameters (depth, width, learning rate, batch size, data covariances) we don't expect to affect the scaling, finding that none do; the only change we observe is an eigenvalue-independent rise of effective optimization time when going to a more shallow network. Results are summarized in Figure 19.

The only hyperparameter we know will change the exponent of $\lambda_\alpha^{\text{exp.}} \sim \eta \cdot n_{\text{iter}}$ is the richness parameter ζ ; with $\zeta \ll 1$ being known to have exponent -1 . What is unknown is the scaling exponent as $\zeta \gg 1$, which we find to be lower than the baseline $1/2$. Lastly, our effective optimization time's η is defined as the base, ζ -independent learning rate. Our findings are shown in Figure 20.

Lastly, we would like to note that the base synthetic case and the CIFAR-5m case were averaged over 3 trials, and upon finding a remarkably small standard deviation in effective optimization time, all further validation experiments were completed with only a single trial.

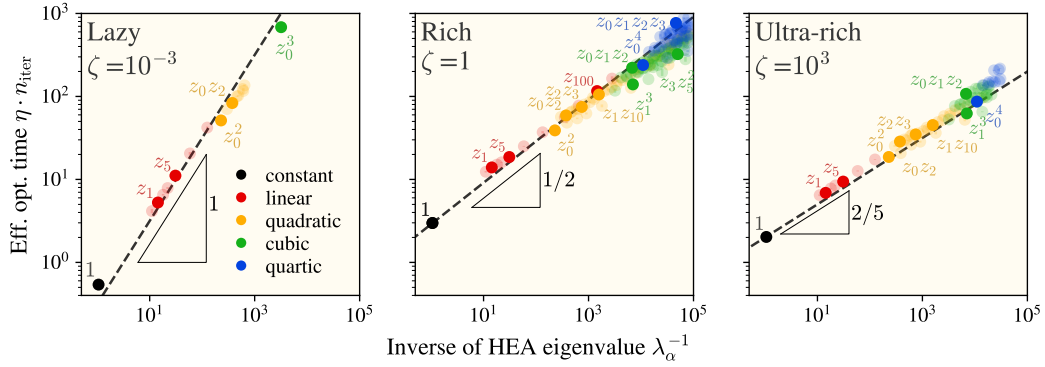


Figure 20: **Sweeping the richness parameter changes the slope.** (Left) We validate taking $\zeta \ll 1$ to bring us into the lazy/NTK regime, where we know the eigenvalue to be inversely proportional to the effective optimization time. (Right) We investigate the so-called "ultra-rich" regime $\zeta \gg 1$ and find there to be a $\lambda_\alpha^{-2/5} \sim \eta \cdot n_{\text{iter}}$ relationship. We have no expectation the $2/5$ factor for the ultra-rich regime holding for other hyperparameter choices, especially under a change of depth.

G REVIEW OF HILBERT SPACE THEORY

In this section we review the basics of Hilbert space theory, which would be very useful in proving theorems in the next few appendices.

Naïvely, a Hilbert space is just $\mathbb{R}^n$ or $\mathbb{C}^n$, where n is allowed to go to infinity. In the standard notation, it is written as $(\mathcal{H}, \langle \cdot | \cdot \rangle)$, where $\mathcal{H}$ is the set of vectors, and the angular brackets $\langle \cdot | \cdot \rangle$ is the inner product on pairs of vectors in the set.

G.1 DIRAC NOTATION

We use the standard Dirac notation.

- Vectors are written as $|v\rangle \in \mathcal{H}$.
- The duals of vectors on $\mathcal{H}$ are written as $\langle v|$, defined via the inner product: $\langle v| : |w\rangle \mapsto \langle v|w\rangle$.
- Linear operators are written as capital letters: A .
- Linear operators act on the right: $\langle v|A|w\rangle := \langle v|(|Aw\rangle)$.

For example, the Euclidean space $\mathbb{R}^d$ is a Hilbert space. Its inner product is the dot-product $\langle v|w\rangle = v^\top w$. Write column vectors as kets and row conjugates as bras:

$$|v\rangle = \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix}, \quad \langle v| = |v\rangle^\top = [v_1 \quad \cdots \quad v_d], \quad \langle v|w\rangle = \langle v| |w\rangle = v^\top w.$$

For a matrix $A \in \mathbb{R}^{d \times d}$, it acts on vectors by multiplication on the right. Rank-one operators are of the form $|v\rangle\langle w|$, with action $|x\rangle \mapsto |v\rangle\langle w|x\rangle$.

G.2 OPERATORS

In a Euclidean space, such as $\mathbb{R}^3$, the geometry is defined by its lengths and angles. A linear operator on $\mathbb{R}^3$ preserves lengths and angles, and therefore preserves its geometry, iff it preserves the dot-product between vectors. Such geometry-preserving linear operators are called orthogonal operators.

Generalized to Hilbert spaces, geometry-preserving linear operators are called **unitary**. An operator $V : \mathcal{H} \rightarrow \mathcal{K}$ is unitary iff

$$\langle Vv|Vw\rangle = \langle v|w\rangle, \quad \|V|v\rangle\| = \||v\rangle\|.$$

For example, a unitary $V : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is just an orthogonal matrix. Unitary maps preserve all geometric relationships, such as angles between vectors, lengths of vectors, and orthonormal bases.

The rank of an operator is the dimension of its range. By the analog of singular value decomposition in Hilbert space, if an operator has finite rank r , then it can be written in the form of $\sum_{k=1}^r |v_k\rangle\langle u_k|$.

In a Euclidean space, operators have transposes. The transposition of operators satisfies $(A^\top w)^\top v = w^\top (Av)$. Generalized to Hilbert space, operators have **adjoints**. The adjoint of an operator A is written as A^* , and satisfies

$$\langle w|Av\rangle = \langle A^*w|v\rangle$$

In a Euclidean space, symmetric operators are operators such that $A = A^\top$. In a Hilbert space, **self-adjoint** operators are operators such that $A = A^*$.

Just like how symmetric operators on $\mathbb{R}^d$ are of the form $\sum_{k=1}^d a_k v_k v_k^\top$, self-adjoint operators are of the form $\sum_k a_k |v_k\rangle\langle v_k|$, although the summation may be infinite.

The **operator norm** is a norm defined on operators. It is defined by $\|A\|_{op} := \sup_{x \in \mathcal{H}, \|x\|=1} \|Ax\|$.

An operator is **positive definite**, written as $A \succ 0$, iff it is self-adjoint, and $\langle v|Av\rangle > 0$ for all nonzero $|v\rangle$. Similarly, an operator is **positive semidefinite**, written as $A \succeq 0$, iff it is self-adjoint, and $\langle v|Av\rangle \geq 0$ for all $|v\rangle$.

G.3 COMPACT OPERATORS

In infinite dimensions, some finite-dimensional intuitions break. A finite-rank operator, by virtue of being finite-rank, behaves essentially the same as a linear operator between two finite-dimensional spaces. However, they are somewhat too trivial. Compact operators are a compromise. They can have infinite rank, but still allow some finite-dimensional intuitions to apply.

An operator $K : \mathcal{H} \rightarrow \mathcal{H}$ is **compact** iff it is the operator-norm limit of a sequence of finite-rank operators.

Symmetric matrices are diagonalizable. Similarly, if an operator is compact and self-adjoint, then it is diagonalizable: There exists an orthonormal basis of eigenvectors $\{|e_j\rangle\}_{j \in J}$ and real eigenvalues $\{\lambda_j\}_{j \in J}$ with

$$K = \sum_{j \in J} \lambda_j |e_j\rangle\langle e_j|, \quad \lambda_j \in \mathbb{R}, \quad \lambda_j \rightarrow 0.$$

Nonzero eigenvalues have finite multiplicity, and the eigenvalues can only accumulate at zero.

We will only study operators of a very specific form that arises naturally from kernel regression.

Let $\{|v_n\rangle\}_{n \geq 0}$ be an orthonormal set in $\mathcal{H}$ and let $a_n \geq 0$ with $\sum_n a_n < \infty$. Define

$$K := \sum_{n=0}^{\infty} a_n |v_n\rangle\langle v_n|,$$

Then:

- K is self-adjoint, positive, has finite trace, and compact.
- K has eigenpairs $\{(a_n, |v_n\rangle) : n \geq 0\}$, counting multiplicities of repeated a_n .
- $\|K\| = \sup_n a_n$, $\text{Tr}[K] = \sum_n a_n$.

G.4 FUNCTION SPACES

Generally, the vector space of functions is infinite-dimensional. Hilbert space theory shows that certain geometric intuitions honed from low-dimensional Euclidean spaces are still useful even in an infinite-dimensional function space.

As a basic example, consider $L^2(dx)$, the set of square-integrable functions on $\mathbb{R} \rightarrow \mathbb{R}$, with inner product

$$\langle f|g\rangle = \int_{\mathbb{R}} f(x)g(x)dx$$

This generalizes to $L^2(d^D x)$, the space of square-integrable functions on $\mathbb{R}^D \rightarrow \mathbb{R}^D$. It has inner product

$$\langle f|g \rangle = \int_{\mathbb{R}^D} f(x)g(x)d^D x$$

Let μ be a probability measure on $\mathbb{R}^d$. Then $L^2(\mu)$ is the space of functions f such that $\int_{\mathbb{R}^d} |f(x)|^2 \mu(dx)$. That is, the space of functions f , such that if $X \sim \mu$, then $f(X)$ has finite second-moment.

The inner product $\langle f|g \rangle = \mathbb{E}_{X \sim \mu}[f(X)g(X)]$.

In this formalism, the expectation operator is simply taking the inner product by the constant-1 vector:

$$\mathbb{E}_{X \sim \mu}[f(X)] = \mathbb{E}_{X \sim \mu}[1 f(X)] = \langle 1|f \rangle$$

We will mostly study $L^2(\mu)$ in the case where μ is a Gaussian distribution $\mathcal{N}(0, \Sigma)$. Those cases are called **Gaussian spaces**.

We will concern ourselves mainly with integral kernel operators over $L^2(\mu)$. An integral kernel operator is a linear operator of the type $K : L^2(\mu) \rightarrow L^2(\mu)$. It is defined using a kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, such that

$$K|v \rangle = |w \rangle : \quad w(x) = \int_{\mathbb{R}^d} K(x, y)v(y)\mu(dy)$$

Solving for the eigensystem of K is equivalent to diagonalizing the kernel operator, such that

$$K = \sum_{n=0}^{\infty} \lambda_n |w_n \rangle \langle w_n|$$

for some orthonormal basis $|w_n \rangle$.

As an example, the dot-product kernel in $\mathbb{R}^1$ is

$$K(x, y) = \sum_{n=0}^{\infty} \frac{c_n}{n!} (xy)^n$$

In operator form, it is equivalent to

$$K = \sum_{n=0}^{\infty} \frac{c_n (2n-1)!!}{n!} |v_n \rangle \langle v_n|$$

where we define the vectors corresponding to the monomial functions:

$$|v_n \rangle, \quad v_n(x) = \frac{x^n}{\sqrt{(2n-1)!!}}$$

The factor of $\sqrt{(2n-1)!!}$ is necessary to make sure the vectors are unit-length: $\langle v_n|v_n \rangle = 1$.

The problem with the representation

$$K = \sum_{n=0}^{\infty} \frac{c_n (2n-1)!!}{n!} |v_n \rangle \langle v_n|$$

is that, while it looks like a diagonalization, it is not, because the vectors are unit-length, but not orthogonal to each other. Diagonalizing the kernel is impossible in general, but it is possible in certain limiting cases. Our main theoretical work in this paper is just working out these cases.

G.5 1D GAUSSIAN SPACE

Let $\mu(dx) = (2\pi)^{-1/2}e^{-x^2/2}dx$ be the standard Gaussian distribution on $\mathbb{R}^1$. We have the 1D Gaussian space $L^2(\mu)$.

The differentiation operator D_x maps $f(x)$ to $f'(x)$, and by integration-by-parts, has adjoint

$$D_x^* = -D_x + x$$

Therefore, we can define the Ornstein–Uhlenbeck (OU) operator

$$L := -D_x^2 + xD_x = (D_x^*D_x).$$

Then L is self-adjoint and positive semidefinite on $L^2(\mu)$.

The eigenvector equation for L is

$$L|v_n\rangle = \lambda_n|v_n\rangle \implies v_n''(x) - xv_n'(x) + \lambda_nv_n(x) = 0$$

This is just Hermite's differential equation, with solutions being the probabilist's Hermite polynomials He_n . They are orthogonal in $L^2(\mu)$ with

$$\mathbb{E}_\mu[\text{He}_m(X)\text{He}_n(X)] = n!\delta_{mn}.$$

but not normalized. They are normalized by

$$h_n(x) := \frac{\text{He}_n(x)}{\sqrt{n!}}, \quad |h_n\rangle \in L^2(\mu).$$

Thus we see that the normalized Hermite polynomials is the eigensystem for the OU operator in Gaussian space:

$$L|h_n\rangle = n|h_n\rangle, \quad L = \sum_{n=0}^{\infty} n|h_n\rangle\langle h_n|$$

Since the OU operator is the simplest way to make a self-adjoint operator out of the differentiation operator D_x , it is a natural object to consider.

G.6 MULTIDIMENSIONAL GAUSSIAN SPACE

The gaussian space over $\mathbb{R}^1$ can be generalized to multiple dimensions. Let $\mu_d = \mathcal{N}(0, I_d)$ be the standard gaussian distribution on $\mathbb{R}^d$. The corresponding OU operator is

$$L = -\Delta + x \cdot \nabla = \sum_{i=1}^d -\partial_{x_i}^2 + x\partial_{x_i}$$

That is, we can write L as a sum of 1-dimensional OU operators, as $L = \sum_{i=1}^d L_i$. Note that the operator operates on the dimensions $x_1, \dots, x_d$ without interfering with each other:

$$L(f_1(x_1) \cdots f_d(x_d)) = (L_1 f_1)(x_1)f_2(x_2) \cdots f_d(x_d) + \cdots + f_1(x_1) \cdots f_{d-1}(x_{d-1})(L_d f_d)(x_d)$$

Because of this, the eigensystem of L is obtained directly by taking the product of the eigensystem of $L_1, \dots, L_d$.

For a multiindex $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$ set

$$\text{He}_\alpha(x) := \prod_{i=1}^d \text{He}_{\alpha_i}(x_i), \quad h_\alpha(x) := \prod_{i=1}^d \frac{\text{He}_{\alpha_i}(x_i)}{\sqrt{\alpha_i!}}, \quad |\alpha| := \sum_i \alpha_i.$$

Then, L is diagonalized as

$$L = \sum_{\alpha \in \mathbb{N}^d} |\alpha| |h_\alpha\rangle\langle h_\alpha|$$

For each $n = 0, 1, 2, \dots$, the eigenspace

$$\mathcal{E}_n := \text{span}\{|h_\alpha\rangle : |\alpha| = n\}$$

has multiplicity

$$\dim \mathcal{E}_n = \#\{\alpha \in \mathbb{N}^d : |\alpha| = n\} = \binom{n+d-1}{d-1}.$$

This degeneracy is due to symmetry. For any orthogonal transformation $M : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and any smooth $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we have $L(f \circ M) = (Lf) \circ M$. Consequently, if $|v_n\rangle$ is a solution to the eigenvector equation $L|v_n\rangle = \lambda_n|v_n\rangle$, so would its spherical rotations, and any vector sum of them would have the same eigenvalue λ_n .

This is similar to the case of solid harmonics in $\mathbb{R}^d$. In that case, Δ is also spherically symmetric, and its eigenspaces accordingly are degenerate.

G.7 UNITARY TRANSFORMATIONS OF GAUSSIAN SPACES

Consider two Gaussian spaces $L^2(\mathcal{N}(0, I_d)), L^2(\mathcal{N}(0, \Sigma))$. One of them is standardized, the other is not. Since many properties of the Hermite polynomials are derived over the standard Gaussian space, we would like to translate statements in $L^2(\mathcal{N}(0, I_d))$ to statements in $L^2(\mathcal{N}(0, \Sigma))$.

By the geometric view point, as long as a statement is cast in the language of Hilbert space geometry, the statement will continue to hold true after any unitary transformation.

Now, suppose we have a matrix M such that $MM^\top = \Sigma$, then

$$\mathbb{E}_{X \sim \mathcal{N}(0, \Sigma)}[f(X)g(X)] = \mathbb{E}_{Z \sim \mathcal{N}(0, I_d)}[f(LZ)g(LZ)]$$

Thus, we have a unitary transformation $V : L^2(\mathcal{N}(0, \Sigma)) \rightarrow L^2(\mathcal{N}(0, I_d))$, defined by

$$(Vf)(x) = f(Mx), \quad (V^*f)(x) = f(M^{-1}x)$$

Note that the matrix M has an ambiguity: If $MM^\top = \Sigma$, then given any orthogonal matrix O , we also have $(OM)(OM)^\top = \Sigma$. Thus, we can obtain an entire family of unitary transformations, one per orthogonal matrix O . We will exploit this degree of ambiguity later.

Vectors can be transformed. So can operators. An operator A is transformed to $VA V^*$, so that $(VA V^*)(V|v\rangle) = V(A|v\rangle)$.

$$\begin{array}{ccc} L^2(\mathcal{N}(0, \Sigma)) & \xrightarrow{A} & L^2(\mathcal{N}(0, \Sigma)) \\ \downarrow V & & \downarrow V \\ L^2(\mathcal{N}(0, I_d)) & \xrightarrow{VA V^*} & L^2(\mathcal{N}(0, I_d)) \end{array}$$

If an operator is represented as $\sum_n a_n |w_n\rangle\langle v_n|$, then its transformed operator is represented as $\sum_n a_n V|w_n\rangle\langle v_n|V^*$. In particular, diagonalized operators stay diagonalized:

$$A = \sum_n \lambda_n |v_n\rangle\langle v_n| \implies VA V^* = \sum_n \lambda_n V|v_n\rangle\langle v_n|V^*$$

Kernels are transformed in the same way as functions, since

$$\begin{aligned} & \mathbb{E}_{X \sim \mathcal{N}(0, \Sigma)} [\mathbb{E}_{Y \sim \mathcal{N}(0, \Sigma)} [f(X)K(X, Y)g(Y)]] \\ &= \mathbb{E}_{X \sim \mathcal{N}(0, I_d)} [\mathbb{E}_{Y \sim \mathcal{N}(0, I_d)} [f(MX)K(MX, MY)g(MY)]] \end{aligned}$$

G.8 SOME USEFUL PROPERTIES OF HERMITE POLYNOMIALS

Let μ be the standard normal distribution on $\mathbb{R}^1$. For each $n \in \mathbb{N}$, define the normalized monomial vector $|v_n\rangle$ by the function $\frac{x^n}{\sqrt{(2n-1)!!}}$, and define the normalized probabilist's Hermite polynomial vector $|h_n\rangle = \frac{1}{\sqrt{n!}} \text{He}_n(x)$.

They are both sequences of unit vectors in $L^2(\mu)$. However, the sequence $|h_0\rangle, |h_1\rangle, \dots$ is orthonormal, but the sequence $|v_0\rangle, |v_1\rangle, \dots$ is not orthonormal, and in fact, is increasingly ill-conditioned.

By looking up a standard reference table, we have the following basic properties:

1. $|h_0\rangle, |h_2\rangle, |h_4\rangle, \dots$ are obtained by Gram-Schmidt orthonormalization of $|v_0\rangle, |v_2\rangle, |v_4\rangle, \dots$
2. $|h_1\rangle, |h_3\rangle, |h_5\rangle, \dots$ are obtained by Gram-Schmidt orthonormalization of $|v_1\rangle, |v_3\rangle, |v_5\rangle, \dots$

3.

$$|h_n\rangle = \sqrt{n!} \sum_{m=0}^{\lfloor \frac{n}{2} \rfloor} \frac{(-1)^m \sqrt{(2n-4m-1)!!}}{2^m m! (n-2m)!} |v_{n-2m}\rangle$$

4. $|v_n\rangle = \sum_{m=0}^{\infty} M_{nm} |h_m\rangle$ where M is an invertible lower-triangular matrix satisfying

$$M_{n,n-2m} = \frac{n!}{\sqrt{(2n-1)!!}} \frac{1}{2^m m! \sqrt{(n-2m)!}}$$

$$M_{n,n-2m}^{-1} = \sqrt{n!} \frac{(-1)^m \sqrt{(2n-4m-1)!!}}{2^m m! (n-2m)!}$$

for $0 \leq 2m \leq n$.

5.

$$\langle v_n | v_m \rangle = \begin{cases} \frac{(n+m-1)!!}{\sqrt{(2n-1)!!(2m-1)!!}} & n \equiv m \pmod{2} \\ 0 & \text{else} \end{cases}$$

In particular,

$$\begin{aligned} & \vdots \\ \langle v_n | v_{n-4} \rangle &= \frac{\sqrt{(2n-1)(2n-3)(2n-5)(2n-7)}}{(2n-1)(2n-3)}, \\ \langle v_n | v_{n-2} \rangle &= \frac{\sqrt{(2n-1)(2n-3)}}{2n-1}, \\ \langle v_n | v_n \rangle &= 1, \\ \langle v_n | v_{n+2} \rangle &= \frac{2n+1}{\sqrt{(2n+1)(2n+3)}}, \\ \langle v_n | v_{n+4} \rangle &= \frac{(2n+1)(2n+3)}{\sqrt{(2n+1)(2n+3)(2n+5)(2n+7)}} \\ & \vdots \end{aligned}$$

A useful operator is M_x , the multiply-by- x operator, which, using the multiplication formula

$$x \text{He}_n = \text{He}_{n+1} + n \text{He}_{n-1}$$

gives

$$M_x = \sum_n \sqrt{n+1} |h_{n+1}\rangle \langle h_n| + \sqrt{n+1} |h_n\rangle \langle h_{n+1}|$$

It is equivalently expressed as $M_x = a + a^*$, where $a = \sum_n \sqrt{n+1} |h_n\rangle \langle h_{n+1}|$ is the lowering ladder operator. Furthermore, because $\text{He}'_n = n \text{He}_{n-1}$, the lowering ladder operator is the differentiation operator D_x . Therefore, $a^* = D_x^* = M_x - a = M_x - D_x$.

H PROOF OF THEOREM 1

In this appendix, we prove [Theorem 1](#), which states that the eigensystem of a Gaussian kernel $K_\sigma(\mathbf{x}, \mathbf{x}') = e^{-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{x}'\|^2}$ on a multidimensional anisotropic Gaussian measure $\mu = \mathcal{N}(0, \Sigma)$ approaches the Hermite eigenstructure given in [Definition 4](#) as $\sigma \rightarrow \infty$. To show this, we will obtain exact expressions for the kernel eigensystem, then simply show that the eigenvalues and eigenvectors are those predicted by the HEA up to terms that vanish as σ grows.

We will proceed in three stages of successive generality:

1. We solve the problem for $L^2(\mu)$, where $\mu = \mathcal{N}(0, 1)$ is the standard Gaussian. This is the hardest stage.
2. We take an outer product of measures to solve the problem for $L^2(\mu)$, where $\mu = \mathcal{N}(0, \mathbf{I}_d)$.
3. We solve the problem for $L^2(\mu)$, where $\mu = \mathcal{N}(0, \Sigma)$, by taking an opportune unitary transformation from $L^2(\mathcal{N}(0, \gamma))$ to $L^2(\mathcal{N}(0, 1))$.

H.1 PART 1: THE 1D UNIT GAUSSIAN

We solve the problem for $L^2(\mu)$, where $\mu = \mathcal{N}(0, 1)$ is the standard Gaussian. Our approach here is essentially the one used by ([Zhu et al., 1997](#)).

We first quote the Mehler formula ([Mehler, 1866](#)):

$$\frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\rho^2(x^2+y^2)-2\rho xy}{2(1-\rho^2)}\right) = \sum_{n=0}^{\infty} \rho^n h_n(x) h_n(y)$$

for any $\rho \in [0, +1)$, where h_k is the [normalized Hermite polynomial](#).

We multiply by $(1-\rho)$, and substitute with $\rho = e^{-t}$, to make the form more beautiful:

$$\sqrt{\tanh(t/2)} \exp\left(-\frac{e^{-t}(x^2+y^2)-2xy}{4 \sinh t}\right) = \sum_{n=0}^{\infty} (1-e^{-t}) e^{-nt} h_n(x) h_n(y)$$

Define an integral kernel operator $K_t : L^2(\mu) \rightarrow L^2(\mu)$ using the expression on the left:

$$(K_t f)(x) = \int_{\mathbb{R}^1} \sqrt{\tanh(t/2)} \exp\left(-\frac{e^{-t}(x^2+y^2)-2xy}{4 \sinh t}\right) f(y) \mu(dy)$$

The Mehler formula then states that K_t is diagonalized as:

$$K_t = \sum_{n=0}^{\infty} (1-e^{-t}) e^{-nt} |h_n\rangle \langle h_n|.$$

Thus, we have successfully obtained a one-parameter family of integral kernel operators, all diagonalized in the Hermite basis. Each operator has a Gaussian kernel. However, none of them matches the form that we want:

$$K(x, y) = e^{-\frac{1}{2\sigma^2}(x-y)^2}$$

In order to reach such a form, we need to construct $\{T_\tau : \tau \in \mathbb{R}\}$, a one-parameter family of unitary transformations $L^2(\mu)$. Then we will solve for τ, t , such that $T_\tau K_t V_\tau^*$ has the desired kernel form.

For any $\tau \in \mathbb{R}$, define the squeezing operator $T_\tau : L^2(\mu) \rightarrow L^2(\mu)$ by

$$(T_\tau f)(x) = e^{\tau/2} e^{-\frac{e^{2\tau}-1}{4}x^2} f(e^\tau x)$$

This is unitary by direct integration and change of variables. It is a one-parameter family, and it satisfies

$$T_\tau^* = T_{-\tau}, \quad T_{\tau_1} \circ T_{\tau_2} = T_{\tau_1+\tau_2}$$

We can interpret this one-parameter family as a continuous “rotation” in $L^2(\mu)$. Except that, because $L^2(\mu)$ has infinitely many dimensions, the rotation need not return to the starting point. Concretely, consider what happens when the 0th Hermite vector $|h_0\rangle$ is rotated. The family of vectors $T_\tau |h_0\rangle$ would not turn back towards $|h_0\rangle$ again. Instead, it continues to rotate further and further away, with

$$\langle h_0 | T_\tau | h_0 \rangle = \frac{1}{\sqrt{\cosh \tau}} \rightarrow 0$$

as $\tau \rightarrow \infty$. Such behavior is possible, because there are infinitely many dimensions to rotate to.

The one-parameter family T_τ can be written as $T_\tau = e^{\tau A}$, where A is the **infinitesimal generator** of the family:

$$A = \partial_\tau|_{\tau=0} T_\tau = xD_x + \frac{1}{2}(1-x^2) = \text{He}_1 D_x + \frac{1}{2} \text{He}_2$$

Using the [multiplication](#) and [differentiation](#) formulas of the Hermite polynomials, we have

$$\begin{aligned} A \text{He}_n &= -\frac{1}{2}(\text{He}_{n+2} - n(n-1)\text{He}_{n-2}), \\ A^2 \text{He}_n &= \frac{1}{4}(\text{He}_{n+4} - 2(n^2 + n + 1)\text{He}_n + n(n-1)(n-2)(n-3)\text{He}_{n-4}) \end{aligned}$$

Thus,

$$\langle h_n | A | h_n \rangle = 0, \quad \langle h_n | A^2 | h_n \rangle = -\frac{n^2 + n + 1}{2}$$

Therefore, the angle between $|h_n\rangle$ and $T_\tau |h_n\rangle$, at the limit of small τ , is

$$\arccos \langle h_n | e^{\tau A} | h_n \rangle = \sqrt{\frac{n^2 + n + 1}{2}} \tau + O(\tau^2)$$

Equivalently, since

$$A | h_n \rangle = -\frac{1}{2} \sqrt{(n+1)(n+2)} | h_{n+2} \rangle + \frac{1}{2} \sqrt{n(n-1)} | h_{n-2} \rangle$$

we see that the operator rotates $|h_n\rangle$ in the direction of $|h_{n+2}\rangle$ and $|h_{n-2}\rangle$ simultaneously. This explains our previous statement that $e^{\tau A}$ rotates $|h_n\rangle$ further and further away, towards $|h_\infty\rangle$.

Now, for any two function f, g , we have

$$\begin{aligned} \langle f | T_\tau K T_\tau^* | g \rangle &= \iint (T_{-\tau} f)(x) K(x, y) (T_{-\tau} g)(y) \mu(dx) \mu(dy) \\ &= \iint f(u) K_\tau(u, v) g(v) \mu(du) \mu(dv) \end{aligned}$$

where the transformed kernel is

$$K_\tau(x, y) = e^\tau e^{-\frac{e^{2\tau}-1}{4}(x^2+y^2)} K(e^\tau x, e^\tau y)$$

Consider the previously solved case of $K = K_t$. Plugging it in, we find that $T_\tau K T_\tau^*$ is an integral kernel operator with kernel function

$$K_{t,\tau}(x, y) = e^\tau \sqrt{\tanh(t/2)} \exp \left(- \left(\frac{e^{2\tau}-1}{4} + \frac{e^{-t}e^{2\tau}}{4 \sinh t} \right) (x^2 + y^2) + \frac{e^{2\tau}}{2 \sinh t} xy \right).$$

To match the target form $K(x, y) = \exp(-(x-y)^2/(2\sigma^2))$,

$$\frac{e^{2\tau}}{2 \sinh t} = \frac{1}{\sigma^2}, \quad \frac{e^{2\tau}-1}{4} + \frac{e^{-t}e^{2\tau}}{4 \sinh t} = \frac{1}{2\sigma^2}.$$

Eliminate $e^{2\tau}$ and solve for t , we have $\sigma^2 = e^t + e^{-t} - 1$, that is,

$$\sigma^2 = 4 \sinh^2(t/2), \quad t = 2 \operatorname{arsinh}(\sigma/2)$$

Plug it back,

$$e^{2\tau} = \frac{2 \sinh t}{\sigma^2} = \tanh(t/2)^{-1} = \sqrt{1 + \frac{4}{\sigma^2}}$$

So the prefactor simplifies to $e^\tau \sqrt{\tanh(t/2)} = 1$, hence

$$(T_\tau K_t T_\tau^*)(x, y) = \exp \left(-\frac{(x-y)^2}{2\sigma^2} \right), \quad t = 2 \operatorname{arsinh}(\sigma/2), \quad \tau = \frac{1}{4} \ln \left(1 + \frac{4}{\sigma^2} \right)$$

The diagonalization of $T_\tau K_t T_\tau^*$ is

$$T_\tau K_t T_\tau^* = \sum_{n=0}^{\infty} (1 - e^{-t}) e^{-nt} T_\tau |h_n\rangle \langle h_n| T_\tau^*$$

Thus, the eigensystem is

$$\text{eigensystem}(\mu, K) = \{(1 - e^{-t}) e^{-nt}, T_\tau |h_n\rangle : n = 0, 1, 2, \dots\}$$

At the $\sigma \rightarrow \infty$ limit, we have

$$t = 2 \ln \sigma + 2/\sigma^2 + O(\sigma^{-4}), \quad \tau = \sigma^{-2} - 2\sigma^{-4} + O(\sigma^{-6})$$

Now, the Hermite eigensystem corresponding to $e^{-\frac{(x-y)^2}{2\sigma^2}}$ is $\mathcal{HE}(1, (c_n))$ with $c_n = \sigma^{-2n}$, we see that the HEA is proven:

$$\frac{(1 - e^{-t}) e^{-nt}}{c_n} = 1 - (2n + 2)\sigma^{-2} + O(\sigma^{-4}),$$

$$\arccos \langle h_n | T_\tau | h_n \rangle = \sqrt{\frac{n^2 + n + 1}{2}} \sigma^{-2} + O(\sigma^{-4})$$

We also see that, for any fixed n , the rate of convergence is on the order of $n\sigma^{-2} = nc_{\ell+1}/c_\ell$. We will show in the next few sections that this is a generic phenomenon.

In general, if a kernel has coefficients c_ℓ decaying at a rate of ϵ , then the n -th entry in the kernel's eigensystem converges to the corresponding n -th entry in the Hermite eigensystem, at a rate of $O(\epsilon)$. The constant in $O(\epsilon)$ increases with n , so that the convergence is not uniform. The higher-order entries converge slower.

H.2 PART 1 BONUS

We can diagonalize $K(x, y) = e^{xy/\sigma^2}$ in the same way:

$$K(x, y) = (1 - 2/\sigma^2)^{-1/2} K_{t,\tau}(x, y), \quad t = \text{arcosh}(\sigma^2/2), \quad \tau = \frac{1}{4} \ln \left(1 - \frac{4}{\sigma^4}\right)$$

Thus, its eigensystem is

$$\text{eigensystem}(\mu, K) = \{(1 - 2/\sigma^2)^{-1/2} (1 - e^{-t}) e^{-nt}, T_\tau |h_n\rangle : n = 0, 1, 2, \dots\}$$

At the $\sigma \rightarrow \infty$ limit,

$$t = 2 \ln \sigma - \sigma^{-4} + O(\sigma^{-8}), \quad \tau = -\sigma^{-4} + O(\sigma^{-8})$$

yielding

$$\frac{(1 - e^{-t}) e^{-nt}}{c_n} = 1 + (n + \frac{1}{2})\sigma^{-4} + O(\sigma^{-8}),$$

$$\arccos \langle h_n | T_\tau | h_n \rangle = -\sqrt{\frac{n^2 + n + 1}{2}} \sigma^{-4} + O(\sigma^{-8})$$

This case is special, in that the convergence is on the order of $n\sigma^{-4} = n(c_{\ell+1}/c_\ell)^2$, which is faster by one order of magnitude than the generic case.

To understand this, we directly expand the operators. The operator for $\exp(-(x - y)^2/2\sigma^2)$ has expansion

$$K_{\exp(-(x-y)^2/2\sigma^2)} = I + \sigma^{-2}(|h_1\rangle\langle h_1| - M_x^2) + O(\sigma^{-4})$$

where we note that the multiplication operator M_x is self-adjoint, with expression

$$M_x = \sum_n \sqrt{n+1} |h_{n+1}\rangle \langle h_n| + \sqrt{n+1} |h_n\rangle \langle h_{n+1}|$$

Because M_x^2 is not diagonal in the Hermite basis, the operator $K_{\exp(-(x-y)^2/2\sigma^2)}$ is not diagonal at the σ^{-2} order, and perturbation occurs at that order.

In contrast, the operator for $\exp(xy/\sigma^2)$ has expansion

$$K_{\exp(xy/\sigma^2)} = I + \sigma^{-2} |h_1\rangle \langle h_1| + \sigma^{-4} (|h_2\rangle + 2^{-1/2} |h_0\rangle) (\langle h_2| + 2^{-1/2} \langle h_0|) + O(\sigma^{-6})$$

Therefore, it is not diagonal at the σ^{-4} order, and perturbation occurs at that order.

H.3 PART 2

We take its product, to solve the problem for $L^2(\mu)$, where $\mu = \mathcal{N}(0, I_d)$.

The kernel $K(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}$ decomposes into a product of kernels:

$$K(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^d e^{-\frac{(x_i - y_i)^2}{2\sigma^2}}$$

Therefore, its kernel operator decomposes into a tensor product of kernel operators:

$$\begin{aligned} K &= \bigotimes_{i=1}^d K_i \\ &= \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d} \bigotimes_{i=1}^d (1 - e^{-t}) e^{-\alpha_i t} T_{\tau} |h_{\alpha_i}\rangle \langle h_{\alpha_i}| T_{\tau}^* \\ &= \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d} (1 - e^{-t})^d e^{-|\boldsymbol{\alpha}|t} T_{\tau}^{\otimes d} |h_{\boldsymbol{\alpha}}\rangle \langle h_{\boldsymbol{\alpha}}| (T_{\tau}^{\otimes d})^* \end{aligned}$$

So, its eigensystem is

$$\{(1 - e^{-t})^d e^{-|\boldsymbol{\alpha}|t}, T_{\tau}^{\otimes d} |h_{\boldsymbol{\alpha}}\rangle : \boldsymbol{\alpha} \in \mathbb{N}^d\}$$

At the $\sigma \rightarrow \infty$ limit, using the previous result, each eigenvalue converges as

$$\frac{(1 - e^{-t})^d e^{-|\boldsymbol{\alpha}|t}}{c_{|\boldsymbol{\alpha}|}} = 1 - (2|\boldsymbol{\alpha}| + 2d)\sigma^{-2} + O(\sigma^{-4})$$

and each eigenvector converges as

$$\arccos \langle h_{\boldsymbol{\alpha}} | T_{\tau}^{\otimes d} | h_{\boldsymbol{\alpha}} \rangle = \sqrt{\frac{\sum_{i=1}^d (\alpha_i^2 + \alpha_i + 1)}{2}} \sigma^{-2} + O(\sigma^{-4})$$

H.4 PART 3

We solve the problem for $L^2(\mu)$, where $\mu = \mathcal{N}(0, \boldsymbol{\Sigma})$, by taking an opportune unitary transformation $V : L^2(\mathcal{N}(0, \boldsymbol{\Sigma})) \rightarrow L^2(\mathcal{N}(0, I_d))$.

As previously stated, if $MM^{\top} = \boldsymbol{\Sigma}$, then $V : L^2(\mathcal{N}(0, \boldsymbol{\Sigma})) \rightarrow L^2(\mathcal{N}(0, I_d))$ defined by

$$(Vf)(x) = f(Mx), (V^*f)(x) = f(M^{-1}x)$$

is unitary.

Now, define the operator using the kernel function $K(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}$. With the unitary transformation, its kernel becomes

$$(VKV^*)(\mathbf{x}, \mathbf{y}) = K(M\mathbf{x}, M\mathbf{y}) = e^{-\frac{(\mathbf{x}-\mathbf{y})^{\top} M^{\top} M (\mathbf{x}-\mathbf{y})}{2\sigma^2}}$$

Therefore, the kernel decomposes if $M^{\top} M$ is diagonalized. This can be solved by taking the SVD of the covariance matrix as $\boldsymbol{\Sigma} = U\Gamma U^{\top}$, where $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_d)$, then we can set

$$M = U\Gamma^{1/2}$$

so that VKV^* has kernel

$$\prod_{i=1}^d e^{-\frac{(x_i - y_i)^2}{2(\sigma/\sqrt{\gamma_i})^2}}$$

Thus, VKV^* diagonalizes as

$$VKV^* = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d} \prod_{i=1}^d (1 - e^{-t_i}) e^{-\alpha_i t_i} \bigotimes_{i=1}^d T_{\tau_i} |h_{\alpha_i}\rangle \langle h_{\alpha_i}| \bigotimes_{i=1}^d T_{\tau_i}^*$$

where t_i, τ_i are defined by

$$t_i = 2 \operatorname{arsinh}(\sigma/2\sqrt{\gamma_i}), \quad \tau_i = \frac{1}{4} \ln \left(1 + \frac{4\gamma_i}{\sigma^2} \right)$$

Now, we convert this back to K to obtain

$$K = \sum_{\alpha \in \mathbb{N}^d} \prod_{i=1}^d (1 - e^{-t_i}) e^{-\alpha_i t_i} V^* \otimes_{i=1}^d T_{\tau_i} |h_{\alpha_i}\rangle (V^* \otimes_{i=1}^d T_{\tau_i} |h_{\alpha_i}\rangle)^*$$

At the limit of $\sigma \rightarrow \infty$, the eigenvalues converge to

$$\prod_{i=1}^d (1 - e^{-t_i}) e^{-\alpha_i t_i} = \left(\sigma^{-2|\alpha|} \prod_i \gamma_i^{\alpha_i} \right) (1 - (2|\alpha| + 2d)\sigma^{-2} + O(\sigma^{-4}))$$

at a rate of σ^{-2} .

As before, the eigenvectors converge to $V^* |h_{\alpha}\rangle$ at a rate of $\sqrt{\frac{\sum_{i=1}^d (\alpha_i^2 + \alpha_i + 1)}{2}} \sigma^{-2}$. Since

$$(V^* h_{\alpha})(\mathbf{x}) = h_{\alpha}(\Gamma^{-1/2} U^{\top} \mathbf{x}) = h_{\alpha}^{(\Sigma)}(\mathbf{x})$$

the theorem is proven fully.

Similarly to the case in [Appendix H.2](#), the kernel $K(\mathbf{x}, \mathbf{y}) = e^{\mathbf{x}^{\top} \mathbf{y} / \sigma^2}$ converges to the Hermite eigensystem at a rate of $O(\sigma^{-4})$, one order of magnitude faster.

I PROOF OF THEOREM 2 IN ONE DIMENSION

In this appendix, we prove Theorem 2 in the special case of one dimension. The general case is proven in Appendix J. The proof of this case is easier than the general case since there is no multiplicity. The ideas in the proof would be reused in the general case.

Appendix I.1 is a reference sheet quoting several theorems we need for the proofs. The reader can skip it and refer back when needed. Appendix I.2 states Theorem 2 rigorously as Theorem 5, then shows that it is a special case of a more general theorem (Theorem 6). The next section proves the general case. Appendix I.3.1 shows that the eigenvalues of the kernel converge to the desired form with relative error decaying at a rate of $O(\epsilon)$. Appendix I.3.2 leverages this to show that the eigenvectors also converge to the desired form at a rate of $O(\epsilon)$.

I.1 SETUP

We need to quote some big-name theorems for later use.

We will need to cut up the spectrum of an operator into segments, each falling within an interval. The following theorem allows us to construct tight enough bounds on the intervals. It is a special case of the general Courant–Fischer–Weyl min-max principle, strong enough for our purpose. Our special case avoids the part about essential spectrum, which makes the general statement inconvenient to use.

Theorem 3 (Courant–Fischer–Weyl min-max principle). *Let A be a compact, positive semidefinite operator over a Hilbert space $\mathcal{H}$. Let its eigenvalues be enumerated as $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$. They can be finite or infinite. Then,*

$$\lambda_k = \max_{\substack{\mathcal{M} \subset \mathcal{H} \\ \dim \mathcal{M} = k}} \min_{\substack{x \in \mathcal{M} \\ \|x\|=1}} \langle x | A | x \rangle \quad (46)$$

$$= \min_{\substack{\mathcal{M} \subset \mathcal{H} \\ \dim \mathcal{M} = k-1}} \max_{\substack{x \in \mathcal{M}^\perp \\ \|x\|=1}} \langle x | A | x \rangle \quad (47)$$

Proof. The second equation is (Teschl, 2009, Theorem 4.10). The first equation is proven by essentially the same technique as the finite-dimensional case.

Let $v_1, v_2, \dots$ be its eigenvectors. If we set $\mathcal{M}$ to be the span of $\{v_1 | \dots, v_k\}$, then we have

$$\lambda_k = \min_{\substack{x \in \mathcal{M} \\ \|x\|=1}} \langle x, Ax \rangle$$

So it remains to prove the other half.

For any subspace with $\dim \mathcal{M} = k$, it must have a nontrivial intersection with $\{v_k, v_{k+1}, \dots\}$, therefore, there exists some unit vector $x \in \mathcal{M}$, such that it has decomposition $x = \sum_{i \geq k} a_i v_i$. With that, we have

$$\langle x | Ax \rangle = \sum_{i \geq k} |a_i|^2 \lambda_i \leq \sum_{i \geq k} |a_i|^2 \lambda_k = \lambda_k$$

□

The min-max principle has a corollary that we will use, more convenient than quoting the full min-max principle.

Corollary 1 (Cauchy interlacing law). *For any $n \times n$ Hermitian matrix A_n with top left $n-1 \times n-1$ minor A_{n-1} , then*

$$\lambda_{i+1}(A_n) \leq \lambda_i(A_{n-1}) \leq \lambda_i(A_n)$$

for all $1 \leq i < n$. (Tao, 2012, Eq. 1.75)

After bounding the eigenvalues, we will use the second theorem to bound the eigenvectors. However, we need something more, because we need to bound the eigenvector rotations of *segments* of the spectrum, and the segment may be badly separated within, even though it is well-separated from other segments.

Therefore, we need to handle the rotations *reducing subspaces*. A reducing subspace for an operator A is a closed subspace V , such that $A(V) \subset V$. For self-adjoint operators, the reducing subspaces are direct sums of eigenspaces. In particular, the span of an eigenvector is a reducing subspace.

Given a segment of the spectrum, $\Lambda \subset \sigma(A)$, we define V_Λ as the reducing space of Λ . For example, if $\Lambda = \{E\}$, then V_Λ is the closed span of all eigenvectors with eigenvalue E . We also define P_Λ as the orthogonal projector to V_Λ .

The following theorem shows that, if a segment of the operator spectrum is separated by $\Omega(1)$ from the rest of the spectrum, then its corresponding reducing subspaces would only rotate by $O(\epsilon)$ under an operator perturbation by $O(\epsilon)$.

Theorem 4 (Davis–Kahan $\sin \Theta$ theorem (Davis & Kahan, 1970)). *Let A, B be self-adjoint operators such that*

- $\sigma(A)$ is partitioned into Λ_0, Λ_1
- $\sigma(B)$ is partitioned into Γ_0, Γ_1
- The spaces $V_{\Lambda_0}, V_{\Gamma_0}$ are of the same dimension. Similarly for $V_{\Lambda_1}, V_{\Gamma_1}$.
- Λ_0 is contained in an interval $[x, y]$.
- Γ_1 is disjoint from the enlarged interval $(x - \delta, y + \delta)$.

Then there exists a self-adjoint “angle” operator Θ , such that the rotation operator

$\begin{bmatrix} \cos \Theta & \sin \Theta \\ -\sin \Theta & \cos \Theta \end{bmatrix}$ rotates V_{Λ_i} to V_{Γ_i} for $i = 0, 1$. Furthermore, the angle operator satisfies

$$\|\sin \Theta\| \leq \frac{1}{\delta} \|A - B\| \quad (48)$$

for any unitarily invariant norm $\|\cdot\|$.

Intuitively, the operator Θ is just a diagonal matrix of angles, in a suitable orthonormal basis. The operator $\begin{bmatrix} \cos \Theta & \sin \Theta \\ -\sin \Theta & \cos \Theta \end{bmatrix}$ performs simultaneous rotation in many (potentially infinitely many) 2-dimensional planes, such that the two reducing subspaces of A are rotated to two reducing subspaces of B .

I.2 REDUCTION TO A GENERAL CASE

We clean up the form of Theorem 2 by performing a few WLOGs and reductions into Theorem 5, so that we can deduce the theorem as a corollary of a more general Theorem 6. Alternatively, the theorem can be proven by proving the more general, multidimensional case. This is done in Appendix J. However, it may be worthwhile to study the following special case before reading the more general case, since most of the proof ideas are present.

We begin with a convenient definition.

Definition 5 (fast-decay). A real-valued sequence $c_0, c_1, \dots$ is **fast-decaying** iff there exists a number $\epsilon \in [0, 1)$, such that $c_{n+1} \leq \epsilon c_n$ for all $n \in \mathbb{N}$.

As in the proof of Theorem 1 in Appendix H, we can perform a unitary transform of $L^2(\mathcal{N}(0, \gamma))$ to $L^2(\mathcal{N}(0, 1))$.

Next, we define $a_n = \frac{(2n-1)!!}{n!} c_n$, so that

$$K = \sum_n a_n |v_n\rangle \langle v_n|$$

By Stirling’s approximation,

$$\frac{(2n-1)!!}{n!} = \frac{1}{2^n} \binom{2n}{n} = \frac{2^n}{\sqrt{\pi n}} (1 + O(1/n))$$

Therefore, if c_n is fast-decaying with parameter ϵ , then a_n is fast-decaying with parameters $2\epsilon, \delta_{low, n}/\sqrt{n}$. So, we can study the kernel $\sum_n a_n |v_n\rangle \langle v_n|$, with the fast-decaying condition directly imposed on a_n . This makes the notation cleaner.

Because $\langle v_n | h_n \rangle = \sqrt{\frac{n!}{(2n-1)!!}}$ (see [Appendix G.8](#)), convergence to c_n is equivalent to convergence to $a_n |\langle v_n | h_n \rangle|^2$.

With that, we can restate the theorem we wish to prove, more rigorously:

Theorem 5 (The HEA holds for a fast-decaying kernel on 1D Gaussian measure). *Let $\mu = \mathcal{N}(0, 1)$ be the standard Gaussian measure, and let*

$$K = \sum_{n=0}^{\infty} a_n |v_n\rangle\langle v_n|$$

be a dot-product kernel with fast-decaying coefficients a_n with parameters ϵ , then for any $n \in \mathbb{N}$, there exists an eigensystem of K , written as $(\lambda_0(K), |v_0(K)\rangle), (\lambda_1(K), |v_1(K)\rangle), \dots$, such that

$$\frac{\lambda_n(K)}{a_n |\langle v_n | h_n \rangle|^2} = 1 + O(\epsilon)$$

$$|\angle(|v_n(K)\rangle, |h_n\rangle)| = O(\epsilon)$$

as $\epsilon \rightarrow 0$. Furthermore, the scaling factor in $O(\epsilon)$ depends only on the Gram matrix of the vectors $|v_0\rangle, \dots, |v_n\rangle$.

The above statement may seem oddly convoluted with “for any $n \in \mathbb{N}$, there exists...”, but this is necessary, because the rate of convergence is not uniform over the sequence. In general, higher-order eigenvectors converge slower than lower-order eigenvectors, which means the constant in the $O(\epsilon)$ terms is larger for larger n , and no uniform-convergence rate exists.

The scaling factor in $O(\epsilon)$ measures the speed of convergence for the n -th eigenpair. It is slower if the Gram matrix of the vectors $|v_0\rangle, \dots, |v_n\rangle$ is ill-conditioned, because in this case, the vectors are almost linearly dependent.

Indeed, the Hankel moment matrices of most commonly used probability distributions, including the uniform distribution, the gaussian distribution, and the exponential distribution, are exponentially ill-conditioned ([Chen & Lawrence, 1999](#)). Therefore, we should expect the constant in $O(\epsilon)$ to grow exponentially with n .

Also, take note of the phrasing “there exists an eigensystem of K denoted as $(\lambda_k(K), |v_k(K)\rangle)_k$ ”, because we allow K to suffer multiplicity. In these cases, the corresponding eigenspace would have more than 1 dimension, and therefore there is freedom in choosing any orthonormal basis of the eigenspace as “the eigenvectors”. The theorem states that, despite the multiplicity, there exists a good choice of eigenvectors, such that they make a small angle with the canonical eigenvectors $|\hat{v}_n\rangle$. This will become especially relevant in the proof of the multidimensional generalization in [Appendix J](#).

Because the orthonormal basis $|h_0\rangle, |h_1\rangle, \dots$ is obtained by Gram–Schmidt orthonormalization on $|v_0\rangle, |v_1\rangle, \dots$, we can generalize it, this time with all epsilons and deltas in place for maximal rigor:

Theorem 6. *Let:*

1. $|v_0\rangle, |v_1\rangle, \dots$ be a sequence of linearly independent unit vectors in a Hilbert space, such that their closed span is the whole space;
2. $|\hat{v}_0\rangle, |\hat{v}_1\rangle, \dots$ be the sequence obtained by performing Gram–Schmidt process on the sequence;

Then, there exists a sequence of constants $C_0, C_1, \dots, > 0$ and $\epsilon_0, \epsilon_1, \dots > 0$, such that for all $n \in \mathbb{N}$, $\epsilon \in [0, \epsilon_n]$, and all $a_0, a_1, \dots$ fast-decaying sequences with parameters ϵ , there exists an eigensystem of the kernel $K := \sum_{n=0}^{\infty} a_n |v_n\rangle\langle v_n|$, denoted as $(\lambda_k(K), |v_k(K)\rangle)_k$, such that

$$\lambda_n(K) \in a_n |\langle \hat{v}_n | v_n \rangle|^2 (1 \pm C_n \epsilon)$$

and

$$|\angle(|v_n(K)\rangle, |\hat{v}_n\rangle)| \leq C_n \epsilon$$

Furthermore, the scaling factor C_n and the bound ϵ_n depend on only on the Gram matrix of the vectors $|v_0\rangle, \dots, |v_n\rangle$.

Intuitively restated, the theorem says that as $\epsilon \rightarrow 0$, the eigensystem of the fast-decaying kernel $\sum_{n=0}^{\infty} a_n |v_n\rangle\langle v_n|$ rotates towards the canonical eigensystem at a rate of ϵ .

The proof has two parts. The first part uses the [min-max principle](#) and lowest-order operator perturbation (commonly used in quantum mechanics), to segment the spectrum of K into small intervals of the form

$$\lambda_n(K) \in a_n |\langle \hat{v}_n | v_n \rangle|^2 (1 + O(\epsilon))$$

In particular, since a_n are fast-decaying, it shows that the eigenvalues are exponentially separated.

The second part applies [Davis–Kahan](#) twice, using this exponential separation of eigenvalues, to show that $|\sin \angle(|v_n(K)\rangle, |\tilde{K}\rangle)| = O(\epsilon)$ and $|\sin \angle(|v_n(\tilde{K})\rangle, |\hat{v}_n\rangle)| = O(\epsilon)$, for a cleverly-chosen operator $\tilde{K}$.

Before we launch into the proof, we should look at a simple case that explains why this should be true.

Consider the case of two dimensions. We only have $|v_0\rangle, |v_1\rangle$. In this case, the kernel $K = a_0 |v_0\rangle\langle v_0| + a_1 |v_1\rangle\langle v_1|$. Diagonalizing the kernel is equivalent to finding the major and minor axes of the contour ellipse defined by $\{|x\rangle : \langle x | K | x \rangle = 1\}$. This ellipse is the unique ellipse tangent to the 4 lines defined by $a_0 |\langle v_0 | x \rangle|^2 = 1$, $a_1 |\langle v_1 | x \rangle|^2 = 1$.

Suppose we fix a_0 , and let $a_1 \rightarrow 0$. Then, the lines of $a_0 |\langle v_0 | x \rangle|^2 = 1$ remain constant, but the lines of $a_1 |\langle v_1 | x \rangle|^2 = 1$ diverge to infinity. The ellipse degenerates to two parallel lines. Its minor semiaxis rotates to become perpendicular to the two parallel lines, i.e. parallel to $|v_0\rangle$. Therefore, the eigenpair converges to $(a_0 |\langle \hat{v}_0 | v_0 \rangle|^2, |\hat{v}_0\rangle)$.

Suppose we fix a_1 and let $a_0 \rightarrow \infty$. Then, the lines of $a_1 |\langle v_1 | x \rangle|^2 = 1$ remain constant, but the lines of $a_0 |\langle v_0 | x \rangle|^2 = 1$ converge to the origin. The ellipse degenerates to two line segments. Its major semiaxis rotates to become the same as that line segment, i.e. parallel to $a_0 |\langle v_0 | x \rangle|^2 = 1$, i.e. perpendicular to $|v_0\rangle$. Therefore, the eigenpair converges to $(a_1 |\langle \hat{v}_1 | v_1 \rangle|^2, |\hat{v}_1\rangle)$.

Intuitively, we see that for a given n , the effect of all $a_{n+1}, a_{n+2}, \dots$ terms in the kernel is a small perturbation on the n -th eigenspace, and negligible because the parallel planes diverge to infinity. The effect of all $a_{n-1}, a_{n-2}, \dots, a_0$ is a large but *fixed* perturbation, forcing the n -th eigenspace to be perpendicular to all of $|v_{n-1}\rangle, \dots, |v_0\rangle$, but once that is done, their effects are also negligible because the parallel planes converge to the origin.

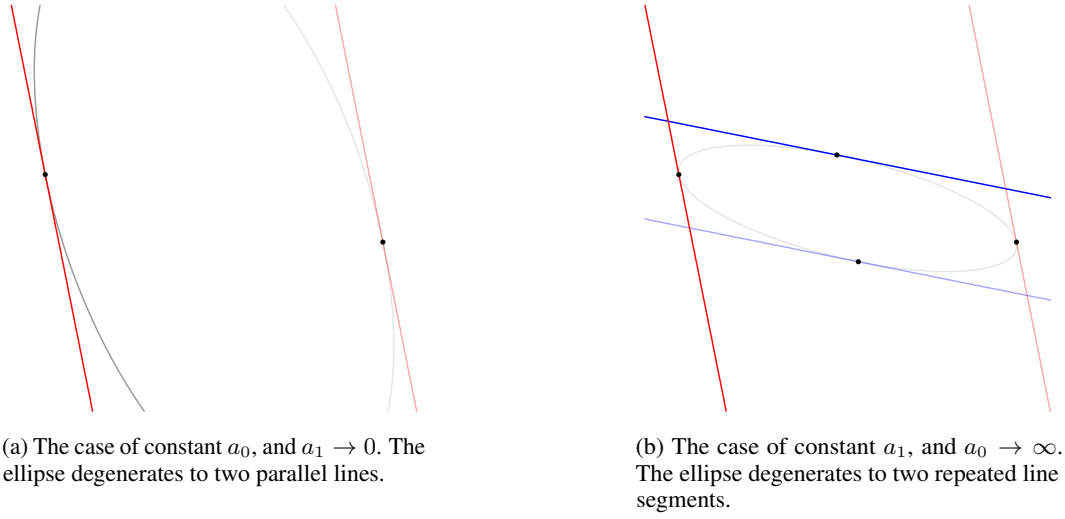


Figure 21: Diagonalizing the kernel in 2 dimensions at the $a_1/a_0 \rightarrow 0$ limit.

I.3 PROOF OF THEOREM 6

I.3.1 PART 1

To show: $\lambda_n(K) = a_n |\langle \hat{v}_n | v_n \rangle|^2 (1 + O(\epsilon))$.

Proof. If $a_n = 0$, then all $\lambda_n, \lambda_{n+1}, \dots = 0$, and the kernel K becomes finite-ranked, with range $\text{Span}(|v_{0:n-1}\rangle)$. So the theorem becomes trivial. Otherwise, we assume $a_n > 0$, which means that all $a_0, \dots, a_n > 0$.

We apply the [min-max principle](#) to obtain an upper bound of the form $\lambda_n(K) \leq a_n |\langle \hat{v}_n | v_n \rangle|^2 (1 + O(\epsilon))$ and a lower bound of the form $\lambda_n(K) \geq a_n |\langle \hat{v}_n | v_n \rangle|^2 (1 + O(\epsilon))$, thus completing the estimate.

For the upper bound, we use $V = \text{Span}(|v_{0:n-1}\rangle)$, then

$$\begin{aligned}
 \lambda_n &\leq \sup_{\substack{x \in \text{Span}(|v_{0:n-1}\rangle)^\perp \\ \|x\|=1}} \langle x | K | x \rangle \\
 &= \sup_{\substack{x \in \text{Span}(|v_{0:n-1}\rangle)^\perp \\ \|x\|=1}} \sum_{k=0}^{\infty} a_k |\langle x | v_k \rangle|^2 \\
 &= \sup_{\substack{x \in \text{Span}(|\hat{v}_{n:\infty}\rangle) \\ \|x\|=1}} \sum_{k=n}^{\infty} a_k |\langle x | v_k \rangle|^2 \\
 &\leq \sup_{\substack{x \in \text{Span}(|\hat{v}_{n:\infty}\rangle) \\ \|x\|=1}} a_n |\langle x | v_n \rangle|^2 + \sup_{\substack{x \in \text{Span}(|\hat{v}_{n:\infty}\rangle) \\ \|x\|=1}} \sum_{k=n+1}^{\infty} a_k |\langle x | v_k \rangle|^2 \\
 &= a_n |\langle \hat{v}_n | v_n \rangle|^2 + \lambda_{\max} \left(\sum_{k=n+1}^{\infty} a_k |v_k\rangle \langle v_k| \right) \\
 &\leq a_n |\langle \hat{v}_n | v_n \rangle|^2 + \text{Tr} \left(\sum_{k=n+1}^{\infty} a_k |v_k\rangle \langle v_k| \right) \\
 &= a_n |\langle \hat{v}_n | v_n \rangle|^2 + \sum_{k=n+1}^{\infty} a_k \\
 &= a_n |\langle \hat{v}_n | v_n \rangle|^2 + a_n O(\epsilon) \\
 &= a_n |\langle \hat{v}_n | v_n \rangle|^2 (1 + O(\epsilon))
 \end{aligned}$$

where the step $\lambda_{\max} \leq \text{Tr}$ is because all $a_k \geq 0$.

Though unnecessary, we can concretely write down the upper bound as

$$\lambda_n \leq a_n |\langle \hat{v}_n | v_n \rangle|^2 (1 + C\epsilon)$$

where

$$C = \frac{1}{|\langle \hat{v}_n | v_n \rangle|^2} \sum_{k=n+1}^{\infty} \frac{a_k}{a_n \epsilon} \leq \frac{1}{|\langle \hat{v}_n | v_n \rangle|^2} \frac{1}{1 - \epsilon}$$

For the lower bound, we use $V = \text{Span}(|v_{0:n}\rangle)$, then

$$\begin{aligned} \lambda_n &\geq \inf_{\substack{x \in \text{Span}(|v_{0:n}\rangle) \\ \|x\|=1}} \langle x | K | x \rangle \\ &= \inf_{\substack{x \in \text{Span}(|v_{0:n}\rangle) \\ \|x\|=1}} \sum_{k=0}^{\infty} a_k |\langle x | v_k \rangle|^2 \\ &\geq \inf_{\substack{x \in \text{Span}(|v_{0:n}\rangle) \\ \|x\|=1}} \sum_{k=0}^n a_k |\langle x | v_k \rangle|^2 \end{aligned}$$

The quantity is a standard problem in quadratic programming, with exact solution $\lambda_{\min}(G^{1/2}AG^{1/2})$, where $A = \text{diag}(a_0, \dots, a_n)$ and $G = (\langle v_i | v_j \rangle)_{i,j=0}^n$ is the Gram matrix.

To see this, write $x = \sum_{j \in 0:n} c_j v_j$, then

$$\|x\|^2 = c^T G c, \quad \sum_{k=0}^n a_k |\langle x | v_k \rangle|^2 = (Gc)^T A (Gc) = c^T G A G c.$$

So the constrained minimum of $\frac{c^T G A G c}{c^T G c}$ equals the smallest eigenvalue of $G^{1/2}AG^{1/2}$ by the Rayleigh–Ritz principle.

Let $M = G^{1/2}AG^{1/2}$. We invert the matrix to use standard operator perturbation theory:

$$M^{-1} = \frac{1}{a_n} u_n u_n^T + \sum_{k=0}^{n-1} \frac{1}{a_k} u_k u_k^T$$

where $u_k = G^{-1/2}e_k$ is the k -th column of $G^{-1/2}$. The perturbation $\sum_{k=0}^{n-1} \frac{1}{a_k} u_k u_k^T$ is order $O(\epsilon)$ compared to the unperturbed part $\frac{1}{a_n} u_n u_n^T$.

The unperturbed operator has maximal eigenvalue $\frac{\|u_n\|^2}{a_n}$ with eigenvector $\hat{u}_n := u_n / \|u_n\|$. The perturbed operator has maximal eigenvalue:

$$\frac{\|u_n\|^2}{a_n} + \hat{u}_n^T \left(\sum_{k=0}^{n-1} \frac{1}{a_k} u_k u_k^T \right) \hat{u}_n + O(\epsilon^2)$$

Inverting the eigenvalue,

$$\lambda_{\min}(G^{1/2}AG^{1/2}) = \frac{a_n}{\|u_n\|^2} \left(1 - \frac{a_n/a_{n-1}}{\|u_n\|^2} |u_{n-1}^T \hat{u}_n|^2 + O(\epsilon^2) \right) = \frac{a_n}{\|u_n\|^2} (1 + O(\epsilon))$$

Now, $\|u_n\|^2 = e_n^T G^{-1} e_n$, and G^{-1} is the Gram matrix of the dual basis of $|v_{0:n}\rangle$ in $\text{Span}(|v_{0:n}\rangle)$. In particular, because $|\hat{v}_n\rangle$ is perpendicular to all of $|v_0\rangle, \dots, |v_{n-1}\rangle$, the n -th dual vector is just $\frac{|\hat{v}_n\rangle}{\langle v_n | \hat{v}_n \rangle}$.

Therefore $\|u_n\|^2 = \left\| \frac{|\hat{v}_n\rangle}{\langle v_n | \hat{v}_n \rangle} \right\|^2 = \frac{1}{|\langle v_n | \hat{v}_n \rangle|^2}$, and we obtain the desired lower bound

$$\lambda_n \geq a_n |\langle \hat{v}_n | v_n \rangle|^2 (1 + O(\epsilon))$$

□

Some comments on the constants in $O(\epsilon)$.

In the above proof, we constructed an upper bound $a_n |\langle \hat{v}_n | v_n \rangle|^2 (1 + O(\epsilon))$ and a lower bound $a_n |\langle \hat{v}_n | v_n \rangle|^2 (1 - O(\epsilon))$.

The constant in the upper bound is

$$\frac{1}{|\langle \hat{v}_n | v_n \rangle|^2} \frac{1}{1 - \epsilon}$$

which depends on $|\langle \hat{v}_n | v_n \rangle|^2 = \sin^2 \theta$, where θ is the angle between $|v_n\rangle$ and $\text{Span}(|v_{0:n-1}\rangle)$. We see that the bound is pushed wider when either the coefficients become less strictly exponentially-decaying, or the vector $|v_n\rangle$ leans into $\text{Span}(|v_{0:n-1}\rangle)$, and thus becomes less orthonormal.

The constant in the lower bound is

$$\begin{aligned} \frac{1}{\|u_n\|^2} |u_{n-1}^T \hat{u}_n|^2 &= \frac{1}{\|u_n\|^4} |u_{n-1}^T u_n|^2 \\ &= \langle v_n | \hat{v}_n \rangle^4 |u_{n-1}^T u_n|^2 \\ &= |\langle v_n | \hat{v}_n \rangle|^4 (G_{n-1,n}^{-1})^2 \end{aligned}$$

Similarly to the previous case, $G_{n-1,n}^{-1}$ gets larger as the vectors $|v_0\rangle, \dots, |v_n\rangle$ get less orthonormal, which worsens eigenvalue convergence.

I.3.2 PART 2

Proof. If $a_n = 0$, then we can trivially select the n -th eigenvector to be $|\hat{v}_n\rangle$. Otherwise, we have $a_0, \dots, a_n > 0$.

By the eigenvalue bound (that is, Part 1 of the theorem), the spectral gap around $\lambda_n(K)$ is

$$\min_{j \neq n} |\lambda_n(K) - \lambda_j(K)| = a_n |\langle \hat{v}_n | v_n \rangle|^2 (1 + O(\epsilon))$$

Define the truncated operator $\tilde{K} = \sum_{k=0}^n a_k |v_k\rangle \langle v_k|$. By [Davis–Kahan](#),

$$\begin{aligned} |\sin \angle(v_n(K), v_n(\tilde{K}))| &\leq \frac{2}{a_n |\langle \hat{v}_n | v_n \rangle|^2 (1 + O(\epsilon))} \|K - \tilde{K}\|_{op} \\ &\leq \frac{2}{a_n |\langle \hat{v}_n | v_n \rangle|^2 (1 + O(\epsilon))} \text{Tr} [K - \tilde{K}] \\ &= \frac{2}{a_n |\langle \hat{v}_n | v_n \rangle|^2 (1 + O(\epsilon))} \sum_{k=n+1}^{\infty} a_k \\ &= O(\epsilon) \end{aligned}$$

Thus, we need only bound $|\sin \angle(v_n(\tilde{K}), \hat{v}_n)|$.

Because $\tilde{K}$ lives inside $\text{Span}(|v_{0:n}\rangle)$, we thenceforth restrict the Hilbert space to just $\text{Span}(|v_{0:n}\rangle)$.

Define the twice-truncated operator $\bar{K} = \sum_{k=0}^{n-1} a_k |v_k\rangle \langle v_k|$. The eigenvalue bound applies to its first n eigenvalues, and its $(n+1)$ -th eigenstate is the ground state, with eigenvalue 0 and eigenvector $|\hat{v}_n\rangle$.

Thus, the spectral gap around its ground state eigenvalue is

$$\min_{j \neq n} |\lambda_n(\bar{K}) - \lambda_j(\bar{K})| = \lambda_{n-1}(\bar{K}) = a_{n-1} |\langle \hat{v}_{n-1} | v_{n-1} \rangle|^2 (1 + O(\epsilon))$$

By [Davis–Kahan](#) again,

$$\begin{aligned} |\sin \angle(v_n(\tilde{K}), \hat{v}_n)| &= |\sin \angle(v_n(\tilde{K}), v_n(\bar{K}))| \\ &\leq \frac{2}{a_{n-1} |\langle \hat{v}_{n-1} | v_{n-1} \rangle|^2 (1 + O(\epsilon))} \|\tilde{K} - \bar{K}\|_{op} \\ &= \frac{2}{a_{n-1} |\langle \hat{v}_{n-1} | v_{n-1} \rangle|^2 (1 + O(\epsilon))} a_n \\ &= O(\epsilon) \end{aligned}$$

2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2750
2751
2752
2753

□

There are two occurrences of $O(\epsilon)$ in the proof. Both can be bounded explicitly, to show that they only depend on the Gram matrix of $|v_0\rangle, \dots, |v_n\rangle$, as in Part 1.

The first $O(\epsilon)$ has explicit upper bound constant

$$\frac{2}{a_n \epsilon |\langle \hat{v}_n | v_n \rangle|^2 (1 + O(\epsilon))} \sum_{k=n+1}^{\infty} a_k \leq \frac{2}{|\langle \hat{v}_n | v_n \rangle|^2} \frac{1}{1 - \epsilon} \frac{1}{1 + O(\epsilon)}$$

where the remaining $O(\epsilon)$ in $\frac{1}{1+O(\epsilon)}$ came from Part 1, which as we showed in Part 1, only depends on the Gram matrix of $|v_0\rangle, \dots, |v_n\rangle$.

The second $O(\epsilon)$ has explicit upper bound constant

$$\frac{2}{|\langle \hat{v}_{n-1} | v_{n-1} \rangle|^2 (1 + O(\epsilon))}$$

where the remaining $O(\epsilon)$ only depends on the Gram matrix of $|v_0\rangle, \dots, |v_{n-1}\rangle$.

J PROOF OF THEOREM 2 IN THE GENERAL CASE

In this appendix, we prove Theorem 2 completely. The general idea is to modify the proof given in Appendix I to account for multiplicity.

Appendix J.1 presents the overall plan of the proof. Appendix J.2 sets up all the machinery needed to handle multiplicity in eigenvalues and eigenvectors, which is a new occurrence in multiple dimensions. Appendix J.3 states the theorem in full rigor as Theorem 7. The next two subsections prove two lemmas that apply to generic operators, not just an integral kernel operator: Appendix J.4 shows that the eigenvalues of a generic fast-decaying kernel splits into segments that are exponentially separated, and Appendix J.5 sharpens this separation into proving that the eigenvalues in the N -th segment are only slightly perturbed by all the other segments. Appendix J.6 specializes to the case of a dot-product kernel, showing convergence of the eigenvalues, and then leverages that convergence into the convergence of eigenspaces.

J.1 PLAN OF THE PROOF

Let $K(\mathbf{x}, \mathbf{y}) = \sum_{n=0}^{\infty} \frac{c_n}{n!} (\mathbf{x}^\top \mathbf{y})^n$ be a dot-product kernel with fast-decaying coefficients c_n .

As in Appendix H.4, to study a spherically symmetric dot-product kernel over a nonstandard Gaussian distribution $\mathcal{N}(0, \Sigma)$, we construct a whitening unitary transform $V : L^2(\mathcal{N}(0, \Sigma)) \rightarrow L^2(\mathcal{N}(0, I_d))$, thus converting the problem to solving for the eigenstructure of a spherically asymmetric kernel over the standard Gaussian distribution.

Let the SVD of Σ be $U\Gamma U^\top$, where $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_d)$ with $\gamma_1, \dots, \gamma_d \geq 0$. Define V by $(Vf)(\mathbf{x}) = f(M\mathbf{x})$. This then converts the operator K to VKV^* . The operator VKV^* is a kernel operator with kernel function satisfying

$$(VKV^*)(\mathbf{x}, \mathbf{y}) = \sum_{n=0}^{\infty} \frac{c_n}{n!} (\gamma_1 x_1 y_1 + \dots + \gamma_d x_d y_d)^n$$

We will prove that the eigensystem of VKV^* converges to

$$\left\{ \left(c_{|\alpha|} \prod_{i=1}^d \gamma_i^{\alpha_i}, |h_\alpha\rangle \right) \text{ for all } \alpha \in \mathbb{N}_0^d \right\}$$

as $\epsilon \rightarrow 0$. Then, by reversing the V transform, we find that the eigensystem of K converges to

$$\left\{ \left(c_{|\alpha|} \prod_{i=1}^d \gamma_i^{\alpha_i}, |h_\alpha^{(\Sigma)}\rangle \right) \text{ for all } \alpha \in \mathbb{N}_0^d \right\}$$

as desired.

We eliminate a pesky special case: one or more of $\gamma_1, \dots, \gamma_d$ may be equal to zero. In this case, VKV^* may not be positive definite, but merely positive semidefinite, which is annoying. For example, what if $\gamma_2 = \gamma_4 = 0$? Then the operator VKV^* splits to two halves. It is the zero operator on $\text{Span}(e_2, e_4)$, and it is positive definite on $\text{Span}(e_1, e_3, e_5, \dots, e_d)$. Then we can separately prove the eigensystem convergence on the two halves, and take their tensor product. The case of zero operator is obviously trivial, since its eigensystem is just

$$\{(0, |h_{(\alpha_2, \alpha_4)}\rangle)\} \text{ for all } (\alpha_2, \alpha_4) \in \mathbb{N}_0^2\}$$

Thus, WLOG, we need only consider the case where $\gamma_1, \dots, \gamma_d > 0$.

We note that, at least in one case, the theorem has been proven: If $c_n = \sigma^{-2n}$ for some σ , then it is just a minor variant of Theorem 1, which has been proven in Appendix H.2.

Thus, if we prove that the difference between the eigensystem of VKV^* and the eigensystem of $ce^{\mathbf{x}^\top \mathbf{y} / \sigma^2}$ vanishes at $\epsilon \rightarrow 0$, for some well-chosen values of σ, c , we are done.

This cannot be done directly, once again due to nonuniform convergence: The higher-order parts of the eigensystem is wilder, and harder to control. To bypass this difficulty, we divide and conquer.

We prove that the eigensystem of VKV^* is “segmented” into exponentially separated intervals, such that each segment is ϵ -insensitive to perturbations in all other segments. This allows us to show

that, for any fixed $n \in \mathbb{N}_0$, the n -th segment of eigensystem (VKV^*) – corresponding to the term $c_n(\mathbf{x}^\top \Gamma \mathbf{y})^n$ – converges to the n -th segment of eigensystem $(K_{ce^e(\mathbf{x}^\top \Gamma \mathbf{y})})$, where $c = c_n e^{-n}$. Since the n -th segment of eigensystem $(K_{ce^e(\mathbf{x}^\top \Gamma \mathbf{y})})$ converges to

$$\left\{ \left(\frac{c_n}{\epsilon^n} \prod_{i=1}^d \gamma_i^{\alpha_i}, |h_\alpha\rangle \right) \text{ for all } \alpha \in \mathbb{N}_0^d, |\alpha| = n \right\}$$

the theorem is proven.

J.2 MACHINERY FOR MULTIPLICITY

The main difficulty, compared to the one-dimensional case, is that we must directly handle the multiplicity of eigensystems. By this, we mean that in $\mathbb{R}^d$, the Hermite basis is no longer of form $\{|h_n\rangle\}_{n \in \mathbb{N}}$, but rather, $\{|h_\alpha\rangle\}_{\alpha \in \mathbb{N}^d}$. This creates degeneracy, that is, if $\sum_i \alpha_i = \sum_i \alpha'_i$, then $|h_\alpha\rangle, |h_{\alpha'}\rangle$ belong to the same eigenspace.

Concretely, define the Ornstein–Uhlenbeck operator $\nabla^2 - x \cdot \nabla$ on $L^2(\mu_d)$. In the $d = 1$ case, its eigenvalues have no multiplicity, and its eigenvectors are precisely the normalized Hermite polynomials $\{|h_n\rangle\}_{n \in \mathbb{N}}$. In the $d > 1$ case, its eigenvalues suffer multiplicity. Its n -th eigenspace is $\{|h_\alpha\rangle : |\alpha| = n\}_{\alpha \in \mathbb{N}^d}$, with $\binom{n+d-1}{d-1}$ dimensions.

This multiplicity is inescapable, because the Ornstein–Uhlenbeck operator is spherically symmetric, and spherical symmetry inevitably leads to multiplicity. In our case, the dot-product kernel $\sum_n \frac{c_n}{n!} (\mathbf{x}^\top \Gamma \mathbf{y})^n$ may have some entries of Γ equal, which leads to (partial) spherical symmetry, and thus multiplicity.

As a more famous example, the spherical harmonics in $L^2(\mathbb{R}^d)$ are the eigenvectors of the Laplacian operator ∇^2 . Due to spherical symmetry of the operator, its eigenvalues have multiplicity, thus it splits $L^2(\mathbb{R}^d)$ into eigenspaces. The n -th eigenspace is spanned by $\frac{(2n+d-2)(n+d-3)!}{n!(d-2)!}$ degree- n homogeneous polynomials.

Let's consider the prototypical case that we want to study: the convergence of dot-product kernels to the Hermite eigensystem. In $\mathbb{R}^d$, there are $\binom{n+d-1}{d-1}$ degree- n monomials, and $\binom{n+d-1}{d-1}$ degree- n Hermite polynomials. To obtain the Hermite polynomials, we cannot simply apply the Gram–Schmidt process on the monomials individually. Instead, we need to apply the Gram–Schmidt process simultaneously on each segment of $\binom{n+d-1}{d-1}$ monomials, to obtain the dimension- $\binom{n+d-1}{d-1}$ subspace spanned by the degree- n Hermite polynomials.

To count the multiplicity, we use a function $m : \mathbb{N} \rightarrow \mathbb{N} \cup \{\infty\}$. It should be interpreted as saying that the n -th eigenspace has dimension $m(n)$. For example, the multiplicity counting function for the Ornstein–Uhlenbeck operator over $L^2(\mu_d)$ is $m(n) = \binom{n+d-1}{d-1}$.

Note that m does not need to be strictly positive. That is, we allow $m(k) = 0$ for some k . We even allow $\sum_k m(k)$ to be finite, in the case that the Hilbert space under consideration is finite-dimensional, although we require $\sum_k m(k) > 0$, for otherwise it would be completely trivial.

For convenience, we will from now on assume that either $m(k) > 0$ for all k , since we do not need more generality. The reader who needs this generality can read the next few sections and mentally generalize the constructions.

The most important condition on the multiplicity counting function is:

Definition 6 (polynomially bounded multiplicity). A multiplicity counting function is **polynomially bounded** iff there exists some $A, d > 0$ such that $m(n) < An^d$ for all $n \in \mathbb{N}$.

This is satisfied by the Hermite basis in any dimension, since its $m(n) = O(n^{d-1})$.

Given a multiplicity counting function, we define a system of vectors that it counts:

Definition 7 (vector systems with multiplicity). Given a multiplicity counting function m , a vector system with multiplicity m is an indexed set of vectors of form $\{v_{k,l} : k \in \mathbb{N}, l \in [1 : m(k)]\}$.

As in the last few sections, we will only consider vector systems that consists of linearly independent unit vectors, and such that the closure of their span is the entire Hilbert space.

Similarly, we define a system of coefficient it counts:

Definition 8 (coefficient systems with multiplicity). Given a multiplicity counting function m , a coefficient system with multiplicity m is an indexed set of real numbers of form $\{v_{k,l} : k \in \mathbb{N}, l \in [1 : m(k)]\}$.

We next generalize the Gram–Schmidt process to handle multiplicity.

Definition 9 (Gram–Schmidt process on vector systems). Given a multiplicity counting function m , and a linearly independent vector system $\{v_{k,l} : k \in \mathbb{N}, l \in [1 : m(k)]\}$, we define the Gram–Schmidt process on the vector system by the following algorithm:

1. Construct an arbitrary orthonormal basis of $\text{Span}(|v_{0,1}\rangle, \dots, |v_{0,m(0)}\rangle)$. Call them $|\hat{v}_{0,1}\rangle, \dots, |\hat{v}_{0,m(0)}\rangle$.
2. Select the next smallest k such that $m(k) > 0$, and project each of $|v_{k,1}\rangle, \dots, |v_{k,m(k)}\rangle$ to $\text{Span}(|\hat{v}_{0,1}\rangle, \dots, |\hat{v}_{0,m(0)}\rangle)^\perp$, to obtain $|v'_{k,1}\rangle, \dots, |v'_{k,m(k)}\rangle$ then construct an arbitrary orthonormal basis of their span. Call them $|\hat{v}_{k,1}\rangle, \dots, |\hat{v}_{k,m(k)}\rangle$.
3. Continue this way inductively.

Note that the Gram–Schmidt process is not uniquely defined, due to the steps where arbitrary orthonormal bases are chosen. However, it constructs a sequence of subspaces $\{\text{Span}(|\hat{v}_{k,1}\rangle, \dots, |\hat{v}_{k,m(k)}\rangle)\}_{k \in \mathbb{N}}$, which are uniquely defined.

Also note that even the traditional Gram–Schmidt process, without multiplicity, still is not uniquely defined, because each $\hat{v}_k$ could have been $-\hat{v}_k$ instead. That is, there is a $\{-1, +1\}$ ambiguity per step. Now, note that $\{-1, +1\}$ is just $O(1)$, the orthogonal group on $\mathbb{R}^1$, and we see that it is a general phenomenon: In general, the Gram–Schmidt process with multiplicity m creates an $O(m(k))$ amount of ambiguity at step k .

Each vector system defines a positive semi-definite kernel

$$K = \sum_{k=0}^{\infty} \sum_{l=1}^{m(k)} a_{k,l} |v_{k,l}\rangle \langle v_{k,l}|$$

for any indexed set of non-negative scalars $\{v_{k,l} : k \in \mathbb{N}, l \in [1 : m(k)]\}$, provided that all $a_{k,l} \geq 0$, and $\sum_{k,l} a_{k,l} < \infty$.

To handle the multiplicity of eigensystem(K), we need to make four changes.

1. Generalize the definition of “fast-decaying coefficients” to fast-decaying *segments* of coefficients.
2. Prove that *segments* of adjacent eigenvalues are exponentially separated.
3. Prove the convergence of reducing subspaces (i.e. direct sums of eigenspaces), rather than eigenvectors.
4. Prove the convergence of whole segments of eigensystem(K), rather than individual entries like $(\lambda_n(K), |v_n(K)\rangle)$.

A fast-decaying sequence of coefficient segments with the specified multiplicity m is obtained by slightly loosening a fast-decaying sequence of coefficients, so that instead of individual coefficients, it is segments of coefficients that are now decaying exponentially. This will force segments of the eigenvalues to be well-separated as well, and thus their corresponding reducing subspaces.

Definition 10 (fast-decay with multiplicity). A coefficient system $c_{k,l}$ is fast-decaying iff there exists a sequence of numbers $\delta_{low,n} \in (0, 1]$, and a number $\epsilon \in [0, 1)$, and a sequence of numbers $\bar{c}_n$, such that

$$c_{n,l} \in [\bar{c}_n \delta_{low,n}, \bar{c}_n], \quad \bar{c}_{n+1} \leq \epsilon \bar{c}_n, \quad \forall n \in \mathbb{N}$$

We say that such a system is a fast-decaying coefficient system with parameters $(\epsilon, \delta_{low,n}, \bar{c}_n)$.

For example, a fast-decaying sequence is a fast-decaying system where the multiplicity counting function is $m(k) = 1$, and all $\delta_{low,n} = 1$. We will show that the coefficient system of a dot-product kernel $\sum_n \frac{c_n}{n!} (\mathbf{x}^\top \Gamma \mathbf{y})^n$ is fast-decaying in [Appendix J.6](#).

Given a polynomially bounded multiplicity m , a vector system for m , and a fast-decaying coefficient system, define

$$K = \sum_k \sum_l a_{k,l} |v_{k,l}\rangle \langle v_{k,l}|$$

It is positive semidefinite and has finite trace. Therefore, its spectrum is discrete, and except for zero, each of its eigenvalue is positive and has only finite multiplicity. Therefore, its eigensystem is well-defined.

Definition 11 (eigensegment). Given K , a positive semidefinite operator with finite trace, enumerate its eigenvalues as $\lambda_{0,1} \geq \dots \geq \lambda_{0,m(0)} \geq \lambda_{1,1} \geq \dots \geq 0$, counting eigenvalue multiplicity. Construct a corresponding orthonormal eigenbasis $|v_{k,l}\rangle$.

If all $\lambda_{n,1}, \dots, \lambda_{n,m(n)}$ are distinct, then the n -th eigensegment of eigensystem(K) is the set $\{(\lambda_{n,l}, \text{Span}(|v_{n,l}\rangle)) : l \in [1 : m(n)]\}$. Otherwise, if $\lambda_{n,1} = \lambda_{n,2}$, and all others are distinct, then the n -th eigensegment of eigensystem(K) is the set

$$\{(\lambda_{n,1}, \text{Span}(|v_{n,1}\rangle, |v_{n,2}\rangle)), (\lambda_{n,2}, \text{Span}(|v_{n,1}\rangle, |v_{n,2}\rangle))\} \cup \{(\lambda_{n,l}, \text{Span}(|v_{n,l}\rangle)) : l \in [3 : m(n)]\}$$

In general, the eigensegment is obtained by merging degenerate eigenspaces.

Despite the valiant effort in removing ambiguity, the above definition still has some residual ambiguity: If unluckily, $\lambda_{k,m(k)} = \lambda_{k+1,1}$, then the reducing subspaces of the k -th and $(k+1)$ -th eigensegments are not uniquely defined, since we can rotate the eigenvector pair $|v_{k+1,1}\rangle, |v_{k,m(k)}\rangle$ arbitrarily.

Fortunately, we will demonstrate that the residual ambiguity disappears in the $\epsilon \rightarrow 0$ limit for fast-decaying kernels.

Finally, we need to define what it means for two eigensegments to be close together:

Definition 12 (eigensegment bulk closeness). Given two eigensegments of equal length, let $\lambda_{n,1}, \dots, \lambda_{n,m(n)}$ be the eigenvalues from the first segment, and $\lambda'_{n,1}, \dots, \lambda'_{n,m(n)}$ from the second. We say that they are ϵ -close in bulk if

$$|\lambda_{n,l} - \lambda'_{n,l}| < \epsilon \min(\lambda_{n,l}, \lambda'_{n,l}) \quad \forall l \in [1 : m(n)]$$

In bulk? Indeed, due to the annoyances of degeneracy, we would first show that the eigenvalues converge as $O(\epsilon)$. After that is done, we can set ϵ small enough so that it will force the eigenspaces to match up as well, and thus does bulk closeness resolve into detailed closeness.

Definition 13 (eigensegment detailed closeness). Given two eigensegments of equal length, we say that the two eigensegments are ϵ -close in detail if they are ϵ -close in bulk, and for each eigenspace V' in the second eigensegment, there exists one or more eigenspaces $V_1, \dots, V_s$ in the first, such that

there exists a unitary angle operator Θ , such that $\begin{bmatrix} \cos \Theta & \sin \Theta \\ -\sin \Theta & \cos \Theta \end{bmatrix}$ rotates $V_1 \oplus \dots \oplus V_s$ to V' , and $\|\sin \Theta\|_{op} < \epsilon$.

See [Appendix I.1](#) and ([Davis & Kahan, 1970](#)) for details on the meaning of the angle operator.

Intuitively, the operator $\begin{bmatrix} \cos \Theta & \sin \Theta \\ -\sin \Theta & \cos \Theta \end{bmatrix}$ is the generalization of multidimensional rotation to a Hilbert space. It performs simultaneous rotation in many (potentially infinitely many) 2-dimensional planes. To say that $\|\sin \Theta\|_{op} < \epsilon$ means that in every single one of these planes, the angle of rotation is $< \arcsin(\epsilon)$.

The definition is by design *asymmetric*, because we will show that if one source eigensegment (think of the dot-product kernel's eigensystem) is always $O(\epsilon)$ -close to a target eigensegment (think of the Hermite eigensystem) in the coarse sense, then as $\epsilon \rightarrow 0$, the source eigensegment will be forced to be $O(\epsilon)$ -close to the target eigensegment in the detailed sense. We would rather not hit a moving target with a static gun, but hit a static target with a moving gun.

We point out that even “convergence in detail” does *not* imply convergence of every and each eigenvector, because a degeneracy in the target eigensegment may stubbornly remain broken in the source segment. For example, it is possible that $\lambda'_{n,1} = \lambda'_{n,2}$ exactly in the target eigensegment, but $\lambda_{n,1} \neq \lambda_{n,2}$ in the source eigensegment. In this case, the eigenvectors corresponding to $\lambda_{n,1}, \lambda_{n,2}$ may be rotated by 45° compared to the chosen eigenvectors for $\lambda'_{n,1}, \lambda'_{n,2}$.

Concretely for our case of the dot-product kernels, this means that the kernel may have no degenerate eigenvectors, but the Hermite eigensystem *has* degenerate eigenvectors. In such a case, the best we can possibly do is proving that each eigenspace of the Hermite eigensystem corresponds to a *direct sum* of the kernel's eigensystem that is a small angle's rotation away.

Concretely, suppose that we have $\gamma_1 = \gamma_2$, then the Hermite eigensystem is degenerate, but we may find that the kernel's eigensystem stubbornly remains both nondegenerate *and* rotated by 45° askew of the Hermite basis. For example, it might stubbornly insist on containing two non-degenerate eigenvectors close to $\sqrt{\frac{1}{2}}(|h_{(0,1)}\rangle + |h_{(1,0)}\rangle)$ and $\sqrt{\frac{1}{2}}(|h_{(0,1)}\rangle - |h_{(1,0)}\rangle)$. This is fine, since their direct sum *does* converge to $\text{Span}(|h_{(0,1)}\rangle, |h_{(1,0)}\rangle)$, and that is the best that can be proven. One cannot expect more to be proven given such degeneracy in the target eigensystem.

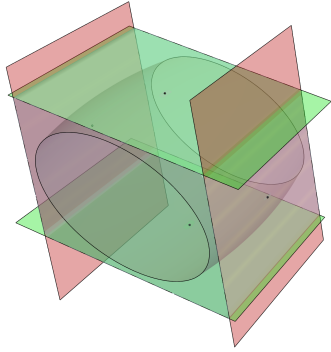
In the case of [Theorem 2](#), before proving it, we saw why it should be true using a 2-dimensional picture with ellipses ([Figure 21](#)). Here, we can also see why it should be true using a 3-dimensional picture with ellipsoids.

Consider the case of three dimensions. We only have $|v_{0,1}\rangle, |v_{1,1}\rangle, |v_{1,2}\rangle$. In this case, the kernel $K = a_{0,1}|v_{0,1}\rangle\langle v_{0,1}| + a_{1,1}|v_{1,1}\rangle\langle v_{1,1}| + a_{1,2}|v_{1,2}\rangle\langle v_{1,2}|$. Diagonalizing the kernel is equivalent to finding the principal axes of the contour ellipsoid defined by $\{|x\rangle : \langle x|K|x\rangle = 1\}$. This ellipsoid is the unique ellipse tangent to the 6 planes defined by $a_{0,1}|v_{0,1}\rangle\langle v_{0,1}| = 1, a_{1,1}|v_{1,1}\rangle\langle v_{1,1}| = 1, a_{1,2}|v_{1,2}\rangle\langle v_{1,2}| = 1$.

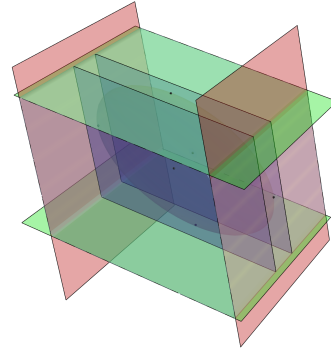
Suppose we fix $a_{1,1}, a_{1,2}$, and let $a_{0,1} \rightarrow 0$. Then, the planes of $a_{1,1}|v_{1,1}\rangle\langle v_{1,1}| = 1, a_{1,2}|v_{1,2}\rangle\langle v_{1,2}| = 1$ remain constant, but the planes of $a_{0,1}|v_{0,1}\rangle\langle v_{0,1}| = 1$ diverge to infinity. The ellipsoid degenerates to a parallelogram prism defined by the 4 planes. Two of its principal axes rotate to become perpendicular to the 2 pairs of parallel planes, and essentially ignore $a_{0,1}$.

Suppose we fix $a_{1,1}, a_{1,2}$, and let $a_{0,1} \rightarrow \infty$. Then, the planes of $a_{1,1}|v_{1,1}\rangle\langle v_{1,1}| = 1, a_{1,2}|v_{1,2}\rangle\langle v_{1,2}| = 1$ remain constant, but the planes of $a_{0,1}|v_{0,1}\rangle\langle v_{0,1}| = 1$ converge to the origin. The ellipsoid degenerates to two flat parallelograms. Two of its principal axes rotate to fall within the parallelogram, perpendicular to its 2 edges.

Intuitively, we see that for a given n , the effect of all $a_{n+1,1}, \dots, a_{n+1,m(n+1)}, a_{n+2,1}, \dots$ terms in the kernel is a small perturbation on the n -th eigenspace, and negligible because the parallel planes diverge to infinity. The effect of all $a_{n-1,m(n-1)-1}, \dots, a_{0,1}$ is a large but *fixed* perturbation, forcing the n -th eigenspace to be perpendicular to all of $|v_{n-1,m(n-1)}\rangle, \dots, |v_{0,0}\rangle$, but once that is done, their effects are also negligible because the parallel planes converge to the origin.



(a) The case of constant $a_{1,1}, a_{1,2}$, and $a_{0,1} \rightarrow 0$. The ellipsoid degenerates to a parallelogram prism.



(b) The case of constant $a_{1,1}, a_{1,2}$, and $a_{0,1} \rightarrow \infty$. The ellipsoid degenerates to two repeated parallelograms.

Figure 22: Diagonalizing the kernel in 3 dimensions at two limits.

Figure 23 shows the eigenstructure of K . On the coarse level, it is divided into exponentially separated segments. The N -th segment contains $m(N)$ eigenvalues clustered within an interval with order of magnitude $\Theta(\bar{a}_N)$. As $\epsilon \rightarrow 0$, the segments become ever cleaner separated, and also converging closer and closer to the N -th segment of a target eigenstructure. The target eigensegment may contain multiple eigenvalues with varying multiplicity. If a target eigenvalue λ has multiplicity 3, then there will be precisely 3 eigenvalues (counting multiplicity) falling within $\lambda(1 \pm O(\epsilon))$.

Notably, these 3 eigenvalues may be degenerate, or not degenerate. In either case, the direct sum of their eigenspaces will have the same dimension as the eigenspace corresponding to λ , and it will make an angle of size $O(\epsilon)$.

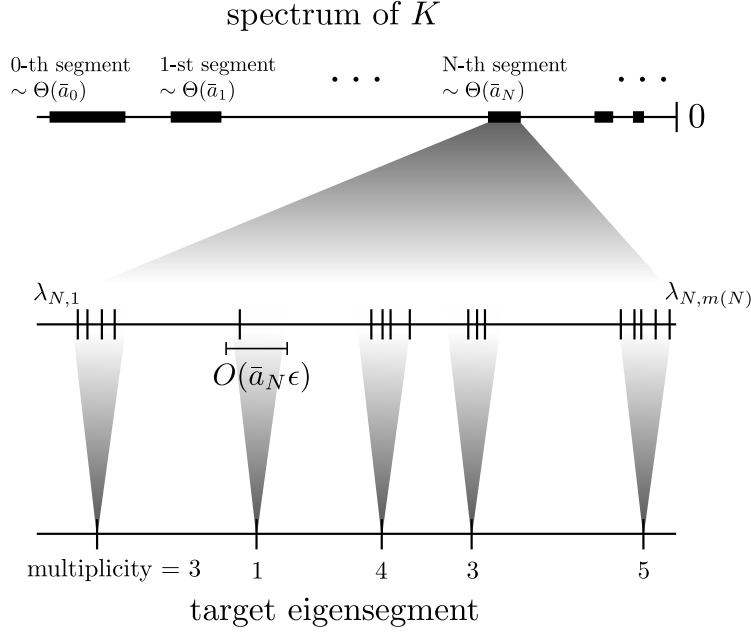


Figure 23: Structure of the kernel spectrum, zooming into one particular N -th segment. They converge to a segment of the target spectrum according to multiplicity of the target spectrum.

J.3 STATEMENT OF THE THEOREM

The statement of the theorem is really unwieldy, so we write it out in a separate section.

A dot-product kernel

$$K(\mathbf{x}, \mathbf{y}) := \sum_{n=0}^{\infty} \frac{c_n}{n!} (\mathbf{x}^\top \mathbf{y})^n$$

is defined by its level-coefficients $c_0, c_1, \dots \geq 0$. Given $\epsilon > 0$, it is ϵ -fast-decaying if $c_{n+1} \leq \epsilon c_n$ for all $n \in \mathbb{N}$.

The multiplicity counting function for it is $m(n) = \binom{n+d-1}{d-1}$. It is polynomially bounded.

We enumerate the eigensystem of K as

$$\begin{aligned} &(\lambda_{0,1}, V_{0,1}), \\ &(\lambda_{1,1}, V_{1,1}), \dots, (\lambda_{1,d}, V_{1,d}), \\ &(\lambda_{2,1}, V_{2,1}), \dots, (\lambda_{2,\binom{d+1}{2}}, V_{2,\binom{d+1}{2}}), \\ &\vdots \\ &(\lambda_{n,1}, V_{n,1}), \dots, (\lambda_{n,\binom{n+d-1}{d-1}}, V_{n,\binom{n+d-1}{d-1}}), \\ &\vdots \end{aligned}$$

where the eigenvalues are arranged in decreasing order:

$$\lambda_{0,1} \geq \lambda_{1,1} \geq \dots$$

In case of multiplicity, the same eigenvalue is written repeatedly. For example, if λ has multiplicity 2, then it is written out twice.

Each $V_{k,l}$ is the eigenspace of $\lambda_{k,l}$. In case of eigenvalue multiplicity, repeat the eigenspace. For example, if $\lambda_{1,1} = \lambda_{1,2}$ is an eigenvalue of multiplicity 2, then it corresponds to a 2-dimensional eigenspace V . In this case, define $V_{1,1} = V_{1,2} = V$.

We also need to perform a *similar* “merging” on the Hermite eigensystem:

$$\left\{ \left(c_n \prod_{i=1}^d \gamma_i^{\alpha_i}, \text{Span} \left(|h_{\alpha}\rangle^{(\Sigma)} \right) \right) : n \in \mathbb{N}, |\alpha| = n \right\}$$

This merging is, unfortunately, *not* exactly the same, because we can’t just say that if $c_n \prod_{i=1}^d \gamma_i^{\alpha_i} = c_{n'} \prod_{i=1}^d \gamma_i^{\alpha'_i}$, then merge their eigenspaces, because we must *not* merge if $n \neq n'$, even if the eigenvalues happen to be the same in this case. This is because, as $\epsilon \rightarrow 0$, eventually c_n and $c_{n'}$ will differ so greatly, that this *accidental* degeneracy is broken. We must only merge eigenspaces that are *non-accidentally* degenerate.

So why didn’t we do this for K ? Because K is much harder to control for! We know everything there could be known about the Hermite eigenstructure, but K is a big unknown that we must laborously control. One thing that is a big unknown about K is that we don’t know which eigenvalues are accidentally the same, and which eigenvalues are non-accidentally the same. So we treat them without special considerations. We know which are accidental and which are not for the Hermite eigenstructure, so we treat them differently.

So, we perform a non-accidental merge of the Hermite eigensystem. For each $n \in \mathbb{N}$, we say that the n -th segment of the Hermite eigensystem is

$$\left\{ \left(c_n \prod_{i=1}^d \gamma_i^{\alpha_i}, \text{Span} \left(|h_{\alpha}\rangle^{(\Sigma)} \right) \right) : |\alpha| = n \right\}$$

and *within* each segment, we merge the degenerate eigenspaces. For example, if it happens that $\prod_{i=1}^d \gamma_i^{\alpha_i} = \prod_{i=1}^d \gamma_i^{\alpha'_i}$ for $|\alpha| = |\alpha'| = n$, but no other α'' with $|\alpha''| = n$, then replace both $\text{Span} \left(|h_{\alpha}\rangle^{(\Sigma)} \right)$ and $\text{Span} \left(|h_{\alpha'}\rangle^{(\Sigma)} \right)$ by their direct sum $\text{Span} \left(|h_{\alpha}\rangle^{(\Sigma)}, |h_{\alpha'}\rangle^{(\Sigma)} \right)$.

With all this set up, we can now state [Theorem 2](#), this time with full rigor.

Theorem 7 (The HEA holds for a fast-decaying dot-product kernel on Gaussian measure). *Let*

1. Σ be a covariance matrix;
2. $\Sigma = U\Gamma U^\top$ be its SVD, with $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_d)$;
3. $\mu = \mathcal{N}(0, \Sigma)$;

For any $N \in \mathbb{N}$, there exists constants C_N, D_N, ϵ_N , such that if $\epsilon \in [0, \epsilon_N]$, and K is an ϵ -fast-decaying dot-product kernel, the following happens.

For any element (λ, V) within the merged N -th segment of the Hermite eigenstructure, there exists exactly $\dim V$ eigenvalues (counting multiplicity) of K that are in the interval $\lambda(1 \pm C_N \epsilon)$. Let their corresponding eigenspaces be $V_1, \dots, V_{\dim V}$, then there exists a unitary operator

$\begin{bmatrix} \cos \Theta & \sin \Theta \\ -\sin \Theta & \cos \Theta \end{bmatrix}$ that rotates $V_1 \oplus \dots \oplus V_{\dim V}$ to V , such that $\|\sin \Theta\|_{op} < D_N \epsilon$.

J.4 PART 1: EXPONENTIAL SEPARATION OF EIGENSYSTEM SEGMENTS

For each segment N , we use the [min-max principle](#) to construct upper and lower bounds on its eigenvalues.

Lemma 1 (exponential separation of eigensystem segments). *Given*

1. *a polynomially bounded multiplicity counting function m ,*
2. *a fast-decaying coefficient system $c_{k,l}$ with parameters $(\epsilon, \delta_{low,n}, \bar{a}_n)$,*
3. *a linearly independent vector system $|v_{k,l}\rangle$,*
4. *operator*

$$K := \sum_{k=0}^{\infty} \sum_{l=1}^{m(l)} a_{k,l} |v_{k,l}\rangle \langle v_{k,l}|$$

then for any n , there exists constants $\epsilon_n, C_n, D_n > 0$, such that

$$\lambda_{n,1}, \dots, \lambda_{n,m(n)} \in [C_n \bar{a}_n, D_n \bar{a}_n]$$

for all $\epsilon \in [0, \epsilon_n]$.

The constants ϵ_n, C_n, D_n depend only on $\delta_{low,0}, \dots, \delta_{low,n}$ and the Gram matrix of the vectors $|v_{0,1}\rangle, \dots, |v_{n,m(n)}\rangle$.

We bound the entire eigensegment $\lambda_{N,1}(K) \geq \dots \geq \lambda_{N,m(N)}(K)$ by the [min-max principle](#). This is analogous to what we did in [Appendix I.3.1](#), but simplified, because we do not need to produce sharp bounds. That comes later.

For the lower bound, we use $V = \text{Span}(|v_{0,1}\rangle, \dots, |v_{N,m(N)}\rangle)$

$$\begin{aligned} \lambda_{N,m(N)}(K) &\geq \min_{\substack{x \in \text{Span}(|v_{0,1}\rangle, \dots, |v_{N,m(N)}\rangle) \\ \|x\|=1}} \langle x | K | x \rangle \\ &= \min_{\substack{x \in \text{Span}(|v_{0,1}\rangle, \dots, |v_{N,m(N)}\rangle) \\ \|x\|=1}} \sum_{k=0}^{\infty} \sum_{l=1}^{m(k)} a_{k,l} |\langle x | v_{k,l} \rangle|^2 \\ &\geq \min_{\substack{x \in \text{Span}(|\hat{v}_{0,1}\rangle, \dots, |\hat{v}_{N,m(N)}\rangle) \\ \|x\|=1}} \sum_{k=0}^N \sum_{l=1}^{m(k)} a_{k,l} |\langle x | v_{k,l} \rangle|^2 \\ &\geq \frac{\bar{a}_N}{\delta_{min,N}} \min_{\substack{x \in \text{Span}(|\hat{v}_{0,1}\rangle, \dots, |\hat{v}_{N,m(N)}\rangle) \\ \|x\|=1}} \sum_{k=0}^N \sum_{l=1}^{m(k)} |\langle x | v_{k,l} \rangle|^2 \\ &= \frac{\bar{a}_N}{\delta_{min,N}} \lambda_{\min}(G) \\ &= \Omega(\bar{a}_N) \end{aligned}$$

where G is the Gram matrix of the vectors $|v_{0,1}\rangle, \dots, |v_{N,m(N)}\rangle$. By assumption, these vectors are linearly independent, so the Gram matrix is positive definite.

For the upper bound, we use $V = \text{Span}(|v_{0,1}\rangle, \dots, |v_{N-1,m(N-1)}\rangle)$.

$$\begin{aligned}
\lambda_{N,1}(K) &\leq \sup_{\substack{x \in \text{Span}(|v_{0,1}\rangle, \dots, |v_{N-1,m(N-1)}\rangle)^\perp \\ \|x\|=1}} \langle x | K | x \rangle \\
&= \sup_{\substack{x \in \text{Span}(|\hat{v}_{N,1}\rangle, \dots) \\ \|x\|=1}} \langle x | K | x \rangle \\
&\leq \sup_{\substack{x \in \text{Span}(|\hat{v}_{N,1}\rangle, \dots) \\ \|x\|=1}} \sum_{l=1}^{m(N)} a_{N,l} \langle x | v_{N,l} \rangle \langle v_{N,l} | x \rangle + \lambda_{\max} \left(\sum_{k=N+1}^{\infty} \sum_{l=1}^{m(k)} a_{k,l} |v_{k,l}\rangle \langle v_{k,l}| \right) \\
&\leq \sup_{\substack{x \in \text{Span}(|\hat{v}_{N,1}\rangle, \dots, |\hat{v}_{N,m(N)}\rangle) \\ \|x\|=1}} \sum_{l=1}^{m(N)} a_{N,l} \langle x | v_{N,l} \rangle \langle v_{N,l} | x \rangle + \text{Tr} \left[\sum_{k=N+1}^{\infty} \sum_{l=1}^{m(k)} a_{k,l} |v_{k,l}\rangle \langle v_{k,l}| \right] \\
&= \lambda_{\max} \left(\left[\sum_{l=1}^{m(N)} a_{N,l} \langle \hat{v}_{N,i} | v_{N,l} \rangle \langle v_{N,l} | \hat{v}_{N,j} \rangle \right]_{i,j=1}^{m(N)} \right) + \underbrace{\sum_{k=N+1}^{\infty} \sum_{l=1}^{m(k)} a_{k,l}}_{\text{use polynomial multiplicity}} \\
&\leq \text{Tr} \left[\left[\sum_{l=1}^{m(N)} a_{N,l} \langle \hat{v}_{N,i} | v_{N,l} \rangle \langle v_{N,l} | \hat{v}_{N,j} \rangle \right]_{i,j=1}^{m(N)} \right] + O(\bar{a}_N \epsilon) \\
&= \sum_{l=1}^{m(N)} \sum_{i=1}^{m(N)} a_{N,l} \langle \hat{v}_{N,i} | v_{N,l} \rangle \langle v_{N,l} | \hat{v}_{N,i} \rangle + O(\bar{a}_N \epsilon) \\
&\leq \bar{a}_N \sum_{l=1}^{m(N)} \sum_{i=1}^{m(N)} \langle \hat{v}_{N,i} | v_{N,l} \rangle \langle v_{N,l} | \hat{v}_{N,i} \rangle + O(\bar{a}_N \epsilon) \\
&= O(\bar{a}_N)
\end{aligned}$$

J.5 PART 2: CONVERGENCE IN BULK

Now that the spectrum is divided into exponentially separated segments, we can perform a surgical extraction of each N -th segment of the spectrum, to cut off both the “head” part $< N$, and the “tail” part $> N$.

Lemma 2 (bulk insensitivity of eigensystem segments). *Under the same assumptions as Appendix J.4, for any n , there exists constants $\epsilon_n > 0$, such that the N -th eigensegment is $O(\epsilon)$ -close in bulk to the spectrum of the matrix*

$$\left[\sum_{l=1}^{m(N)} a_{N,l} \langle \hat{v}_{N,i} | v_{N,l} \rangle \langle v_{N,l} | \hat{v}_{N,j} \rangle \right]_{i,j=1}^{m(N)}$$

for all $\epsilon \in [0, \epsilon_n]$.

The constant ϵ_n and the constant in $O(\epsilon)$ depend only on $\delta_{\text{low},0}$, $\delta_{\text{low},n}$ and the Gram matrix of the vectors $|v_{0,1}\rangle, \dots, |v_{n,m(n)}\rangle$.

Intuitively, the lemma states that the eigensystem segments separate very cleanly. First, relative to the N -th eigensegment, all the higher-order segments are $O(\epsilon)$ -small, and thus ignorable. Second, relative to the N -th segment, the *only* effect of the lower-order segments is to force the N -th eigensegment into a safe subspace very close to the orthogonal subspace $\text{Span}(|\hat{v}_{N,1}\rangle, \dots, |\hat{v}_{N,m(N)}\rangle)$, within which the lower-order terms $a_{0,1}|v_{0,1}\rangle\langle v_{0,1}|, \dots, a_{N-1,m(N-1)}|v_{N-1,m(N-1)}\rangle\langle v_{N-1,m(N-1)}|$ all vanish.

Stated in another way, the lemma states that the N -th segment of the spectrum of K is $O(\epsilon)$ -close in bulk to $\tilde{K}$. To obtain $\tilde{K}$, we first remove its tail $\sum_{n=N+1}^i \sum_{l=1}^{m(n)} a_{n,l} |v_{n,l}\rangle \langle v_{n,l}|$, then project to the

space orthogonal to $|v_{0,1}\rangle, \dots, |v_{N-1,m(N-1)}\rangle$ to remove its head, to obtain

$$\tilde{K} = \sum_{i,j=1}^{m(N)} \sum_{l=1}^{m(N)} a_{N,l} |\hat{v}_{N,i}\rangle \langle \hat{v}_{N,i} | v_{N,l} \rangle \langle v_{N,l} | \hat{v}_{N,j}\rangle \langle \hat{v}_{N,j}|$$

The key of the proof is to ensure that each cut perturbs the eigenvalues by $O(\bar{a}_N \epsilon)$, so that we would extract something that is $O(\epsilon)$ -close in bulk to the original eigensegment.

J.5.1 CUTTING OFF THE TAIL

We need:

Theorem 8 (Wielandt–Hoffman inequality (Kato, 1987)). *Let A, B be self-adjoint operators, such that $C := B - A$ is a trace-class operator, then we can enumerate the eigenvalues of A, B, C as $\alpha_i, \beta_i, \gamma_i$ (including eigenvalue multiplicity) such that*

$$\sum_i |\alpha_i - \beta_i| \leq \sum_i |\gamma_i| \quad (49)$$

The tail of the operator K is the part that comes after the $a_{N,l}$ coefficients. It is bounded in trace norm:

$$\begin{aligned} \text{Tr} \left[\sum_{k=N+1}^{\infty} \sum_{l=1}^{m(k)} a_{k,l} |v_{k,l}\rangle \langle v_{k,l}| \right] &= \sum_{k=N+1}^{\infty} \sum_{l=1}^{m(k)} a_{k,l} \\ &\leq \sum_{k=N+1}^{\infty} \bar{a}_N \epsilon^{k-N} m(k) \\ &\leq \bar{a}_N \sum_{k=N+1}^{\infty} \epsilon^{k-N} A k^D \\ &= O(\bar{a}_N \epsilon) \end{aligned}$$

where we use the polynomial bound $m(k) \leq A k^D$.

Thus, by [Wielandt–Hoffman](#), removing the tail perturbs the spectrum by only $O(\bar{a}_N \epsilon)$. Note particularly:

1. all segments from the 0-th to the N -th remain exponentially separated;
2. the perturbed N -th segment is $O(\epsilon)$ -close in bulk to the original N -th segment.

J.5.2 CUTTING OFF THE HEAD

Cutting off the head is simply cutting off the inverted tail.

Having cut off the tail, we have a finite-rank operator

$$\tilde{K} := \sum_{k=0}^N \sum_{l=1}^{m(k)} a_{k,l} |v_{k,l}\rangle \langle v_{k,l}|$$

that splits into two reducing subspaces $V, V^\perp$, where $V = \text{Span}(|0,1\rangle, \dots, |N, m(N)\rangle)$. It is zero on $V^\perp$ and positive-definite on V . Therefore, we drop down from the full Hilbert space to just V , where we can reason with matrices.

We express $\tilde{K}$ in matrix form in the orthonormal basis $|\hat{v}_{k,l}\rangle$, ordered so that $|\hat{v}_{N,1}\rangle, \dots, |\hat{v}_{N,m(N)}\rangle$ come before $|\hat{v}_{0,1}\rangle, \dots, |\hat{v}_{N-1,m(N-1)}\rangle$:

$$[\tilde{K}] = \begin{bmatrix} A & B \\ B^\top & C + D \end{bmatrix}$$

where the four matrices are:

$$\begin{aligned}
A &= \left[\sum_{l=1}^{m(N)} a_{N,l} \langle \hat{v}_{N,i} | v_{N,l} \rangle \langle v_{N,l} | \hat{v}_{N,j} \rangle \right]_{i,j=1}^{m(N)}, \\
B &= \left[\sum_{l=1}^{m(N)} a_{N,l} \langle \hat{v}_{N,i} | v_{N,l} \rangle \langle v_{N,l} | \hat{v}_{n,j} \rangle \right]_{i \in 1:m(N), (n,j) \in [0:N-1, 1:m]} \\
C &= \left[\sum_{l=1}^{m(N)} a_{N,l} \langle \hat{v}_{n,j} | v_{N,l} \rangle \langle v_{N,l} | \hat{v}_{n',j'} \rangle \right]_{(n,j), (n',j') \in [0:N-1, 1:m]} \\
D &= \left[\sum_{n=0}^{N-1} \sum_{l=1}^{m(N)} a_{n,l} \langle \hat{v}_{n,j} | v_{n,l} \rangle \langle v_{n,l} | \hat{v}_{n',j'} \rangle \right]_{(n,j), (n',j') \in [0:N-1, 1:m]}
\end{aligned}$$

For a symmetric matrix in block form,

$$\begin{bmatrix} A & B \\ B^\top & C + D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} A^{-1} B S^{-1} B^\top A^{-1} & -A^{-1} B S^{-1} \\ -S^{-1} B^\top A^{-1} & S^{-1} \end{bmatrix}$$

where $S = D + C - B^\top A^{-1} B$. We need several crude bounds on A, B, C, D , to prove that the first term really is the bulk term, and the second term really is an order $O(\epsilon)$ perturbation upon the bulk term, and thus we can safely cut it off according to the [Wielandt–Hoffman inequality](#).

1. For each of A, B, C , their entries are all bounded in absolute values by $O(\bar{a}_N)$. Thus, their spectral radii are bounded by $O(\bar{a}_N)$.
2. By [Cauchy interlacing law](#), $\lambda_{\min}(A) \geq \lambda_{\min}([\tilde{K}]) = \Omega(\bar{a}_N)$. Thus, the spectrum of A is $\Theta(\bar{a}_N)$, and the spectrum of A^{-1} is $\Theta(\bar{a}_N^{-1})$.
3. Notice that the matrix D is constructed similarly to the matrix $[\tilde{K}]$, except with one more truncation. Therefore, by the same argument, its least eigenvalue is on the order of $\Omega(\bar{a}_{N-1}) = \Omega(\bar{a}_N/\epsilon)$.
4. Therefore, $B^\top A^{-1}, A^{-1} B$ have entries bounded in order $O(1)$, and $B^\top A^{-1} B$ has entries bounded in order $O(\bar{a}_N)$.
5. Therefore, S has the same spectrum as D with an order $O(\bar{a}_N)$ perturbation. Since D has smallest eigenvalue $\Omega(\bar{a}_{N-1})$, so does S . Therefore, S^{-1} has largest eigenvalue $O(1/\bar{a}_{N-1}) = O(\epsilon/\bar{a}_N)$.

Thus, the N -th segment of the eigenstructure of $\tilde{K}$, inverted, is $O(\epsilon)$ -close in bulk to the eigenstructure of A^{-1} . Inverting again, we find that it is $O(\epsilon)$ -close in bulk to the eigenstructure of A .

J.6 PART 3: THE SPECIAL CASE OF DOT-PRODUCT KERNELS

Recall, we need to show that, in the standard Gaussian space $L^2(\mathcal{N}(0, I_d))$, an operator K defined by kernel $\sum_n \frac{c_n}{n!} (\mathbf{x}^\top \Gamma \mathbf{y})^n$ converges to the Hermite eigensystem. Here, $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_d)$ with $\gamma_1, \dots, \gamma_d > 0$.

Before applying the two generic lemmas. We need to first show that the kernel does have a fast-decaying coefficient system. This is because, even though c_n is a fast-decaying coefficient sequence, it does not automatically imply that K has a fast-decaying coefficient system if we express it in the monomial vector system.

Once this is done, we apply [Appendix J.4](#) to conclude that the kernel's spectrum is divided into exponentially separated segments. Then, we “hopscotch” through several eigenstructures, until we show that the N -th eigensegment of K is $O(\epsilon)$ -close in bulk to the N -th eigensegment of a stretched dot-product kernel $ce^{\epsilon(\mathbf{x}^\top \Gamma \mathbf{y})}$. This argument is nearly the same as the proof of [Appendix J.5](#), with a small extension to account for the special structure of dot-product kernels. Finally, we hopscotch again, applying [Davis–Kahan](#) at every step, to prove that the eigenspaces also converge as $O(\epsilon)$.

J.6.1 COMBINATORICS WITH THE MONOMIAL BASIS

In analogy with the multidimensional Hermite vector system

$$|h_{\alpha}\rangle = \bigotimes_{i=1}^d |h_{\alpha_i}\rangle = \prod_{i=1}^d h_{\alpha_i}(x_i)$$

we define the multidimensional normalized monomial vector system, by taking the tensor product of the single-dimensional normalized monomial vectors:

$$|v_{\alpha}\rangle = \bigotimes_{i=1}^d |v_{\alpha_i}\rangle = \prod_{i=1}^d \frac{x_i^{\alpha_i}}{\sqrt{(2\alpha_i - 1)!!}}$$

Both of them are vector systems over the polynomially bounded multiplicity counting function

$$m(n) := \binom{n + d - 1}{d - 1}$$

The n -th segment of the monomial vector system consists of $\{|v_{\alpha}\rangle : |\alpha| = n\}$, and similarly, for Hermite, $\{|h_{\alpha}\rangle : |\alpha| = n\}$.

The Hermite vector system is obtained by the Gram-Schmidt process on the monomial vector system.

Now, consider the form of the dot-product kernel function:

$$K(x, y) := \sum_{n=0}^{\infty} \frac{c_n}{n!} \left(\sum_{i=1}^d \gamma_i x_i y_i \right)^n = \sum_{n=0}^{\infty} \frac{c_n}{n!} \sum_{|\alpha|=n} \frac{n!}{\alpha_1! \cdots \alpha_d!} \left(\prod_{i=1}^d \gamma_i^{\alpha_i} x_i^{\alpha_i} y_i^{\alpha_i} \right)$$

In bra-ket notation, the operator is

$$K = \sum_{n=0}^{\infty} \sum_{\alpha: |\alpha|=n} c_n \prod_{i=1}^d \left(\frac{\gamma_i^{\alpha_i} (2\alpha_i - 1)!!}{\alpha_i!} \right) |v_{\alpha}\rangle \langle v_{\alpha}|$$

Since c_n is a fast-decaying sequence, it remains to show that the quantity

$$\max_{\alpha: |\alpha|=n} \prod_{i=1}^d \left(\frac{\gamma_i^{\alpha_i} (2\alpha_i - 1)!!}{\alpha_i!} \right)$$

grows *at most* exponentially with n . Once we show that, we know that the coefficient system is also fast-decaying. Equivalently, we need to show that

$$\max_{\alpha: |\alpha|=n} \sum_{i=1}^d \left(\alpha_i \ln(\gamma_i/2) + \ln \binom{2\alpha_i}{\alpha_i} \right)$$

grows *at most* linearly with n . Note that, because such a bound on does not depend on the choice of the fast-decaying sequence c_n , it allows us to change c_n to any other c'_n with the same decay rate ϵ , and still get a fast-decaying coefficient system.

First term is trivial:

$$\sum_{i=1}^d \alpha_i \ln(\gamma_i/2) \in \left[n \min_{i \in [1:d]} \ln(\gamma_i/2), n \max_{i \in [1:d]} \ln(\gamma_i/2) \right] = \Theta(n)$$

To bound the second term, we do some combinatorics. Let $a_k := \binom{2k}{k}$, then we have $\frac{f(k+1)}{f(k)} = 4 - \frac{2}{k+1}$ strictly monotonically increasing. Therefore, f is strictly log-convex. Therefore, $\alpha \mapsto \sum_{i=1}^d \ln f(\alpha_i)$ is strictly Schur-convex. Thus,

$$\sum_{i=1}^d \ln f(\alpha_i)$$

3402 achieves its upper bound at $(n, 0, \dots, 0)$, and lower bound when all α_k are as close to equal as
3403 possible.

3404 Upper bound:

3405
3406
$$\ln f(n) = \ln \binom{2n}{n} = 2n \ln 2 - \frac{1}{2} \ln(\pi n) + O(1/n) = \Theta(n)$$

3407
3408 Lower bound:

3409
3410
$$d \ln f(n/d) = d \ln \binom{2n/d}{n/d} = 2n \ln 2 - \frac{1}{2} \ln(\pi n/d) + O(1/n) = \Theta(n)$$

3411
3412 In fact, we have shown that the coefficient block is very tightly clustered:

3413
3414
$$\sum_{i=1}^d \ln f(\alpha_k) = 2n \ln 2 - \frac{1}{2} \ln(\pi n) + [-\ln \sqrt{d}, 0] + O(1/n)$$

3415
3416 for all α satisfying $|\alpha| = n$.

3417
3418
3419
3420
3421
3422
3423
3424
3425
3426
3427
3428
3429
3430
3431
3432
3433
3434
3435
3436
3437
3438
3439
3440
3441
3442
3443
3444
3445
3446
3447
3448
3449
3450
3451
3452
3453
3454
3455

J.6.2 CONVERGENCE IN BULK

Like the general case, the proof has multiple parts. First, we apply [Appendix J.4](#) to conclude that K has exponentially separated eigensegments. Then, we hopscotch through multiple eigensystem segments to show eigensegment convergence in bulk, as in [Appendix J.5](#).

$$\begin{aligned}
& N\text{-th segment of } K(\mathbf{x}, \mathbf{y}) = \sum_{n=0}^{\infty} \frac{c_n}{n!} (\gamma_1 x_1 y_1 + \cdots + \gamma_d x_d y_d)^n \\
& \quad \parallel \\
& \quad O(\epsilon) \\
& N\text{-th segment of } \tilde{K}(\mathbf{x}, \mathbf{y}) = \sum_{n=0}^N \frac{c_n}{n!} (\gamma_1 x_1 y_1 + \cdots + \gamma_d x_d y_d)^n \\
& \quad \parallel \\
& \quad (0\text{-th segment of } [\tilde{K}]^{-1})^{-1} \\
& \quad \parallel \\
& \quad O(\epsilon) \\
& \quad (\text{eigensystem of just the } a_N \text{ part of } [\tilde{K}]^{-1})^{-1} \\
& \quad \parallel \\
& \quad \text{provided that } c'_N = c_N \\
& \quad (\text{eigensystem of just the } a_N \text{ part of } [\tilde{K}']^{-1})^{-1} \\
& \quad \parallel \\
& \quad O(\epsilon) \\
& \quad (0\text{-th segment of } [\tilde{K}']^{-1})^{-1} \\
& \quad \parallel \\
& N\text{-th segment of } \tilde{K}'(\mathbf{x}, \mathbf{y}) = \sum_{n=0}^N \frac{c'_n}{n!} (\gamma_1 x_1 y_1 + \cdots + \gamma_d x_d y_d)^n \\
& \quad \parallel \\
& \quad O(\epsilon) \\
& N\text{-th segment of } K'(\mathbf{x}, \mathbf{y}) = \sum_{n=0}^{\infty} \frac{c'_n}{n!} (\gamma_1 x_1 y_1 + \cdots + \gamma_d x_d y_d)^n \\
& \quad \parallel \\
& \quad \text{set } K'(\mathbf{x}, \mathbf{y}) = c e^{\epsilon(\mathbf{x}^\top \Gamma \mathbf{y})}, \text{ where } c = c_N / \epsilon^N \\
& N\text{-th segment of } K'(\mathbf{x}, \mathbf{y}) = \frac{c_N}{\epsilon^N} e^{\epsilon(\mathbf{x}^\top \Gamma \mathbf{y})} \\
& \quad \parallel \\
& \quad O(\epsilon) \\
& N\text{-th segment of } \left\{ \left(\frac{c_N}{\epsilon^N} \epsilon^{|\alpha|} \prod_{i=1}^d \gamma_i^{\alpha_i}, |h_\alpha\rangle \right) \text{ for all } \alpha \in \mathbb{N}_0^d \right\} \\
& \quad \parallel \\
& \quad \left\{ \left(c_N \prod_{i=1}^d \gamma_i^{\alpha_i}, |h_\alpha\rangle \right) \text{ for all } \alpha \in \mathbb{N}_0^d, |\alpha| = N \right\}
\end{aligned}$$

Figure 24: Hopscotching through eigensystems.

The new trick here is that we avoid solving for the N -th eigensegment of $\tilde{K}$, by going down then up again, arriving at a previously solved kernel as if taking a subway.

J.6.3 CONVERGENCE IN DETAIL

We have successfully proven the source eigensegment converges in bulk to the target eigensegment. It remains to prove convergence in detail. This requires breaking degeneracies in the source eigensegment whenever the degeneracy is broken in the target eigensegment (but not vice versa), followed by applying [Davis–Kahan](#) once per target eigenspace.

Fix some N . The N -th target eigensegment is

$$\left\{ \left(c_N \prod_{i=1}^d \gamma_i^{\alpha_i}, |h_{\alpha}\rangle \right) \text{ for all } \alpha \in \mathbb{N}_0^d, |\alpha| = N \right\}$$

Some of the eigenvalues in it may be equal, because $\{\ln \gamma_1, \dots, \ln \gamma_d\}$ may be $\mathbb{N}$ -linearly dependent.¹³ Therefore, merge the eigenvectors with equal eigenvalue into eigenspaces. The eigenspaces have distinct eigenvalues. Note in particular that this constitutes a *static target*: The coefficients $c_0, c_1, \dots$ may change, but if $c_N \prod_{i=1}^d \gamma_i^{\alpha_i} = c_N \prod_{i=1}^d \gamma_i^{\alpha'_i}$ for one choice of the coefficients, then it is so for all choices.

Now, fix one such eigenspace V and its corresponding eigenvalue $c_N \zeta$. Let the N -th segment of K be

$$\{(c_N \zeta_{N,1}(K), |v_{N,1}(K)\rangle), \dots, (c_N \zeta_{N,m(N)}(K), |v_{N,m(N)}(K)\rangle)\}$$

where we do not yet demand that the eigenvalues are all distinct. If some of the eigenvalues are unluckily degenerate, we just tolerate an arbitrary choice of the eigenvectors for now.

As previously argued, it is $O(\epsilon)$ -close to the target eigensegment in bulk. Therefore, for small enough ϵ , the source eigenvalues $c_N \zeta_{N,1}(K), \dots, c_N \zeta_{N,m(N)}(K)$ will be corralled around their corresponding target eigenvalues, like iron filings around magnets. Because there are only $m(N)$ source eigenvalues to go around, each target eigenvalue can only grab exactly as many source eigenvalues as its multiplicity. In particular, this means that these “herds” of eigenvalues may be highly degenerate together, they are *separated* from other herds by $\bar{a}_N \Theta(1)$.

In particular, this means $c_N \zeta$ will be able to grab exactly $\dim V$ source eigenvalues, so that they are all stuck within an interval $c_N(\zeta \pm O(\epsilon))$. Enumerate these as $c_N \zeta_{N,i_1}(K), \dots, c_N \zeta_{N,i_{\dim V}}(K)$. Let their eigenvectors be $|v_{N,i_1}(K)\rangle, \dots, |v_{N,i_{\dim V}}(K)\rangle$, and let the vector space spanned by them be V_K . This is a reducing subspace of K , and one that we wish to show as $O(\epsilon)$ -close to V .

Follow through the [hopscotching diagram](#) again. At each step in the hopscotching, either there is an exact equality, in which case the reducing subspace is unchanged, or there is a perturbation on the operator that is $O(\epsilon)$ relative to the operator, in which case, by [Davis–Kahan](#), the reducing subspace is perturbed by an angle of only $O(\epsilon)$.

We spell this out explicitly for the top half of the diagram.

At the first step, $K = \sum_{n=0}^{\infty} \frac{c_n}{n!} (x^\top \Gamma y)^n$ is perturbed by truncating the tail $\sum_{n=N+1}^{\infty} \frac{c_n}{n!} (x^\top \Gamma y)^n$. This has operator norm $O(c_N \epsilon)$. Since the gap between $c_N \zeta_{N,i_1}(K), \dots, c_N \zeta_{N,i_{\dim V}}(K)$ and all other eigenvalues of K is on the order of $\bar{a}_N \Theta(1)$. Thus, by [Davis–Kahan](#), truncating the tail only perturbs the reducing subspace V_K by an angle $O(\epsilon)$.

The second step is an exact identity. Under a matrix inverse, the eigenspaces are preserved, even though the eigenvalues are inverted.

In the third step, the inverted head is truncated. The inverted head has operator norm on the order of $O(\epsilon/\bar{a}_N)$, while the remaining operator has operator norm on the order of $O(1/\bar{a}_N)$. Now, since the gap between $(c_N \zeta_{N,i_1}(K))^{-1}, \dots, (c_N \zeta_{N,i_{\dim V}}(K))^{-1}$ and all other inverted eigenvalues of $[K]^{-1}$ is on the order $\frac{1}{\bar{a}_N} \Theta(1)$, truncating the inverted head only perturbs the reducing subspace by another angle $O(\epsilon)$.

□

¹³Even if they are $\mathbb{N}$ -linearly independent, the gaps between successive eigenvalues will get smaller for larger N if $d \geq 3$. By the standard counting argument in Diophantine approximation theory, the gap decays at least as fast as $N^{-(d-2)}$.