How does feature learning change the function space evolution?

João Lobo * **Bruno Loureiro** † **Long Tran-Thanh*** **Fanghui Liu** ‡

Abstract

Feature learning is widely regarded as the key mechanism distinguishing neural networks from fixed-kernel methods, yet its impact on the induced function space remains poorly understood. In this work, we characterize how the function space spanned by the features of a two-layer neural network evolves during gradient descent early training. After a large gradient step, we show that the feature distribution induced by the structured rank-one component of the update is, at the population level, well approximated by a target-dependent spiked Gaussian covariance. This change induces a data-adaptive kernel with a different spectral structure that reshapes the function space. We investigate this structure, showing that it selectively amplifies eigenvalues aligned with the target direction and mixes leading eigenfunctions, coupling the top radial mode with a target-aligned quadratic harmonic. Overall, our results provide a function-space perspective on early-stage feature learning: rather than just rescaling a fixed kernel, gradient descent induces a data-adaptive deformation that preferentially enhances directions aligned with the signal in the data.

1 Introduction

The success of modern neural networks (NNs) is often attributed to feature learning, namely the ability of models to adapt to the structure of the data during training (Bach, 2017a; Suzuki, 2019; Damian et al., 2022). This stands in contrast to non-adaptive approaches, such as kernel methods and neural networks operating in the so-called lazy training regime (Jacot et al., 2018; Chizat et al., 2019), where features remain effectively frozen at their random, data-independent initialization.

This contrast becomes explicit in random feature models (RFMs) (Rahimi and Recht, 2007) and two-layer NNs, both of which can be written in the form the form $f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} a_i \phi(\boldsymbol{x}, \boldsymbol{w}_i)$ with a feature map $\phi : \mathcal{X} \times \mathcal{W} \to \mathbb{R}$. The key difference lies in whether the features are learned: in RFMs, the weights $\{\boldsymbol{w}_i\}_{i=1}^m \in \mathcal{W}$ are sampled i.i.d. from a probability measure μ and kept fixed, and only the second layer weights $\boldsymbol{a} := \{a_i\}_{i=1}^m$ are optimized, whereas in NNs both layers are jointly trained.

From a function-space perspective, RFMs with ℓ_2-regularization on $\boldsymbol{a}$ are equivalent to kernel methods via the representer theorem (Schölkopf et al., 2001), with empirical kernel $\widehat{k}(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{m}\sum_{i=1}^{m} \phi(\boldsymbol{x}, \boldsymbol{w}_i)\phi(\boldsymbol{x}', \boldsymbol{w}_i)$, which approximates, as $m \to \infty$, the population kernel

$$k_0(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathcal{W}} \phi(\boldsymbol{x}, \boldsymbol{w})\phi(\boldsymbol{x}', \boldsymbol{w}) \, \mathrm{d}\mu(\boldsymbol{w}) \, .$$

*Department of Computer Science, University of Warwick, United Kingdom; Emails: {joao.lobo-pevidor, long.tran-thanh}@warwick.ac.uk

†Departement d'Informatique, École Normale Supérieure, PSL & CNRS; Email: bruno.loureiro@di.ens.fr

‡School of Mathematical Sciences, Institute of Natural Sciences and MOE-LSC, Shanghai Jiao Tong University, China. Email: fanghui.liu@sjtu.edu.cn (Corresponding author)

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The associated reproducing kernel Hilbert space (RKHS), denoted $\mathcal{H}_0(\mu)$, is the closure of functions expressible as weighted averages of the features $\phi(\cdot, \boldsymbol{w})$.

When optimizing $\boldsymbol{a}$, weaker regularization yield strictly larger spaces (Celentano et al., 2021; Chen et al., 2025), up to the $\mathcal{F}_1$-type space of ℓ_1-regularization Bach (2017a), which are closely related to the Barron space $\mathcal{B}$ (Barron, 1993; E et al., 2021). In Chen et al. (2025), it is shown that $\mathcal{B} = \bigcup_{\mu \in \mathcal{P}(\mathcal{W})} \mathcal{H}_0(\mu)$, where $\mathcal{P}(\mathcal{W})$ denotes the set of probability measures on $\mathcal{W}$, so the function space of a two-layer network is governed by its feature distribution. Notably, under the mean-field regime (Chizat, 2021), the function space induced by two-layer neural networks forms a subset of the Barron space (Wojtowytsch and E, 2020). These perspectives describe the spaces induced by different measures, but not how the measure, and hence the space, evolves during training.

In parallel, recent work has shown that one step of gradient descent induces informative low-rank structure in the first-layer weights. The update is approximately a rank-one spike plus a residual of smaller operator norm (Ba et al., 2022; Cui et al., 2024; Dandi et al., 2025), which injects outliers into the spectrum of the feature matrix and aligns it with the target Moniri et al. (2023); Moniri and Hassani (2026). These works describe how parameters and features evolve, but leave the consequences for the induced function space unexplored.

In this work, we connect these two perspectives. Since the updated weights depend on the training data, one gradient step changes the initial weight distribution μ to a data dependent distribution μ^*. We study the function space consequence of this through the population kernel of μ^*, its integral operator, and their spectral geometry. In particular, we work with the Gaussian single-index model, which has been shown to capture the most general class of functions learnable in the proportional high-dimensional regime (Damian et al., 2022; Ba et al., 2022; Dandi et al., 2024). In this context, we study the following question:

How does (one-step) gradient descent change the function space?

1.1 Contributions

Our contributions can be summarized as follows:

1. Approximation from learned features to a target-dependent kernel: In Section 3, we prove that the ideal post-update feature distribution is well approximated by a target-dependent spiked Gaussian, leading to a data-adaptive kernel k_1 (*c.f.* Theorem 3.1). It shows that feature learning can be approximately expressed as target-dependent distribution transformation either in the parameter space or in the input-data space on such kernel.

2. Expansion of the data-adaptive kernel: In Section 4, we study the spectrum of the kernel k_1. After conducting a Taylor expansion of k_1 around an isotropic kernel, we isolate the contribution of the spike. We prove that the higher-order remainder terms are asymptotically negligible and characterize the leading deformation in terms of projections onto a distinguished direction determined by the target vector (*c.f.* Theorem 4.1).

3. Feature learning in the top eigenspaces for the ReLU activation: In Section 5, we consider the case of the ReLU activation function, to give an explicit characterization of the kernel k_1, its spectrum and the spanned function space. Our results theoretically prove that the spike transforms primarily the top and linear eigenspaces of the operator (*c.f.* Theorem 5.3, 5.5). Our numerical validations support our theoretical findings and also illustrate the connection between data-adaptive kernels and neural networks.

Notably, the strength of these effects is governed by the step size: larger steps induce stronger deformations, with growing contributions of both linear and higher-order nonlinear features.

1.2 Related works

The study of function spaces in deep learning theory stems from RKHS, as in Jacot et al. (2018) and Lee et al. (2018). However, these naturally operate in a lazy training regime (Chizat et al., 2019),

hence their ability for active feature learning is limited (Ghorbani et al., 2019) and they are not able to access any of the one-step parameter-space results mentioned earlier. In parallel, Xu and Zheng (2024) studies feature learning via the *feature geometry* perspective, unifying statistical dependence and feature representations in an inner-product space. However, their focus is on learning optimal features with standard networks rather than understanding how features evolve during training. Furthermore, Dou and Liang (2021) studied the conjugate kernel RKHS $\mathcal{H}_t$ and the neural tangent kernel RKHS $\mathcal{K}_t$ through a signed measure. Our focus is instead on the spectral geometry induced by the representation learned from one-step gradient descent.

2 Problem Setting and Preliminaries

Consider a supervised regression problem with training data $\mathcal{D} = (\boldsymbol{x}_i, y_i)_{i=1}^{n+N}$, which we will assume is drawn from a Gaussian single-index model:

$$y_i = f^*(\boldsymbol{x}_i) + \varepsilon_i := g(\langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle) + \varepsilon_i, \quad \boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)\,, \tag{2.1}$$

where $\boldsymbol{w}^* \in \mathbb{S}^{d-1}$, $g : \mathbb{R} \to \mathbb{R}$ is a link function and ε_i are i.i.d. sub-Gaussian noise random variables with zero mean and variance σ_ε^2. This synthetic data model has been the subject of several different works in the theoretical literature, where it was studied as a testbed for separation between lazy and feature learning regimes in the high-dimensional limit (Ben Arous et al., 2021; Damian et al., 2022; Ba et al., 2020; Dandi et al., 2024). In particular, it was shown that despite being efficiently learnable with $n = \Theta(d)$ samples for generic g (Li, 1991; Babichev and Bach, 2018; Barbier et al., 2019; Damian et al., 2024; Troiani et al., 2025), non-adaptive kernel methods require infinite data to learn it with arbitrary precision. Given the training data, we will consider the problem of learning it with a two-layer neural network defined by $f(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{a}) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \sigma(\boldsymbol{w}_j^\top \boldsymbol{x})$, where $\boldsymbol{W} \in \mathbb{R}^{d \times m}$, $\boldsymbol{a} \in \mathbb{R}^m$ are the first and second layer weights, respectively, and $\sigma : \mathbb{R} \to \mathbb{R}$ is an element-wise activation function. We make the following standard assumptions:

Assumption 2.1 (Main assumptions)**.**

1. (**Initialization**) At initialization, the first layer weights are distributed as $\boldsymbol{w}_j^0 \sim \mathcal{N}(0, \frac{1}{d}\boldsymbol{I}_d)$ and the second-layer weights are distributed as $\boldsymbol{a}^0 \sim \mathcal{N}(0, \frac{1}{m}\boldsymbol{I}_m)$.
2. (**High-dimensional proportional regime**) We work under the proportional regime scaling, defined as the limit where $n, m, \eta, d \to \infty$ at fixed ratios:
$$\alpha := \frac{n}{d}, \qquad \beta := \frac{m}{d}, \qquad \widetilde{\eta} := \frac{\eta}{d^\zeta} \tag{2.2}$$
where the step size parameter admits $\eta = \Theta(d^\zeta)$ for some $\zeta \in [1/2, 1)$.
3. (**Activation and target functions**) The activation function σ is uniformly L_σ-Lipschitz and g is uniformly bounded and L_g-Lipschitz with $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[g(z)] = 0$ and $c_1 := \mathbb{E}_{z\sim\mathcal{N}(0,1)}[g'(z)] \neq 0$. Equivalently, in the language of Ben Arous et al. (2021) we assume g has information exponent 1.

We employ a two-stage training scheme under the empirical risk minimization via the squared loss as in Ba et al. (2022); Dandi et al. (2024); Moniri et al. (2023); Cui et al. (2024); Dandi et al. (2025). Splitting the training data into two parts we have the following framework:

- Assume we use n i.i.d samples from Eq. (2.1) to train the first-layer by one single step, while keeping the second-layer fixed:

$$\begin{aligned} \boldsymbol{w}_j^1 &= \boldsymbol{w}_j^0 - \eta \boldsymbol{g}_j^0, \quad \forall j \in [m] \\ \boldsymbol{g}_j^0 &= \frac{1}{n\sqrt{m}} \sum_{i=1}^{n} \left(f(\boldsymbol{x}_i; \boldsymbol{W}^0, \boldsymbol{a}^0) - y_i \right) a_j^0 \boldsymbol{x}_i \sigma'({\boldsymbol{w}_j^0}^\top \boldsymbol{x}_i)\,. \end{aligned} \tag{2.3}$$

- Given the updated weights $\boldsymbol{W}^1$, we update the second-layer weights via ridge regression using another N samples i.i.d from Eq. (2.1)

$$\widehat{\boldsymbol{a}}_\lambda = \underset{\boldsymbol{a}\in\mathbb{R}^m}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - f(\boldsymbol{x}_i; \boldsymbol{a}, \boldsymbol{W}^1)\right)^2 + \lambda\|\boldsymbol{a}\|_2^2 = \left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}/m + \lambda I_m\right)^{-1} \boldsymbol{\Phi}^\top\boldsymbol{y}/\sqrt{m}\,, \tag{2.4}$$

  where the feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{N\times m}$ with elements $\phi_{ij} = \sigma(\boldsymbol{x}_i^\top \boldsymbol{w}_j^1)$ and the label vector $\boldsymbol{y} = [y_1, y_2, \cdots, y_N]^\top$.

The effect of training on the first-layer weights is known to depend on the scaling of η. Considering the data generating process from Eq. (2.1), denote by $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$ the data matrix and label vector seen during the training step in Eq. (2.3).

After one gradient step, the first-layer weights admit the decomposition (Ba et al., 2022; Moniri et al., 2023; Dandi et al., 2025; Moniri and Hassani, 2026):

$$\boldsymbol{W}_1 = \boldsymbol{W}_0 + \bar{\boldsymbol{A}} + \boldsymbol{E}, \qquad \bar{\boldsymbol{A}} := \frac{\mu_1\eta}{\sqrt{m}} \frac{\boldsymbol{X}^\top \boldsymbol{y}\boldsymbol{a}_0^\top}{n}\,, \tag{2.5}$$

where $\bar{\boldsymbol{A}}$ is the structured update and $\boldsymbol{E}$ collects residual fluctuations, with $\|\boldsymbol{E}\|_{op}$ remaining below $\|\bar{\boldsymbol{A}}\|_{op}$ by a factor of $\widetilde{\mathcal{O}}(d^{-1/2})$. At the first step, the formulation Eq. (2.5) also applies to the intermediate regime $\eta = \Theta(d^\zeta)$ with $\zeta \in [1/2, 1)$ (Moniri et al., 2023; Moniri and Hassani, 2026), as well as $\eta = \Theta(d)$ (Cui et al., 2024; Dandi et al., 2025). When $\eta = \Theta(\sqrt{d})$, it also persists over any number of fixed t steps (Ba et al., 2022; Wang et al., 2024). Although $\bar{\boldsymbol{A}}$ has rank one, applying the nonlinear activation can generate additional spectral components in the learned feature matrix $\sigma(\boldsymbol{X}\boldsymbol{W}_1)$ (Moniri et al., 2023).

Notation: We denote vectors in high dimensional spaces by bold lowercase letters ($\boldsymbol{v}$) and matrices/operators by bold uppercase letters ($\boldsymbol{A}$). Functions (f) and functional operators (T) are represented in standard typeface. We let $\rho_{\mathcal{X}}$ be the measure induced by the input distribution $\mathcal{N}(0, \boldsymbol{I}_d)$, and $L^2(\rho_{\mathcal{X}})$ denote the associated Hilbert function space. The notation $\|\cdot\|$ refers specifically to the standard Euclidean norm in $\mathbb{R}^d$. Whenever an alternative norm is employed, it will be explicitly denoted with the appropriate subscript (e.g., $\|\cdot\|_{L^2(\rho_{\mathcal{X}})}$ or $\|\cdot\|_{\text{op}}$). Using standard asymptotic notation $\mathcal{O}, o, \Omega, \Theta$, we track dependencies on d and on the spike strength B. We further use the shorthand notation $o_d(f) := o(f \cdot d^{-c})$ for some $c \in (0, 1)$ to identify terms that vanish slower than $o(f/d)$.

3 Feature learning as a distribution transformation in the kernels

Under Gaussian initialization, the network implicitly induces the baseline kernel $k_0(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{1}{d}\boldsymbol{I}_d)}[\sigma(\langle\boldsymbol{w}, \boldsymbol{x}\rangle)\sigma(\langle\boldsymbol{w}, \boldsymbol{x}'\rangle)]$ as mentioned before. To capture the network's capacity for feature learning, our analysis isolates the structural deformation of this kernel driven solely by a rank-one update to the weights.

After one gradient step under Assumption 2.1, we have $(\boldsymbol{W}_1^\top \boldsymbol{x})_j = \langle \boldsymbol{w}_j^0, \boldsymbol{x}\rangle + \langle \boldsymbol{v}_j, \boldsymbol{x}\rangle$ for vectors $\boldsymbol{v}_j = \frac{\mu_1\eta a_j^{(0)}}{\sqrt{m}} \frac{\boldsymbol{\varsigma}}{n} \in \mathbb{R}^d$, so all neurons share the same empirical update direction $\boldsymbol{\varsigma}/n$. Each updated neuron $\boldsymbol{z} = \boldsymbol{w} + \boldsymbol{v}$, $\boldsymbol{v} = \frac{\mu_1\eta}{n\sqrt{m}} a\boldsymbol{\varsigma}$, is thus a random vector whose law μ^* is induced jointly by initialization $(\boldsymbol{w}, a)$ and the first-stage training set $\mathcal{D}$. With this in mind, we study the random feature kernel of μ^*:

$$k_1^*(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathcal{W}} \phi(\boldsymbol{x}, \boldsymbol{w})\, \phi(\boldsymbol{x}', \boldsymbol{w})\, \mathrm{d}\mu^*(\boldsymbol{w}) = \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\boldsymbol{w},a}[\phi(\boldsymbol{x}, \boldsymbol{w} + \boldsymbol{v})\phi(\boldsymbol{x}', \boldsymbol{w} + \boldsymbol{v})]\,. \tag{3.1}$$

Writing $k_{\mathcal{D}}(\boldsymbol{x}, \boldsymbol{x}') := \mathbb{E}_{\boldsymbol{w},a}\left[\phi(\boldsymbol{x}, \boldsymbol{w} + \boldsymbol{v})\, \phi(\boldsymbol{x}', \boldsymbol{w} + \boldsymbol{v}) \mid \mathcal{D}\right]$ for the kernel obtained from a fixed training set $\mathcal{D}$, Fubini's theorem gives $k_1^* = \mathbb{E}_{\mathcal{D}}[k_{\mathcal{D}}]$. The kernel k_1^* therefore depends on the data only through its distribution, not through the particular sample, and we call it the *idealized* kernel of the update. The neurons of one trained network all share the same $\mathcal{D}$, so their empirical distribution follows the

Note that, k_t^* for constant steps t under the small step-size $\eta = \Theta(\sqrt{d})$ also admits this formulation. But for a unifying analysis framework, we study the impact of the learning rate coming from the regime Eq. (2.2) under the first step. In the following, we will investigate how the spike $\boldsymbol{v}$ impact the kernel as well as the associated function spaces.

As two positive semi-definite kernels, k_0 and k_1^* uniquely determine two different RKHS: $\mathcal{H}_0, \mathcal{H}_1^* \subset L^2(\rho_{\mathcal{X}})$. We let ρ be an unknown probability distribution on $\mathcal{X} \times \mathcal{Y}$ satisfying $\int_{\mathcal{X}\times\mathcal{Y}} y^2 \mathrm{d}\rho(\boldsymbol{x}, y) < \infty$, and denote its corresponding marginal distribution over the inputs as $\rho_{\mathcal{X}}$. Following Bach (2017b), we can associate each kernel with a self-adjoint, positive semi-definite, trace-class integral operator. For the two conjugate kernels $k_0, k_1^\star$, the operators $T_0, T_1^* : L^2(\rho_{\mathcal{X}}) \to L^2(\rho_{\mathcal{X}})$ are given by:

$$(T_0 f)(\boldsymbol{x}) = \int_{\mathcal{X}} k_0(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \quad \text{and} \quad (T_1^* f)(\boldsymbol{x}) = \int_{\mathcal{X}} k_1^*(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \,. \tag{3.2}$$

By the spectral theorem, both operators admit different spectral decompositions, and Mercer's theorem states that we can represent each kernel by its respective spectral decomposition. Therefore, we will study the distribution of the learned features, the formulation of the kernel k_1^*, as well as the spectrum of function spaces from $\mathcal{H}_0$ to $\mathcal{H}_1^*$.

3.1 Distribution and moments of the learned features

Let $\boldsymbol{\varsigma} := \sum_{i=1}^n y_i \boldsymbol{x}_i$. The learned feature is given by $\boldsymbol{z} := \boldsymbol{w} + \frac{\mu_1 \eta}{n\sqrt{m}} a \boldsymbol{\varsigma}$. Since a is a scalar Gaussian random variable and $\boldsymbol{w}$ is a Gaussian random vector, both independent, $\boldsymbol{z}$ has zero mean. Moreover, conditioned on the dataset (or equivalently conditioned on $\boldsymbol{\varsigma}$) we have:

$$\boldsymbol{z} \mid \boldsymbol{\varsigma} \sim \mathcal{N}\left(0,\ \frac{1}{d}\boldsymbol{I}_d + \frac{\mu_1^2 \eta^2}{n^2 m^2} \boldsymbol{\varsigma}\boldsymbol{\varsigma}^\top\right) .$$

Therefore, conditionally on $\boldsymbol{\varsigma}$, $\boldsymbol{z}$ is a Gaussian vector with a spiked covariance along the random direction $\boldsymbol{\varsigma}$. Without conditioning, the distribution of $\boldsymbol{z}$ is complex. In particular, its covariance matrix is given by

$$\frac{1}{d}\boldsymbol{\Gamma} := \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \underbrace{\left(1 + \frac{\eta^2 \lambda_{\alpha,\beta}}{d^2}\mathbb{E}[f^*(\boldsymbol{x})^2] + \frac{\eta^2 \lambda_{\alpha,\beta}}{d^2}\sigma_\varepsilon^2\right)}_{:=A} \frac{\boldsymbol{I}_d}{d} + \underbrace{\left[\frac{\eta^2 \lambda_{\alpha,\beta} c_1^2}{d}\left(\alpha - \frac{1}{d} + \frac{2s}{c_1^2 d}\right)\right]}_{:=B} \frac{\boldsymbol{w}^*(\boldsymbol{w}^*)^\top}{d} , \tag{3.3}$$

where $\lambda_{\alpha,\beta} := \frac{\mu_1^2}{\alpha\beta^2}$, α, β are constants as defined in Eq. (2.2), c_1 is the first Hermite coefficient of g, μ_1 is the first Hermite coefficient of σ and, by denoting $h(t) := [g(t)]^2$, we have

$$s := \mathbb{E}[g'(\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle)^2] + \mathbb{E}[g(\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle) g''(\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle)] = \frac{1}{2}\mathbb{E}[h''(\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle)] \,.$$

For short, we write $\boldsymbol{\Gamma} := A\boldsymbol{I}_d + B\boldsymbol{w}^*(\boldsymbol{w}^*)^\top$, and defer the full derivation of these quantities to Appendix A.1. See the discussion on when $s \gtreqless 0$ and examples for particular choices of g in Appendix A.2, some of which may violate Assumption 2.1 and are presented to provide intuition.

Note that B has three terms and α will dominate in the high-dimensional asymptotic regime. Accordingly, we can ensure $\boldsymbol{\Gamma}$ to be positive definite.

As previously discussed, the (unconditional) distribution of $\boldsymbol{z}$ is complex, but in the next subsection we will show that it can be well approximated by a Gaussian distribution with matching covariance $\boldsymbol{\Gamma}$. This will allow us to derive an approximation for the associated kernel.

3.2 The kernel formulation

We emphasize that k_1^* (from Eq. (3.1)) denotes the kernel induced by the ideal rank-one feature update, which the next theorem shows can be approximated by a Gaussian kernel k_1 with matching covariance (proof available in Appendix C.1).

Theorem 3.1. Denote $\varsigma := \sum_{i=1}^{n} y_i \boldsymbol{x}_i$ and $\boldsymbol{z} := \boldsymbol{w} + \frac{\mu_1 \eta}{n\sqrt{m}} a \varsigma$ and let μ^* be the distribution of $\boldsymbol{z}$. Then, under Assumption 2.1, given the matrix $\boldsymbol{\Gamma}$ defined by Eq. (3.3), k_1^* under μ^* can be approximated by the following kernel

$$k_1(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{w} \sim \mathcal{N}(0, \frac{1}{d}\boldsymbol{\Gamma})}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)\sigma(\langle \boldsymbol{w}, \boldsymbol{x}' \rangle)] = k_0(\boldsymbol{\Gamma}^{1/2}\boldsymbol{x}, \boldsymbol{\Gamma}^{1/2}\boldsymbol{x}') \,, \tag{3.4}$$

such that their respective integral operators T_1^* and T_1 admit

$$\|T_1^* - T_1\|_{\text{HS}} = \|k_1^* - k_1\|_{L^2(\rho_{\mathcal{X}}) \times L^2(\rho_{\mathcal{X}})} = \mathcal{O}\left(\frac{\eta^4}{d^4}\right) .$$

Moreover, defining the pushforward measure of the standard Gaussian measure $\nu = (\rho_{\mathcal{X}} \circ \boldsymbol{\Gamma}^{-1/2})$ such that $k_0(\boldsymbol{z}, \boldsymbol{z}') = \sum_{i=0}^{\infty} \omega_i \boldsymbol{e}_i(\boldsymbol{z}) \boldsymbol{e}_i(\boldsymbol{z}')$ with respect to ν, the spectral decomposition of k_1 with respect to $\rho_{\mathcal{X}}$ is given by $k_1(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=0}^{\infty} \omega_i \boldsymbol{e}_i(\boldsymbol{\Gamma}^{1/2}\boldsymbol{x}) \boldsymbol{e}_i(\boldsymbol{\Gamma}^{1/2}\boldsymbol{x}')$.

Remark 3.2. Our result builds upon macroscopic results of order $\mathcal{O}(|B|/d)$, by repeatedly showing the trailing terms of our approximations decay with rate $o_d(|B|/d)$. This bound shows that the approximation of k_1^* via k_1 respects the decay needed in order for the result to hold. Indeed, since $|B| = \Theta\left(\eta^2/d\right)$, we have $\mathcal{O}\left(\frac{\eta^4}{d^4}\right) \subset o_d\left(\frac{|B|}{d}\right)$. Thanks to Theorem 3.1, we can work with k_1 throughout this paper.

Data-adaptive kernel: The relationship between k_0 and k_1 in Eq. (3.4) can be characterized by a distribution transformation in either parameters or data. From the **parameter-space** perspective, feature learning transforms the parameter distribution from its initial standard Gaussian distribution into a target-dependent distribution shaped by the training objective. From the **data-space** perspective, feature learning induces a shift in the representation of the input data, moving it toward a structure that is more aligned with the target function or labels. This can be a data-adaptive kernel, similar in spirit to Follain and Bach (2025); Huang et al. (2025) with a fixed base kernel with a low-dimensional linear map. In contrast, our transformation acts in the full ambient space, leading to anisotropic amplification of signal directions rather than dimensionality reduction.

4 Feature learning as an additional target-dependent kernel

4.1 The spiked covariance expansion framework

To understand the connection between k_1 and k_0, we present an Taylor-like expansion with the goal of isolating the contributions of the spike, with proof deferred to Appendix C.2.

Theorem 4.1. Let $\boldsymbol{\Sigma} := \gamma_1 \boldsymbol{I} + \gamma_2 \boldsymbol{u}\boldsymbol{u}^\top$ be a covariance matrix with $\gamma_1, \gamma_2 > 0$ and $\|\boldsymbol{u}\| = 1$, and $G : \mathbb{R}^d \to \mathbb{R}$ be a measurable function of at most polynomial growth. Denote $D_{\boldsymbol{u}}^{(j)} G$ as the j-th order directional derivative of G along $\boldsymbol{u}$ in the sense of tempered distributions, defined by its action on any test function $\varphi \in \mathcal{S}(\mathbb{R}^d)$ as

$$\langle D_{\boldsymbol{u}}^{(j)} G, \varphi \rangle := (-1)^j \langle G, D_{\boldsymbol{u}}^{(j)} \varphi \rangle \,, \text{where } D_{\boldsymbol{u}} \varphi(\boldsymbol{w}) = \langle \nabla \varphi(\boldsymbol{w}), \boldsymbol{u} \rangle \,.$$

Then, whenever the series on the right-hand side converges, we have that

$$\mathbb{E}_{\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{\Sigma})}[G(\boldsymbol{w})] = \mathbb{E}_{\boldsymbol{w} \sim \mathcal{N}(0, \gamma_1 \boldsymbol{I}_d)}[G(\boldsymbol{w})] + \sum_{j=1}^{\infty} \frac{1}{j!} \left(\frac{\gamma_2}{2}\right)^j \mathbb{E}_{\boldsymbol{w} \sim \mathcal{N}(0, \gamma_1 \boldsymbol{I}_d)}[D_{\boldsymbol{u}}^{(2j)} G(\boldsymbol{w})] \,.$$

Remark 4.2. Since we are interested in analyzing $\boldsymbol{\Gamma}$, we note that the conditions on γ_1 and γ_2 naturally translate to our case since $A, B > 0$ in the asymptotic regime. Also note that because the growth of G is polynomially bounded (Assumption 2.1) and the Gaussian density belongs to $\mathcal{S}(\mathbb{R}^d)$, the terms $\mathbb{E}_{\boldsymbol{w} \sim \mathcal{N}(0, \gamma_1 \boldsymbol{I}_d)}[D_{\boldsymbol{u}}^{(2n)} G(\boldsymbol{w})]$ are well-defined for all $n \geq 0$.

Remark 4.3. This expansion can be intuitively understood as a Taylor series of the scalar function $f(t) = \mathbb{E}_{\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{\Sigma}(t))}[G(\boldsymbol{w})]$ where $\boldsymbol{\Sigma}(t) = \gamma_1 \boldsymbol{I}_d + t\gamma_2 \boldsymbol{u}\boldsymbol{u}^\top$. Iteratively applying Price's Theorem

(Price, 1958; McMahon, 1964) to obtain the derivatives of f with respect to t reveals that this infinite sum is exactly the formal Taylor series of $f(t)$ around $t = 0$. This series need not converge at $t = 1$, in particular when the spike γ_2 is large compared to γ_1, so the identity in Theorem 4.1 should be understood whenever the right-hand side converges. Our analysis does not rely on convergence as we truncate the expansion after the first-order term and control the rest through the integral form of the Taylor remainder: $f(1) = f(0) + f'(0) + \int_0^1 (1-t)\, f''(t)\, \mathrm{d}t$ (Lemma 4.4).

4.2 Expansion of the updated kernel k_1

To bring Theorem 4.1 into our context, for a fixed pair $(\boldsymbol{x}, \boldsymbol{x}')$, we define the function $G(\boldsymbol{w}) := G_{\boldsymbol{x},\boldsymbol{x}'}(\boldsymbol{w}) = \sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{w}, \boldsymbol{x}'\rangle)$. Given Remark 4.2, setting $\gamma_1 := \frac{A}{d}$ and $\gamma_2 := \frac{B}{d}$, we apply the theorem to obtain the expansion:

$$k_1(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{1}{d}\boldsymbol{\Gamma})}[G(\boldsymbol{w})] = \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{A}{d}\boldsymbol{I}_d)}[G(\boldsymbol{w})] + \sum_{j=1}^{\infty} \frac{1}{j!}\left(\frac{B}{2d}\right)^j \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{A}{d}\boldsymbol{I}_d)}[D^{(2j)}_{\boldsymbol{w}^*} G(\boldsymbol{w})]\,.$$

In this series, the $j = 0$ term corresponds to the scaled isotropic kernel

$$k_0^{(A)}(\boldsymbol{x}, \boldsymbol{x}') := \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{A}{d}\boldsymbol{I}_d)}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{w}, \boldsymbol{x}'\rangle)]\,,$$

and the $j = 1$ term admits an exact expression by applying Leibniz rule

$$\begin{aligned} S(\boldsymbol{x}, \boldsymbol{x}') := \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{A}{d}\boldsymbol{I}_d)}[D^{(2)}_{\boldsymbol{w}^*} G(\boldsymbol{w})] &= \langle \boldsymbol{w}^*, \boldsymbol{x}'\rangle^2 \mathbb{E}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\sigma^{(2)}(\langle \boldsymbol{w}, \boldsymbol{x}'\rangle)] \\ &\quad + 2\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle\langle \boldsymbol{w}^*, \boldsymbol{x}'\rangle \mathbb{E}[\sigma'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\sigma'(\langle \boldsymbol{w}, \boldsymbol{x}'\rangle)] \\ &\quad + \langle \boldsymbol{w}^*, \boldsymbol{x}\rangle^2 \mathbb{E}[\sigma^{(2)}(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{w}, \boldsymbol{x}'\rangle)]\,. \end{aligned} \tag{4.1}$$

Hence we can isolate the effect of the spike by taking k_1 as a first-order perturbation of $k_0^{(A)}$ with

$$k_1(\boldsymbol{x}, \boldsymbol{x}') = k_0^{(A)}(\boldsymbol{x}, \boldsymbol{x}') + \frac{B}{2d} S(\boldsymbol{x}, \boldsymbol{x}') + R(\boldsymbol{x}, \boldsymbol{x}')\,, \tag{4.2}$$

where $S(\boldsymbol{x}, \boldsymbol{x}')$ is defined by Eq. (4.1), and $R(\boldsymbol{x}, \boldsymbol{x}')$ are the residual terms:

$$R(\boldsymbol{x}, \boldsymbol{x}') := k_1(\boldsymbol{x}, \boldsymbol{x}') - k_0^{(A)}(\boldsymbol{x}, \boldsymbol{x}') - \frac{B}{2d} S(\boldsymbol{x}, \boldsymbol{x}') \tag{4.3}$$

By Assumption 2.1, k_1, $k_0^{(A)}$ and S grow at most polynomially in $\|\boldsymbol{x}\|$ and $\|\boldsymbol{x}'\|$. Since Gaussian moments are finite, they are square-integrable with respect to $\rho_{\mathcal{X}} \otimes \rho_{\mathcal{X}}$ and define Hilbert–Schmidt integral operators on $L^2(\rho_{\mathcal{X}})$; in particular, so does R. The following result bounds the operator norm of the remainder, with the full proof available in Appendix C.3.

Lemma 4.4. Under Assumption 2.1, the function R defined by Eq. (4.3), with bounded integral operator $T_R f(\boldsymbol{x}) = \int R(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}')$, admits the following representation

$$R(\boldsymbol{x}, \boldsymbol{x}') = \left(\frac{B}{2d}\right)^2 \int_0^1 (1-t)\, \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{\Gamma}(t)/d)} \left[D^{(4)}_{\boldsymbol{w}^*} G(\boldsymbol{w})\right] \mathrm{d}t\,,$$

where $\boldsymbol{\Gamma}(t) := A\boldsymbol{I}_d + tB\boldsymbol{w}^*(\boldsymbol{w}^*)^\top$. Furthermore, we have $\|T_R\|_{\mathrm{op}} = \mathcal{O}\left(B^2/d^2\right)$.

5 Example under the ReLU activations

To make the previous discussion on k_1 concrete, we now look closer to the particular case of a ReLU network, giving an explicit characterization of its eigenfunctions as well as numerical illustration. Defining the warped cosine similarity $\gamma_{\boldsymbol{\Gamma}} := \frac{\boldsymbol{x}^\top \boldsymbol{\Gamma} \boldsymbol{x}'}{\sqrt{(\boldsymbol{x}^\top \boldsymbol{\Gamma} \boldsymbol{x})(\boldsymbol{x}'^\top \boldsymbol{\Gamma} \boldsymbol{x}')}}$, we can write k_1 for the ReLU activation explicitly:

$$k_1(\boldsymbol{x}, \boldsymbol{x}') = \frac{\sqrt{(\boldsymbol{x}^\top \boldsymbol{\Gamma} \boldsymbol{x})(\boldsymbol{x}'^\top \boldsymbol{\Gamma} \boldsymbol{x}')}}{2\pi d} \left[\gamma_{\boldsymbol{\Gamma}}(\pi - \arccos(\gamma_{\boldsymbol{\Gamma}})) + \sqrt{1 - \gamma_{\boldsymbol{\Gamma}}^2}\right]\,. \tag{5.1}$$

The calculation leverages the coordinate transformation trick (Liao and Couillet, 2018, Appendix A) and we omit the details here. In the next, we discuss the results from Section 4 in the ReLU case.

5.1 Specialization of the expansion for the ReLU kernel

Next, we consider the expansion of the kernel for the ReLU activation, and determine the first two terms from Eq. (4.2).

When $\sigma(t) = \max(0,t)$, the scaling effect is captured by the identity $k_0^{(A)}(\boldsymbol{x},\boldsymbol{x}') = Ak_0(\boldsymbol{x},\boldsymbol{x}')$. To determine S, if we let $\theta_{\boldsymbol{x},\boldsymbol{x}'}$ be the angle between $\boldsymbol{x}$ and $\boldsymbol{x}'$, we know that $\sigma'' = \delta$ in the distributional sense, thus $\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{A}{d}\boldsymbol{I}_d)}[\sigma''(\langle\boldsymbol{w},\boldsymbol{x}\rangle)\sigma(\langle\boldsymbol{w},\boldsymbol{x}'\rangle)] = \frac{\|\boldsymbol{x}'\|}{\|\boldsymbol{x}\|}\frac{\sin\theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi}$ and similarly $\mathbb{E}[\sigma(\langle\boldsymbol{w},\boldsymbol{x}\rangle)\sigma''(\langle\boldsymbol{w},\boldsymbol{x}'\rangle)] = \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{x}'\|}\frac{\sin\theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi}$. Also, $\sigma'(t) = \mathbf{1}_{\{t\geq 0\}}$ so $\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{A}{d}\boldsymbol{I}_d)}[\sigma'(\langle\boldsymbol{w},\boldsymbol{x}\rangle)\sigma'(\langle\boldsymbol{w},\boldsymbol{x}'\rangle)] = \frac{\pi-\theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi}$.[4]

Therefore, we have the following

$$S(\boldsymbol{x},\boldsymbol{x}') = \frac{\pi-\theta_{\boldsymbol{x},\boldsymbol{x}'}}{\pi}[\langle\boldsymbol{x},\boldsymbol{w}^*\rangle\langle\boldsymbol{x}',\boldsymbol{w}^*\rangle] + \frac{\sin\theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi}\left(\frac{\|\boldsymbol{x}'\|}{\|\boldsymbol{x}\|}\langle\boldsymbol{x},\boldsymbol{w}^*\rangle^2 + \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{x}'\|}\langle\boldsymbol{x}',\boldsymbol{w}^*\rangle^2\right), \quad (5.2)$$

and the ReLU k_1 kernel is given by

$$\begin{aligned} k_1(\boldsymbol{x},\boldsymbol{x}') =& Ak_0(\boldsymbol{x},\boldsymbol{x}') + \frac{B}{2d}\frac{(\pi-\theta_{\boldsymbol{x},\boldsymbol{x}'})}{\pi}[\langle\boldsymbol{x},\boldsymbol{w}^*\rangle\langle\boldsymbol{x}',\boldsymbol{w}^*\rangle] \\ &+ \frac{B}{2d}\frac{\sin\theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi}\left(\frac{\|\boldsymbol{x}'\|}{\|\boldsymbol{x}\|}\langle\boldsymbol{x},\boldsymbol{w}^*\rangle^2 + \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{x}'\|}\langle\boldsymbol{x}',\boldsymbol{w}^*\rangle^2\right) + R(\boldsymbol{x},\boldsymbol{x}')\,. \end{aligned}$$

Since k_0 is the dominant term of the expansion, the following lemma details the spectral basis of T_0, which we will use to approximate the eigenfunctions of the new operator.

Lemma 5.1. The normalized eigenfunctions with nonzero eigenvalues of the integral operator T_0 when $\sigma(t) = \max(0,t)$ are strictly of the form $\psi_{k,m}(\boldsymbol{x}) = \frac{\|\boldsymbol{x}\|}{\sqrt{d}}Y_{k,m}(\boldsymbol{\omega})$, where $\boldsymbol{\omega} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ and $Y_{k,m}$ are the orthonormal spherical harmonics on $\mathbb{S}^{d-1}$.

Approximating the action of the first order kernel S

Lemma 4.4 establishes that the effect of the spike onto the kernel is driven by S. To understand it, we leverage the concentration properties of the Gaussian measure to approximate the action of its integral operator, with proof deferred to Appendix C.5.

Lemma 5.2. Consider the function from Eq. (5.2) with integral operator $T_S : L^2(\rho_{\mathcal{X}}) \to L^2(\rho_{\mathcal{X}})$ defined by $T_S f(\boldsymbol{x}) = \int S(\boldsymbol{x},\boldsymbol{x}')f(\boldsymbol{x}')\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}')$. Let $\{\psi_{k,m}\}$ be the eigenbasis of T_0 and f be a normalized function that can expressed in that basis. If we define the operators $T_S^{(1*)}, T_S^{(2*)} : L^2(\rho_{\mathcal{X}}) \to L^2(\rho_{\mathcal{X}})$ by

$$T_S^{(1*)}f(\boldsymbol{x}) = \frac{\langle\boldsymbol{x},\boldsymbol{w}^*\rangle}{2}\langle\langle\cdot,\boldsymbol{w}^*\rangle, f\rangle \quad \text{and} \quad T_S^{(2*)}f(\boldsymbol{x}) = \frac{1}{2\pi}\left(\frac{\langle\boldsymbol{x},\boldsymbol{w}^*\rangle^2}{\|\boldsymbol{x}\|}\langle\|\cdot\|, f\rangle + \|\boldsymbol{x}\|\left\langle\frac{\langle\cdot,\boldsymbol{w}^*\rangle^2}{\|\cdot\|}, f\right\rangle\right),$$

then, under Assumption 2.1, we have $T_S f(\boldsymbol{x}) = T_S^{(1*)}f(\boldsymbol{x}) + T_S^{(2*)}f(\boldsymbol{x}) + E(\boldsymbol{x})$, such that $\|E\|_{L^2(\rho_{\mathcal{X}})} = o_d(1)$.

These simplified operators can be exclusively defined in terms of the basis from Lemma 5.1, allowing us to track how they combine to shape the functional geometry.

[4] Note that the expectation is always taken under $\mathcal{N}(0,\frac{A}{d}\boldsymbol{I}_d)$, however that does not affect the result since for every $c>0$ we have $\sigma(ct) = c\sigma(t)$, $\sigma'(ct) = \mathbf{1}_{\{ct\geq 0\}} = \sigma'(t)$ and $\delta(ct) = \frac{\delta(t)}{|c|}$.

Emergence of feature learning in the top eigenspaces

Given the explicit form of the ReLU kernel, we can compute the action of the operator T_1 on linear functions, with proof available in Appendix C.6. This reveals a clean orthogonal splitting.

Theorem 5.3 (Feature learning in the linear eigenspace)**.** Under Assumption 2.1, the function $\psi_*(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{w}^* \rangle$ is an eigenfunction of T_1 with eigenvalue $\lambda_{\psi_*}(T_1) = A\lambda_{\psi_*}(T_0) + \frac{B}{4d}$. Furthermore, for any function $\psi_\perp(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{v} \rangle$ such that $\boldsymbol{v} \perp \boldsymbol{w}^*$, we have that $\psi_\perp$ is an eigenfunction of T_1 with eigenvalue $\lambda_{\psi_\perp}(T_1) = A\lambda_{\psi_\perp}(T_0)$.

Remark 5.4. Theorem 5.3 implies that the other eigenfunctions of T_1 must be orthogonal to the linear function. Furthermore, while the top eigenfunction of T_0 is directly related to the constant harmonic Y_0, the terms in S interact with the degree-2 zonal harmonic Y_2. This results in a superposition of Y_0 and Y_2 within the new top eigenspace, as detailed in the following theorem (proof in Appendix C.7).

Theorem 5.5 (Feature learning in the top eigenspace)**.** Let Ψ be the top eigenfunction of the integral operator T_1 with associated eigenvalue $\lambda_{\max}(T_1)$. Also, define $\boldsymbol{\omega} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ and consider the functions given by $Y_0(\boldsymbol{\omega}) = 1$ and $\widehat{Y}_2(\boldsymbol{\omega}, \boldsymbol{w}^*) = \langle \boldsymbol{\omega}, \boldsymbol{w}^* \rangle^2 - \frac{1}{d}$ such that $T_0\Big[\|\boldsymbol{x}\| Y_0(\boldsymbol{\omega})\Big] = \lambda_{\max}(T_0)\Big[\|\boldsymbol{x}\| Y_0(\boldsymbol{\omega})\Big]$ and $T_0\Big[\|\boldsymbol{x}\| \widehat{Y}_2(\boldsymbol{\omega}, \boldsymbol{w}^*)\Big] = \lambda_2(T_0)\Big[\|\boldsymbol{x}\| \widehat{Y}_2(\boldsymbol{\omega}, \boldsymbol{w}^*)\Big]$. We define the approximate eigenvalue and eigenfunction as

$$\widetilde{\lambda} = A\lambda_{\max}(T_0) + \frac{B}{2\pi d} + o_d\left(\frac{|B|}{d}\right), \quad \widetilde{\Psi}(\boldsymbol{x}) = \frac{1}{\sqrt{Z}} \frac{\|\boldsymbol{x}\|}{\sqrt{d}} \Big[Y_0(\boldsymbol{\omega}) + \tau Y_2(\boldsymbol{\omega}, \boldsymbol{w}^*)\Big], \tag{5.3}$$

where Z is a normalization constant, $\tau := \tau(B) = \frac{1}{d}\sqrt{\frac{2d-2}{d+2}}\left[\frac{B}{4\pi(\widetilde{\lambda} - A\lambda_2(T_0))}\right]$ and $Y_2 = \frac{\widehat{Y}_2}{\|\widehat{Y}_2\|}$ is the normalized quadratic zonal harmonic. Then, under Assumption 2.1, we have $T_1\widetilde{\Psi} = \widetilde{\lambda}\widetilde{\Psi} + e$ where $\|e\|_{L^2(\rho_{\mathcal{X}})} = o_d\left(\frac{|B|}{d}\right)$. Also, $\|\Psi - \widetilde{\Psi}\|_{L^2(\rho_{\mathcal{X}})} = o_d\left(\frac{|B|}{d}\right)$ and $|\lambda_{\max}(T_1) - \widetilde{\lambda}| = o_d\left(\frac{|B|}{d}\right)$.

Remark 5.6. Theorems 5.3 and 5.5 show the boost to the linear eigenfunction is of the same order of the boost to the top eigenvalue and of the coefficient of Y_2 in the top eigenfunction. This conforms with the empirical findings demonstrated in Moniri et al. (2023), where the effects of the quadratic and the linear terms are introduced with similar strength on the spectrum of the feature matrix.

5.2 Experiments

In this section we provide experimental results, under Assumption 2.1, with respect to the link function $g(z) = \tanh(2z) + 2\big(e^{-1/2} - \cos z\big)$. We also provide robustness tests of the phenomenon in Appendix D by purposefully violating the assumptions on the link functions, injecting varying levels of noise into the system or changing the input distribution. A link to the code repository along with specific details about implementation are also available at the end of Appendix D.

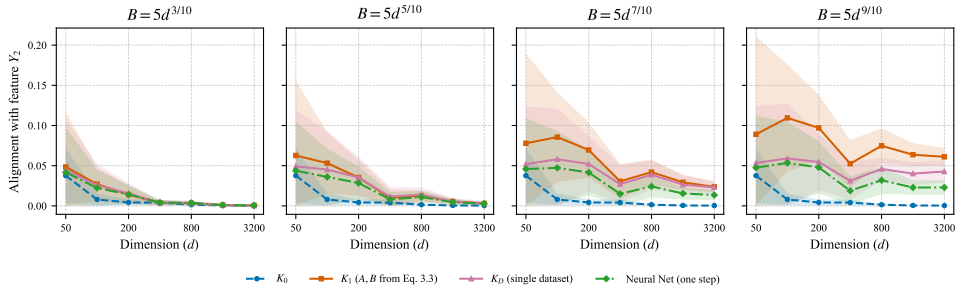


Figure 1: Alignment against d for each B, with the network after one step and kernels $K_{\mathcal{D}}$, K_1.

Alignment (Fig. 1). We compute the squared alignment $\langle \boldsymbol{v}_{\text{top}}, Y_2(\boldsymbol{Z}) \rangle^2$ between the top eigenvector of the kernel matrix on $N = d$ Gaussian points and the quadratic zonal harmonic along $\boldsymbol{w}^*$. For each

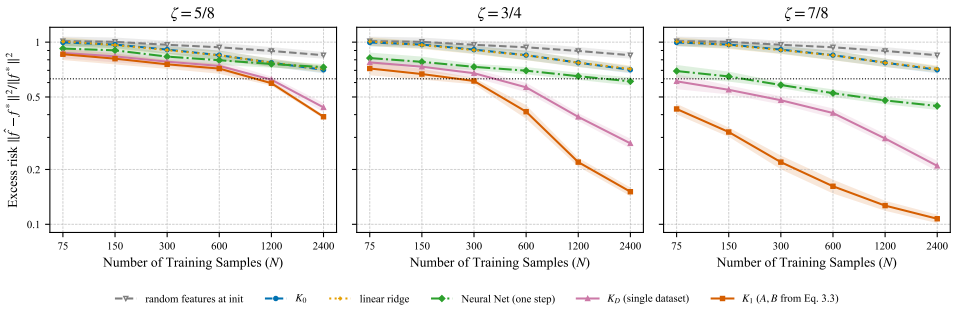


Figure 2: Excess risk against N at $d = 300$, one panel per step size.

spike level $B = 5d^{p/10}$, the network (width 400, $n = 2d$ first-stage samples) takes one gradient step with η chosen so that its idealized kernel k_1 has this B. The alignment of k_1 grows with B, while that of k_0 stays near zero, and the network shows the same growth.

End-to-end risk (Fig. 2). We train the network exactly as in Section 2: width $m = 400$, one full-batch gradient step on the first layer from $n = 8d$ samples with $\eta = 3d^{\zeta}$, followed by ridge regression of the second layer on N fresh samples, and report the excess risk $\|\widehat{f} - f^*\|^2/\|f^*\|^2$ on fresh test points. Each predictor is compared with the reference that isolates the effect of the step: the network with random features at the same initialization $\boldsymbol{W}_0$ (same width, no step), and the kernels $k_{\mathcal{D}}$ (single training set) and k_1 with the initialization kernel k_0. One gradient step lowers the risk of the network below that of its initialization at every N and every step size, and the gain grows with ζ. The same ordering holds for the infinite-width kernels, whose risks satisfy $k_1 < k_{\mathcal{D}} < k_0$. The network has finite width, and it is trained on a single dataset, whereas k_1 is an infinite-width kernel averaged over datasets, which places its spike exactly along $\boldsymbol{w}^*$. Therefore it is not expected to match k_1; the theory predicts the direction of the change, and the trained network follows it.

6 Conclusion

In this work, we characterized how one gradient step reshapes the function space of a two-layer network at the population level: the learned features induce a spiked Gaussian kernel whose deformation, of order B/d, boosts the target-aligned linear eigenvalue and mixes the top radial mode with a target-aligned quadratic harmonic. Our results suggest that working under a distribution shift could be a way to "skip" the early phases of training of these networks.

Limitations. Our analysis is restricted to the Gaussian single-index setting, the ideal structured rank-one update (averaged over the first-stage data) and early steps of training. Thus, the results characterize the functional consequences of the learned feature structure rather than the full finite-sample early training dynamics. Natural next steps include the regime $\eta = \Theta(d)$, where higher-order terms remain relevant, dealing with the residual noise $\boldsymbol{E}$ and quantifying the disparity between our kernel and the kernel of a single trained network without the population average of the data.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 12671631. B.L. was supported by the French government, managed by the National Research Agency (ANR), under the France 2030 program with the project references "ANR-23-IACL-0008" (PR[AI]RIE-PSAI) and "ANR-25-CE23-5660" (MAPLE), as well as the Choose France - CNRS AI Rising Talents program.

References

BA, J., ERDOGDU, M. A., SUZUKI, T., WANG, Z., WU, D. and YANG, G. (2022). High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, vol. 35.

BA, J., ERDOGDU, M. A., SUZUKI, T., WU, D. and ZHANG, T. (2020). Generalization of two-layer neural networks: an asymptotic viewpoint. In *International Conference on Learning Representations*.

BABICHEV, D. and BACH, F. (2018). Slice inverse regression with score functions. *Electronic Journal of Statistics* **12** 1507 – 1543.

BACH, F. (2017a). Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research* **18** 629–681.

BACH, F. (2017b). On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research* **18** 714–751.

BARBIER, J., KRZAKALA, F., MACRIS, N., MIOLANE, L. and ZDEBOROVÁ, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences* **116** 5451–5460.

BARRON, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory* **39** 930–945.

BEN AROUS, G., GHEISSARI, R. and JAGANNATH, A. (2021). Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research* **22** 1–51.

CELENTANO, M., MISIAKIEWICZ, T. and MONTANARI, A. (2021). Minimum complexity interpolation in random features models. *arXiv preprint arXiv:2103.15996* .

CHEN, H., LONG, J. and WU, L. (2025). A duality framework for generalization analysis of random feature models and two-layer neural networks. *Annals of Statistics* .

CHIZAT, L. (2021). Convergence rates of gradient methods for convex optimization in the space of measures. *arXiv preprint arXiv:2105.08368* .

CHIZAT, L., OYALLON, E. and BACH, F. (2019). On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*.

CUI, H., PESCE, L., DANDI, Y., KRZAKALA, F., LU, Y., ZDEBOROVA, L. and LOUREIRO, B. (2024). Asymptotics of feature learning in two-layer networks after one gradient-step. In *Proceedings of the 41st International Conference on Machine Learning* (R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett and F. Berkenkamp, eds.), vol. 235 of *Proceedings of Machine Learning Research*. PMLR.

DAMIAN, A., LEE, J. and SOLTANOLKOTABI, M. (2022). Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*. PMLR.

DAMIAN, A., PILLAUD-VIVIEN, L., LEE, J. D. and BRUNA, J. (2024). The computational complexity of learning gaussian single-index models. *arXiv preprint arXiv:2403.05529* .

DANDI, Y., KRZAKALA, F., LOUREIRO, B., PESCE, L. and STEPHAN, L. (2024). How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research* **25** 1–65.

DANDI, Y., PESCE, L., CUI, H., KRZAKALA, F., LU, Y. and LOUREIRO, B. (2025). A random matrix theory perspective on the spectrum of learned features and asymptotic generalization capabilities. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics* (Y. Li, S. Mandt, S. Agrawal and E. Khan, eds.), vol. 258 of *Proceedings of Machine Learning Research*. PMLR.

DOU, X. and LIANG, T. (2021). Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. *Journal of the American Statistical Association* **116** 1507–1520.

E, W., MA, C. and WU, L. (2021). The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation* 1–38.

FOLLAIN, B. and BACH, F. (2025). Enhanced feature learning via regularisation: Integrating neural networks and kernel methods. *Journal of Machine Learning Research* **26** 1–56.

GHORBANI, B., MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2019). Limitations of lazy training of two-layers neural network. In *NeurIPS*.

HARRIS, C. R., MILLMAN, K. J., VAN DER WALT, S. J., GOMMERS, R., VIRTANEN, P., COURNAPEAU, D., WIESER, E., TAYLOR, J., BERG, S., SMITH, N. J., KERN, R., PICUS, M., HOYER, S., VAN KERKWIJK, M. H., BRETT, M., HALDANE, A., DEL RÍO, J. F., WIEBE, M., PETERSON, P., GÉRARD-MARCHANT, P., SHEPPARD, K., REDDY, T., WECKESSER, W., ABBASI, H., GOHLKE, C. and OLIPHANT, T. E. (2020). Array programming with NumPy. *Nature* **585** 357–362.

HUANG, S., LABARRIÈRE, H., VITO, E. D., POGGIO, T. and ROSASCO, L. (2025). Learning multi-index models with hyper-kernel ridge regression. *arXiv preprint arXiv:2510.02532* .

JACOT, A., GABRIEL, F. and HONGLER, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*.

LEE, J., BAHRI, Y., NOVAK, R., SCHOENHOLZ, S., PENNINGTON, J. and SOHL-DICKSTEIN, J. (2018). Deep neural networks as Gaussian Processes. In *ICLR*.

LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86** 316–327.

LIAO, Z. and COUILLET, R. (2018). On the spectrum of random features maps of high dimensional data. In *International Conference on Machine Learning*.

MCMAHON, E. (1964). An extension of price's theorem (corresp.). *IEEE Transactions on Information Theory* **10** 168–168.

MONIRI, B. and HASSANI, H. (2026). Feature learning in linear-width two-layer networks: Two vs. one step of gradient descent.

MONIRI, B., LEE, D., HASSANI, H. and DOBRIBAN, E. (2023). A theory of non-linear feature learning with one gradient step in two-layer neural networks. *arXiv preprint arXiv:2310.07891* .

PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., KILLEEN, T., LIN, Z., GIMELSHEIN, N., ANTIGA, L., DESMAISON, A., KOPF, A., YANG, E., DEVITO, Z., RAISON, M., TEJANI, A., CHILAMKURTHY, S., STEINER, B., FANG, L., BAI, J. and CHINTALA, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett, eds.), vol. 32. Curran Associates, Inc.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M. and DUCHESNAY, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12** 2825–2830.

PRICE, R. (1958). A useful theorem for nonlinear devices having gaussian inputs. *IRE Transactions on Information Theory* **4** 69–72.

RAHIMI, A. and RECHT, B. (2007). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*.

SCHÖLKOPF, B., HERBRICH, R. and SMOLA, A. J. (2001). A generalized representer theorem. In *International Conference on Computational Learning Theory*. Springer.

SUZUKI, T. (2019). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*.

TROIANI, E., DANDI, Y., DEFILIPPIS, L., ZDEBOROVA, L., LOUREIRO, B. and KRZAKALA, F. (2025). Fundamental computational limits of weak learnability in high-dimensional multi-index models. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, vol. 258 of *Proceedings of Machine Learning Research*. PMLR.

VIRTANEN, P., GOMMERS, R., OLIPHANT, T. E., HABERLAND, M., REDDY, T., COURNAPEAU, D., BUROVSKI, E., PETERSON, P., WECKESSER, W., BRIGHT, J., VAN DER WALT, S. J., BRETT, M., WILSON, J., MILLMAN, K. J., MAYOROV, N., NELSON, A. R. J., JONES, E., KERN, R., LARSON, E., CAREY, C. J., POLAT, İ., FENG, Y., MOORE, E. W., VANDERPLAS, J., LAXALDE, D., PERKTOLD, J., CIMRMAN, R., HENRIKSEN, I., QUINTERO, E. A., HARRIS, C. R., ARCHIBALD, A. M., RIBEIRO, A. H., PEDREGOSA, F., VAN MULBREGT, P. and SCIPY 1.0 CONTRIBUTORS (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17** 261–272.

WANG, Z., WU, D. and FAN, Z. (2024). Nonlinear spiked covariance matrices and signal propagation in deep neural networks. In *Conference on Learning Theory*. PMLR.

WOJTOWYTSCH, S. and E, W. (2020). Can shallow neural networks beat the curse of dimensionality? a mean field training perspective. *IEEE Transactions on Artificial Intelligence* **1** 121–129.

XU, X. and ZHENG, L. (2024). Neural feature learning in function space. *Journal of Machine Learning Research* **25** 1–76.

Contents

A Properties of the Spiked Covariance Matrix **15**

A.1 Derivation of the second moment matrix 15

A.2 Characterization of the coefficient s in the new covariance matrix 16

B Technical Lemmas and their proofs **17**

B.1 Closed form of the infinite sums 17

B.2 Concentration for Gaussian integrals 19

C Proofs of the main results **20**

C.1 Proof of Theorem 3.1 20

C.2 Proof of Theorem 4.1 29

C.3 Proof of Lemma 4.4 30

C.4 Proof of Lemma 5.1 34

C.5 Proof of Lemma 5.2 35

C.6 Proof of Theorem 5.3 40

C.7 Proof of Theorem 5.5 40

D Experimental details **46**

D.1 Different models 46

D.2 Metrics and significance 47

D.3 Alignment experiment 47

D.4 End-to-end experiment 48

D.5 Stress tests 48

D.6 Implementation, Software Stack and Hardware 49

A Properties of the Spiked Covariance Matrix

A.1 Derivation of the second moment matrix

In the random design, we have $y_i = f^*(\boldsymbol{x}_i) + \varepsilon_i$ and $\boldsymbol{x}_i \sim N(0, \boldsymbol{I}_d)$ where ε_i are i.i.d. sub-Gaussian noise random variables with zero mean and variance σ_ε^2. This way $\boldsymbol{\varsigma} = \sum_{i=1}^n y_i \boldsymbol{x}_i$ is random, the random vector $\boldsymbol{z} = \boldsymbol{w} + \frac{\mu_1 \eta}{n\sqrt{m}} a \boldsymbol{\varsigma}$ has mean 0 and covariance

$$\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top] + \frac{\mu_1^2 \eta^2}{n^2 m} \mathbb{E}[a^2 \boldsymbol{\varsigma}\boldsymbol{\varsigma}^\top] \,.$$

Since $a \sim \mathcal{N}(0, \frac{1}{m})$ independently, we have

$$\mathbb{E}[a^2 \boldsymbol{\varsigma}\boldsymbol{\varsigma}^\top] = \frac{1}{m} \mathbb{E}[\boldsymbol{\varsigma}\boldsymbol{\varsigma}^\top] = \frac{1}{m} \Big\{ n \mathbb{E}[y^2 \boldsymbol{x}\boldsymbol{x}^\top] + n(n-1) \mathbb{E}[y\boldsymbol{x}] \mathbb{E}[y\boldsymbol{x}^\top] \Big\} \,.$$

Using that $y = f^*(\boldsymbol{x}) + \varepsilon$ we have

$$\mathbb{E}[y^2 \boldsymbol{x}\boldsymbol{x}^\top] = \mathbb{E}[f^*(\boldsymbol{x})^2 \boldsymbol{x}\boldsymbol{x}^\top] + \sigma_\varepsilon^2 \boldsymbol{I}_d \,.$$

Thus,

$$\mathbb{E}[a^2 \boldsymbol{\varsigma}\boldsymbol{\varsigma}^\top] = \frac{1}{m} \Big(n \Big\{ \mathbb{E}[f^*(\boldsymbol{x})^2 \boldsymbol{x}\boldsymbol{x}^\top] + \sigma_\varepsilon^2 \boldsymbol{I}_d \Big\} + n(n-1) \mathbb{E}[y\boldsymbol{x}] \mathbb{E}[y\boldsymbol{x}^\top] \Big) \,.$$

Hence, the covariance of $\boldsymbol{z}$ is given by the second moment matrix

$$\begin{aligned}
\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] &= \mathbb{E}_{(\boldsymbol{w},\boldsymbol{x},a,\varepsilon)} \left[\left(\boldsymbol{w} + \frac{\mu_1 \eta}{n\sqrt{m}} a \boldsymbol{\varsigma} \right) \left(\boldsymbol{w} + \frac{\mu_1 \eta}{n\sqrt{m}} a \boldsymbol{\varsigma} \right)^\top \right] \\
&= \frac{1}{d} \boldsymbol{I}_d + \frac{\mu_1^2 \eta^2}{n^2 m^2} \Big(n \Big\{ \mathbb{E}[f^*(\boldsymbol{x})^2 \boldsymbol{x}\boldsymbol{x}^\top] + \sigma_\varepsilon^2 \boldsymbol{I}_d \Big\} + n(n-1) \mathbb{E}[y\boldsymbol{x}] \mathbb{E}[y\boldsymbol{x}^\top] \Big) \\
&= \frac{1}{d} \boldsymbol{I}_d + \frac{\mu_1^2 \eta^2}{nm^2} \Big(\Big\{ \mathbb{E}[f^*(\boldsymbol{x})^2 \boldsymbol{x}\boldsymbol{x}^\top] + \sigma_\varepsilon^2 \boldsymbol{I}_d \Big\} + (n-1) \mathbb{E}[y\boldsymbol{x}] \mathbb{E}[y\boldsymbol{x}^\top] \Big) \,.
\end{aligned}$$

Finally, using $n = \alpha d$ and $m = \beta d$, we can rewrite this as

$$\begin{aligned}
&= \frac{1}{d} \boldsymbol{I}_d + \frac{\mu_1^2 \eta^2}{\alpha \beta^2 d^3} \Big(\Big\{ \mathbb{E}[f^*(\boldsymbol{x})^2 \boldsymbol{x}\boldsymbol{x}^\top] + \sigma_\varepsilon^2 \boldsymbol{I}_d \Big\} + (\alpha d - 1) \mathbb{E}[y\boldsymbol{x}] \mathbb{E}[y\boldsymbol{x}^\top] \Big) \\
&= \frac{1}{d} \left(\boldsymbol{I}_d + \frac{\mu_1^2 \eta^2}{\alpha \beta^2 d^2} \mathbb{E}[f^*(\boldsymbol{x})^2 \boldsymbol{x}\boldsymbol{x}^\top] + \frac{\mu_1^2 \eta^2}{\alpha \beta^2 d^2} \sigma_\varepsilon^2 \boldsymbol{I}_d + \frac{\mu_1^2 \eta^2}{\alpha \beta^2 d^2} (\alpha d - 1) \mathbb{E}[y\boldsymbol{x}] \mathbb{E}[y\boldsymbol{x}^\top] \right) \\
&= \frac{1}{d} \left(\boldsymbol{I}_d + \frac{\mu_1^2 \eta^2}{\alpha \beta^2 d^2} \mathbb{E}[f^*(\boldsymbol{x})^2 \boldsymbol{x}\boldsymbol{x}^\top] + \frac{\mu_1^2 \eta^2}{\alpha \beta^2 d^2} \sigma_\varepsilon^2 \boldsymbol{I}_d + \frac{\mu_1^2 \eta^2}{\alpha \beta^2 d} \left(\alpha - \frac{1}{d} \right) \mathbb{E}[y\boldsymbol{x}] \mathbb{E}[y\boldsymbol{x}^\top] \right) \\
&= \frac{1}{d} \left(\boldsymbol{I}_d + \frac{\eta^2 \lambda_{\alpha,\beta}}{d^2} \mathbb{E}[f^*(\boldsymbol{x})^2 \boldsymbol{x}\boldsymbol{x}^\top] + \frac{\eta^2 \lambda_{\alpha,\beta}}{d^2} \sigma_\varepsilon^2 \boldsymbol{I}_d + \frac{\eta^2 \lambda_{\alpha,\beta}}{d} \Big(\alpha - \frac{1}{d} \Big) \mathbb{E}[f^*(\boldsymbol{x})\boldsymbol{x}] \mathbb{E}[f^*(\boldsymbol{x})\boldsymbol{x}^\top] \right) \,,
\end{aligned}$$

where $\lambda_{\alpha,\beta} := \frac{\mu_1^2}{\alpha\beta^2}$ and α, β are the same constants as defined in Eq. (2.2). By using Stein's Lemma the last term simplifies to

$$\mathbb{E}[f^*(\boldsymbol{x})\boldsymbol{x}] \mathbb{E}[f^*(\boldsymbol{x})\boldsymbol{x}^\top] = \mathbb{E}[g'(\langle \boldsymbol{w}^*, \boldsymbol{x} \rangle)]^2 \boldsymbol{w}^* (\boldsymbol{w}^*)^\top = c_1^2 \boldsymbol{w}^* (\boldsymbol{w}^*)^\top \,.$$

We can further rewrite the second term by letting $h(\boldsymbol{x}) = f^*(\boldsymbol{x})^2 \boldsymbol{x} \in \mathbb{R}^d$ such that $J_h \in \mathbb{R}^{d \times d}$, and applying Stein's Lemma:

$$\begin{aligned}
\mathbb{E}[f^*(\boldsymbol{x})^2 \boldsymbol{x}\boldsymbol{x}^\top] &= \mathbb{E}[h(\boldsymbol{x})\boldsymbol{x}^\top] \\
&= \mathbb{E}[J_h(\boldsymbol{x})] = \mathbb{E}[J_{\{f^*(\boldsymbol{x})^2 \boldsymbol{x}\}}] \\
&= \mathbb{E}[\boldsymbol{x} (\nabla f^*(\boldsymbol{x})^2)^\top] + \mathbb{E}[f^*(\boldsymbol{x})^2] \boldsymbol{I}_d \\
&= 2\mathbb{E}[g(\langle \boldsymbol{w}^*, \boldsymbol{x} \rangle) g'(\langle \boldsymbol{w}^*, \boldsymbol{x} \rangle) \boldsymbol{x}] (\boldsymbol{w}^*)^\top + \mathbb{E}[f^*(\boldsymbol{x})^2] \boldsymbol{I}_d.
\end{aligned}$$

In particular, if g is smooth enough, we can apply the same idea to obtain

$$\begin{aligned}\mathbb{E}[f^*(\boldsymbol{x})^2\boldsymbol{x}\boldsymbol{x}^\top] &= 2\mathbb{E}[\nabla f^*(\boldsymbol{x})\nabla f^*(\boldsymbol{x})^\top] + 2\mathbb{E}[f^*(\boldsymbol{x})\boldsymbol{H}_{f^*}(\boldsymbol{x})] + \mathbb{E}[f^*(\boldsymbol{x})^2]\boldsymbol{I}_d \\ &= 2\{\mathbb{E}[g'(\langle\boldsymbol{w}^*,\boldsymbol{x}\rangle)^2] + \mathbb{E}[g(\langle\boldsymbol{w}^*,\boldsymbol{x}\rangle)g''(\langle\boldsymbol{w}^*,\boldsymbol{x}\rangle)]\}\boldsymbol{w}^*(\boldsymbol{w}^*)^\top + \mathbb{E}[f^*(\boldsymbol{x})^2]\boldsymbol{I}_d \\ &:= 2s\boldsymbol{w}^*(\boldsymbol{w}^*)^\top + \mathbb{E}[f^*(\boldsymbol{x})^2]\boldsymbol{I}_d\,,\end{aligned}$$

where $\boldsymbol{H}_{f^*}$ is the Hessian matrix of f^*. Remarkably in this case we have an isotropic term dependent on $\mathbb{E}[f^*(\boldsymbol{x})^2]$, but also an explicit projection onto the one dimensional space created by the target weights $\boldsymbol{w}^*$. Note that even when g is not twice differentiable, the fact that g is Lipschitz and bounded implies g has weak derivatives in the sense of distribution; thus the quantity above is well-defined regardless of the existence of the classical derivatives of g.

As a conclusion, the second moment matrix can be written as

$$\begin{aligned}\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] &= \frac{1}{d}\left(\boldsymbol{I}_d + \frac{\eta^2\lambda_{\alpha,\beta}}{d^2}\left\{2s\boldsymbol{w}^*(\boldsymbol{w}^*)^\top + \mathbb{E}[f^*(\boldsymbol{x})^2]\boldsymbol{I}_d\right\} + \frac{\eta^2\lambda_{\alpha,\beta}}{d^2}\sigma_\varepsilon^2\boldsymbol{I}_d + \frac{\eta^2\lambda_{\alpha,\beta}}{d}\Big(\alpha - \frac{1}{d}\Big)c_1^2\boldsymbol{w}^*(\boldsymbol{w}^*)^\top\right) \\ &= \frac{1}{d}\left(\boldsymbol{I}_d + \frac{\eta^2\lambda_{\alpha,\beta}}{d^2}\mathbb{E}[f^*(\boldsymbol{x})^2]\boldsymbol{I}_d + \frac{\eta^2\lambda_{\alpha,\beta}}{d^2}\sigma_\varepsilon^2\boldsymbol{I}_d + \frac{\eta^2\lambda_{\alpha,\beta}}{d}\Big(\alpha - \frac{1}{d} + \frac{2s}{c_1^2 d}\Big)c_1^2\boldsymbol{w}^*(\boldsymbol{w}^*)^\top\right)\,.\end{aligned}$$

A.2 Characterization of the coefficient s in the new covariance matrix

Remember that, for $h(t) := [g(t)]^2$, we have

$$s = \mathbb{E}[g'(\langle\boldsymbol{w}^*,\boldsymbol{x}\rangle)^2] + \mathbb{E}[g(\langle\boldsymbol{w}^*,\boldsymbol{x}\rangle)g''(\langle\boldsymbol{w}^*,\boldsymbol{x}\rangle)] = \frac{1}{2}\mathbb{E}[h''(\langle\boldsymbol{w}^*,\boldsymbol{x}\rangle)]\,.$$

To characterize the behavior of s, considering that $\|\boldsymbol{w}^*\| = 1$ without loss of generality, we write $Z := \langle\boldsymbol{w}^*,\boldsymbol{x}\rangle \sim \mathcal{N}(0,1)$ and note that for the standard Gaussian measure the following identity holds

$$\mathbb{E}_Z[h''(Z)] = \mathbb{E}_Z[(Z^2-1)h(Z)] = \mathbb{E}_Z[(Z^2-1)g(Z)^2]\,.$$

We have

$$\mathbb{E}_Z[(Z^2-1)g(Z)^2] = \mathbb{E}_Z[Z^2g(Z)^2] - \mathbb{E}_Z[g(Z)^2] = \mathrm{Cov}(Z^2, g(Z)^2)\,.$$

Also note that g is bounded, therefore

$$\mathbb{E}_Z[(Z^2-1)g(Z)^2] \le M_g^2\mathbb{E}[|Z^2-1|] \le C_g\,,$$

where C_g depends exclusively on M_g and not on the dimension d.

Since $g(Z)^2$ is always non-negative, the sign of s depends entirely on the (Z^2-1) weighting factor: if $Z \in (-1,1)$, then (Z^2-1) is negative; if $|Z| > 1$, then (Z^2-1) is positive. Therefore, s will be negative if $g(Z)^2$ concentrates most of its mass inside the interval $(-1,1)$ and decays to zero before the positive regions ($|Z| > 1$) can outweigh it. In other words, $s < 0$ when the link function g is highly localized or "bump-like" near the origin.

To illustrate the behavior of s across different link functions, we discuss some concrete examples.

Cases when $s < 0$: To guarantee a negative s, we need functions that heavily prioritize the interval $(-1,1)$ and ignore the tails.

Indicator function: Let $g(t) = 1$ if $|t| < 1$, and 0 elsewhere. This choice makes $g(t)^2$ be exactly 1 entirely inside the region where (Z^2-1) is negative, and it evaluates to 0 everywhere (Z^2-1) is positive. We have $s = \frac{1}{2}\mathbb{E}[(Z^2-1)g(Z)^2] = \frac{1}{2}\int_{-1}^{1}(t^2-1)\frac{1}{\sqrt{2\pi}}e^{-t^2/2}\mathrm{d}t$. Because the integrand is strictly negative everywhere in this domain, s must be negative.

Gaussian bump: Let $g(t) = \exp(-t^2/2)$. This is a very a localized function. It smoothly peaks at the origin and decays rapidly, heavily weighting the "negative zone" and suppressing the "positive zone. For this case we have $g(t)^2 = \exp(-t^2)$ and evaluating the quantity gives $s = \frac{1}{2}\mathbb{E}[(Z^2-1)\exp(-Z^2)]$. This is equivalent to taking the integral of (t^2-1) against a tighter Gaussian density

$\mathcal{N}(0, 1/3)$. Since the variance is $\frac{1}{3}$, the integral evaluates to something proportional to $(\frac{1}{3} - 1) = -\frac{2}{3}$, yielding a strictly negative s.

Cases when $s > 0$: To guarantee a positive s, we need monotonic functions, or functions that deliberately target the extreme tails of the distribution.

Identity function: Let $g(t) = t$. This is the simplest possible case. We have $h(t) = t^2$ and taking the second derivative $h''(t) = 2$. Using the definition $s = \frac{1}{2}\mathbb{E}[h''(Z)]$, we get $s = \frac{1}{2}\mathbb{E}[2] = 1 > 0$.

Indicator function of the complement: Let $g(t) = 1$ if $|t| > 1$, and 0 elsewhere. This function completely zeroes out the region where $(Z^2 - 1)$ is negative. Computing $s = \frac{1}{2}\mathbb{E}[(Z^2 - 1)g(Z)^2] = \int_1^\infty (t^2 - 1)\frac{1}{\sqrt{2\pi}} e^{-t^2/2} \mathrm{d}t$. Because $(t^2 - 1)$ is strictly positive for all $t > 1$, the integral is definitively positive.

Quadratic function: Let $g(t) = t^2$. Polynomials naturally place massive weight on the tails where numbers grow large, easily overpowering the center. If we compute the quantities necessary $h(t) = t^4$, so $h''(t) = 12t^2$. Evaluating $s = \frac{1}{2}\mathbb{E}[12Z^2] = 6\mathbb{E}[Z^2] = 6 > 0$.

ReLU function: Here we consider the case that g is the ReLU activation, i.e. $g(t) = \max(0, t)$. Let

$$Z = \langle \boldsymbol{w}^*, \boldsymbol{x} \rangle \sim \mathcal{N}(0, \|\boldsymbol{w}^*\|^2)$$

then

$$\mathbb{E}[g(\langle \boldsymbol{w}^*, \boldsymbol{x} \rangle) g''(\langle \boldsymbol{w}^*, \boldsymbol{x} \rangle)] = \mathbb{E}[g(Z)g''(Z)]$$

and the second derivative must be interpreted distributionally. Since g is continuous and $g(0) = 0$, the multiplication of the Dirac distribution by g is well-defined and

$$gg'' = g\delta_0 = g(0)\delta_0 = 0.$$

Therefore,

$$\mathbb{E}[g(Z)g''(Z)] = 0.$$

Moreover $g'(t) = \mathbf{1}_{\{t \geq 0\}}$, so by the symmetry of the centered Gaussian

$$\mathbb{E}[g'(Z)^2] = \mathbb{P}(Z \geq 0) = \frac{1}{2}\,.$$

In the end, combining both results into the definition of s we get

$$s = \mathbb{E}[g'(Z)^2] + \mathbb{E}[g(Z)g''(Z)] = \frac{1}{2}\,.$$

B Technical Lemmas and their proofs

B.1 Closed form of the infinite sums

Lemma B.1. Consider a point $y \in (-\infty, 1/2)$, then for a fixed $i \geq 0$, we have the following identity

$$\sum_{k=i}^{\infty} \binom{2k}{2i} \frac{(2k - 2i - 1)!!}{k!} y^k = \frac{y^i}{i!}(1 - 2y)^{-(i+1/2)}.$$

Proof. First, we check

$$C_{i,k} := \binom{2k}{2i} \frac{(2k - 2i - 1)!!}{k!}.$$

We note that $2k - 2i - 1$ is necessarily odd, thus we can write

$$(2k - 2i - 1)!! = \frac{(2k - 2i)!}{2^{k-i}(k - i)!},$$

and if we expand the binomial coefficient we have

$$\binom{2k}{2i} = \frac{2k!}{2i!(2k-2i)!}.$$

Therefore

$$C_{i,k} = \frac{(2k-2i)!}{2^{k-i}(k-i)!}\frac{2k!}{2i!(2k-2i)!}\frac{1}{k!} = \frac{1}{2^{k-i}(k-i)!}\frac{2k!}{2i!}\frac{1}{k!}.$$

Now, we use the identity $2n! = 2^n(2n-1)!!$ to obtain

$$C_{i,k} = \frac{2^k k!(2k-1)!!}{2^i i!(2i-1)!!}\frac{1}{2^{k-i}(k-i)!}\frac{1}{k!} = \frac{(2k-1)!!}{i!(k-i)!(2i-1)!!},$$

so our goal will be to prove that

$$\frac{y^i}{i!}(1-2y)^{-(i+1/2)} = \sum_{k=i}^{\infty}\frac{(2k-1)!!}{i!(k-i)!(2i-1)!!}y^k \tag{B.1}$$

Next, we study the function

$$(1-2y)^{-(i+1/2)}.$$

Consider the Maclaurin series of the function

$$(1+x)^r = \sum\binom{r}{n}x^r,$$

which is defined for all $|x| < 1$ and real number r.

If we let $x = -2y$ and $r = -(i+1/2)$ we have

$$\begin{aligned}\binom{-(i+1/2)}{n} &= \frac{-(i+1/2)-(i+3/2)\cdots-(i+n-1/2)}{n!}\\ &= (-1)^k\frac{[(2i+1)(2i+3)\ldots(2i+2n-1)]}{2^n n!}\\ &= (-1)^k\frac{(2i+2n-1)!!}{2^n n!(2i-1)!!}.\end{aligned}$$

And substituting everything back

$$\begin{aligned}(1-2y)^{-(i+1/2)} &= \sum_{n=0}^{\infty}(-1)^n\frac{(2i+2n-1)!!}{2^n n!(2i-1)!!}(-2y)^n\\ &= \sum_{n=0}^{\infty}(-1)^{2n}\frac{(2i+2n-1)!!}{2^n n!(2i-1)!!}2^n y^n\\ &= \sum_{n=0}^{\infty}\frac{(2i+2n-1)!!}{n!(2i-1)!!}y^n.\end{aligned}$$

Multiplying the expansion by $\frac{y^i}{i!}$ we get

$$\frac{y^i}{i!}(1-2y)^{-(i+1/2)} = \sum_{n=0}^{\infty}\frac{(2i+2n-1)!!}{i!n!(2i-1)!!}y^{n+i},$$

and if we let $k = n+i$, rearranging the indexes leads to

$$\frac{y^i}{i!}(1-2y)^{-(i+1/2)} = \sum_{k=i}^{\infty}\frac{(2k-1)!!}{i!(k-i)!(2i-1)!!}y^k,$$

which is exactly the expression from Equation B.1. □

B.2 Concentration for Gaussian integrals

Lemma B.2. Consider a fixed $\boldsymbol{x} \in \mathbb{R}^d$. Let $\boldsymbol{x}' \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and define the set

$$\mathcal{A}_\epsilon = \left\{ \boldsymbol{x}' \in \mathbb{R}^d : \left| \theta_{\boldsymbol{x},\boldsymbol{x}'} - \frac{\pi}{2} \right| < \epsilon \right\}, \epsilon \in \left(0, \frac{\pi}{2}\right).$$

Then

$$\rho_{\mathcal{X}}(\mathcal{A}_\epsilon^c) < 3e^{-c_0 d\epsilon^2},$$

and consequently, for any function $g \in L^2(\rho_{\mathcal{X}})$, we have

$$\left| \int_{\mathcal{A}_\epsilon^c} \frac{(\pi - \theta_{\boldsymbol{x},\boldsymbol{x}'})}{\pi} g(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \right| \leq \sqrt{3} e^{-c_1 d\epsilon^2/2} \|g\|_{L^2(\rho_{\mathcal{X}})},$$

and

$$\left| \int_{\mathcal{A}_\epsilon^c} \frac{\sin\theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi} g(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \right| \leq \sqrt{3} e^{-c_2 d\epsilon^2/2} \|g\|_{L^2(\rho_{\mathcal{X}})}.$$

where $c_0, c_1, c_2 > 0$ are absolute constants.

Proof. If we consider the set

$$\mathcal{A}_\epsilon = \left\{ \left| \theta_{\boldsymbol{x},\boldsymbol{x}'} - \frac{\pi}{2} \right| < \epsilon \right\},$$

for a fixed $\boldsymbol{x}$ we define

$$\cos\theta_{\boldsymbol{x},\boldsymbol{x}'} = \frac{\langle \boldsymbol{x}, \boldsymbol{x}' \rangle}{\|\boldsymbol{x}\|\|\boldsymbol{x}'\|} := \frac{Z}{\sqrt{Z^2 + \|\boldsymbol{y}_\perp\|^2}}$$

where $\boldsymbol{y}_\perp$ is a $(d-1)$ dimensional vector orthogonal to $\boldsymbol{x}$. We define the bad event

$$\mathcal{A}_\epsilon^c = \left\{ \left| \theta_{\boldsymbol{x},\boldsymbol{x}'} - \frac{\pi}{2} \right| > \epsilon \right\} = \left\{ |\cos\theta_{\boldsymbol{x},\boldsymbol{x}'}| > \sin\epsilon \right\}.$$

Next we note that

$$|\cos\theta_{\boldsymbol{x},\boldsymbol{x}'}| = \frac{|Z|}{\sqrt{Z^2 + \|\boldsymbol{y}_\perp\|^2}} \leq \frac{|Z|}{\|\boldsymbol{y}_\perp\|}$$

thus

$$\mathbb{P}(\mathcal{A}_\epsilon^c) \leq \mathbb{P}\left(\frac{|Z|}{\|\boldsymbol{y}_\perp\|} \geq \sin\epsilon \right).$$

Splitting the event according to $\|\boldsymbol{y}_\perp\|^2 \geq \frac{d-1}{2}$ we have

$$\mathbb{P}(\mathcal{A}_\epsilon^c) \leq \mathbb{P}\left(|Z| \geq \sin\epsilon \sqrt{\frac{d-1}{2}} \right) + \mathbb{P}\left(\|\boldsymbol{y}_\perp\|^2 < \frac{d-1}{2} \right).$$

Using the standard Gaussian tail bound and the lower-tail concentration bound for the chi-squared distribution, there exists an absolute constant $c > 0$ such that

$$\mathbb{P}(|Z| > t) \leq 2e^{-t^2/2}, \quad \mathbb{P}\left(\|\boldsymbol{y}_\perp\|^2 < \frac{d-1}{2} \right) \leq e^{-c(d-1)}$$

and we have

$$\mathbb{P}(\mathcal{A}_\epsilon^c) \leq 2e^{-\sin(\epsilon)^2(d-1)/4} + e^{-c(d-1)}.$$

Using the bound

$$\sin(\epsilon) \geq \frac{2}{\pi}\epsilon, \quad \forall \epsilon \in [0, \pi/2],$$

and assuming without loss of generality that $\epsilon > 0$ is chosen such that $c > \frac{\epsilon^2}{\pi}$ we can write

$$\mathbb{P}(\mathcal{A}_\epsilon^c) \leq 3e^{-c_0 d\epsilon^2}$$

for some absolute constant $c_0 > 0$.

Given this, we study the integral

$$\int_{\mathcal{A}_\epsilon^c} \frac{(\pi - \theta_{\boldsymbol{x},\boldsymbol{x}'})}{\pi} g(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \,.$$

Noting that $\theta_{\boldsymbol{x},\boldsymbol{x}'} \in [0, \pi]$, we have $\left|\frac{\pi - \theta_{\boldsymbol{x},\boldsymbol{x}'}}{\pi}\right| \leq 1$ and by the Cauchy-Schwarz inequality we can bound the integral over $\mathcal{A}^c$:

$$\left|\int_{\mathcal{A}_\epsilon^c} \frac{(\pi - \theta_{\boldsymbol{x},\boldsymbol{x}'})}{\pi} g(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}')\right| \leq \sqrt{\rho_{\mathcal{X}}(\mathcal{A}_\epsilon^c)} \cdot \sqrt{\int |g(\boldsymbol{x}')|^2 \mathrm{d}\rho_{\mathcal{X}}} \leq \sqrt{3} e^{-c_1 d\epsilon^2/2} \|g\|_{L^2(\rho_{\mathcal{X}})}.$$

Lastly, the result for

$$\left|\int_{\mathcal{A}_\epsilon^c} \frac{\sin\theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi} g(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}')\right|$$

follows from the fact that $|\sin\theta| \leq 1$ and the same argument under the concentration of the measure. □

C Proofs of the main results

C.1 Proof of Theorem 3.1

Proof. We split the proofs into several parts. We will first prove the kernels are close in norm in the product space $L^2(\rho_{\mathcal{X}}) \times L^2(\rho_{\mathcal{X}})$, which leads to the same conclusion about the Hilbert-Schmidt norm and the operator norm of the difference of integral operators. Since these objects live in the $L^2(\rho_{\mathcal{X}})$ space, the inputs are inherently unbounded, thus we separate this case to deal with a bounded set and use the exponential decay of the measure to control the tails.

Lastly, we show the distribution shift identity by introducing a pushforward measure dictated by $\boldsymbol{\Gamma}$ to translate the weights and inputs to the context of isotropic weights.

Ultimately, we want analyze the integral

$$\iint |k_1^*(\boldsymbol{x},\boldsymbol{x}') - k_1(\boldsymbol{x},\boldsymbol{x}')|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}) \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \,.$$

Reduction to a Jensen gap. Recall that $\boldsymbol{z} = \boldsymbol{w} + s_0\, a\, \boldsymbol{\varsigma}$, where $s_0 := \frac{\mu_1 \eta}{n\sqrt{m}}$, and $\boldsymbol{w} \sim \mathcal{N}(0, \frac{1}{d}\boldsymbol{I}_d)$, $a \sim \mathcal{N}(0, \frac{1}{m})$ and $\boldsymbol{\varsigma} = \sum_{i=1}^n y_i \boldsymbol{x}_i$ are mutually independent. Unconditionally, $\boldsymbol{z}$ is not Gaussian in general, but conditionally on $\boldsymbol{\varsigma}$ we have

$$\boldsymbol{z} \mid \boldsymbol{\varsigma} \sim \mathcal{N}\Big(0, \tfrac{1}{d}\boldsymbol{I}_d + \tfrac{s_0^2}{m}\boldsymbol{\varsigma}\boldsymbol{\varsigma}^\top\Big).$$

Fix a pair $(\boldsymbol{x},\boldsymbol{x}')$, let $\widetilde{\boldsymbol{X}} = [\boldsymbol{x},\boldsymbol{x}'] \in \mathbb{R}^{d\times 2}$, and consider the pre-activations $\boldsymbol{v} = \widetilde{\boldsymbol{X}}^\top \boldsymbol{z} = (\langle \boldsymbol{z},\boldsymbol{x}\rangle, \langle \boldsymbol{z},\boldsymbol{x}'\rangle)^\top \in \mathbb{R}^2$. We define

$$\boldsymbol{\Sigma} := \frac{1}{d}\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}} = \frac{1}{d}\begin{bmatrix} \|\boldsymbol{x}\|^2 & \langle \boldsymbol{x},\boldsymbol{x}'\rangle \\ \langle \boldsymbol{x},\boldsymbol{x}'\rangle & \|\boldsymbol{x}'\|^2 \end{bmatrix}, \qquad \boldsymbol{p} := \widetilde{\boldsymbol{X}}^\top \boldsymbol{\varsigma} = \begin{bmatrix} \langle \boldsymbol{\varsigma},\boldsymbol{x}\rangle \\ \langle \boldsymbol{\varsigma},\boldsymbol{x}'\rangle \end{bmatrix},$$

so that, conditionally on $\boldsymbol{\varsigma}$,

$$\boldsymbol{v} \mid \boldsymbol{\varsigma} \sim \mathcal{N}\big(0, \boldsymbol{Q}(\boldsymbol{\varsigma})\big), \qquad \boldsymbol{Q}(\boldsymbol{\varsigma}) := \boldsymbol{\Sigma} + \frac{s_0^2}{m}\boldsymbol{p}\boldsymbol{p}^\top.$$

For a positive semidefinite matrix $\boldsymbol{Q} \in \mathbb{R}^{2\times 2}$ we define

$$\kappa(\boldsymbol{Q}) := \mathbb{E}_{\boldsymbol{h}\sim\mathcal{N}(0,\boldsymbol{Q})}\big[\sigma(h_1)\sigma(h_2)\big].$$

Since $(\boldsymbol{w}, a)$ is independent of $\boldsymbol{\varsigma}$, Fubini's theorem gives

$$k_1^*(\boldsymbol{x},\boldsymbol{x}') = \mathbb{E}_{\boldsymbol{\varsigma}}\big[\kappa(\boldsymbol{Q}(\boldsymbol{\varsigma}))\big].$$

On the other hand, by Eq. (3.3) we have $\frac{1}{d}\boldsymbol{\Gamma} = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \frac{1}{d}\boldsymbol{I}_d + \frac{s_0^2}{m}\mathbb{E}[\boldsymbol{\varsigma}\boldsymbol{\varsigma}^\top]$, so under $\boldsymbol{w}\sim\mathcal{N}(0,\frac{1}{d}\boldsymbol{\Gamma})$ the pre-activations have covariance

$$\bar{\boldsymbol{Q}} := \frac{1}{d}\widetilde{\boldsymbol{X}}^\top\boldsymbol{\Gamma}\widetilde{\boldsymbol{X}} = \boldsymbol{\Sigma} + \frac{s_0^2}{m}\mathbb{E}[\boldsymbol{p}\boldsymbol{p}^\top] = \mathbb{E}_{\boldsymbol{\varsigma}}\big[\boldsymbol{Q}(\boldsymbol{\varsigma})\big],$$

and hence $k_1(\boldsymbol{x},\boldsymbol{x}') = \kappa(\bar{\boldsymbol{Q}})$. Therefore

$$k_1^*(\boldsymbol{x},\boldsymbol{x}') - k_1(\boldsymbol{x},\boldsymbol{x}') = \mathbb{E}_{\boldsymbol{\varsigma}}\big[\kappa(\boldsymbol{Q}(\boldsymbol{\varsigma}))\big] - \kappa\big(\mathbb{E}_{\boldsymbol{\varsigma}}[\boldsymbol{Q}(\boldsymbol{\varsigma})]\big), \tag{C.1}$$

that is, the difference between both kernels is a Jensen gap for the map κ.

Bound of the pointwise difference over a bounded set: To avoid singularities near the origin, we redefine our bounded domain as the set

$$B_{R,c} = \{\boldsymbol{x}\in\mathbb{R}^d : c\sqrt{d} \le \|\boldsymbol{x}\| < R\sqrt{d}\}$$

and we consider the bounded product set $B_{R,c}\times B_{R,c}$. Then, we will provide a truncation argument again inside this set. For $\tau = \frac{\langle \boldsymbol{x},\boldsymbol{x}'\rangle}{\|\boldsymbol{x}\|\|\boldsymbol{x}'\|}$, we can see that

$$\operatorname{Tr}(\boldsymbol{\Sigma}) = \frac{\|\boldsymbol{x}\|^2+\|\boldsymbol{x}'\|^2}{d}, \quad \det\boldsymbol{\Sigma} = \frac{\|\boldsymbol{x}\|^2\|\boldsymbol{x}'\|^2}{d^2}(1-\tau^2).$$

Inside the bounded set we have

$$c^2 \le \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}) \le \lambda_{\max} \le \operatorname{Tr}(\boldsymbol{\Sigma}) \le 2R^2, \quad \lambda_{\min} = \frac{\det\boldsymbol{\Sigma}}{\lambda_{\max}} \ge \frac{c^4(1-\tau^2)}{2R^2}.$$

Now consider the set $\mathcal{S} = \big\{(\boldsymbol{x},\boldsymbol{x}')\in B_{R,c}\times B_{R,c} : |\tau| < \frac{1}{2}\big\}$. In this case, the points are well-separated and $(1-\tau^2)\ge 3/4$ so

$$\lambda_{\min} \ge \frac{3c^4}{8R^2}.$$

Pointwise bound on $\mathcal{S}$. Fix $(\boldsymbol{x},\boldsymbol{x}')\in\mathcal{S}$ and write

$$\boldsymbol{\Delta} := \boldsymbol{Q}(\boldsymbol{\varsigma}) - \bar{\boldsymbol{Q}} = \frac{s_0^2}{m}\big(\boldsymbol{p}\boldsymbol{p}^\top - \mathbb{E}[\boldsymbol{p}\boldsymbol{p}^\top]\big), \qquad \boldsymbol{Q}_s := \bar{\boldsymbol{Q}} + s\boldsymbol{\Delta} = (1-s)\bar{\boldsymbol{Q}} + s\boldsymbol{Q}(\boldsymbol{\varsigma}), \quad s\in[0,1].$$

Both $\bar{\boldsymbol{Q}}$ and $\boldsymbol{Q}(\boldsymbol{\varsigma})$ dominate $\boldsymbol{\Sigma}$ in the Loewner order, hence so does their convex combination, and therefore

$$\boldsymbol{Q}_s \succeq \boldsymbol{\Sigma}, \qquad \|\boldsymbol{Q}_s^{-1}\|_{\mathrm{op}} \le \lambda_{\min}(\boldsymbol{\Sigma})^{-1} \le \frac{8R^2}{3c^4} \qquad \text{for all } s\in[0,1] \text{ and all } \boldsymbol{\varsigma}. \tag{C.2}$$

We stress that this control of the inverse is deterministic and uniform in both s and $\boldsymbol{\varsigma}$.

Let $F(\boldsymbol{h}) := \sigma(h_1)\sigma(h_2)$. By Eq. (C.2) the density of $\mathcal{N}(0,\boldsymbol{Q}_s)$ is smooth in s, and since F has at most polynomial growth, differentiation under the expectation is justified exactly as in the proof of Theorem 4.1. By Price's Theorem, for fixed $\boldsymbol{\varsigma}$,

$$\frac{d}{ds}\kappa(\boldsymbol{Q}_s) = \frac{1}{2}\sum_{i,j}\Delta_{ij}\,\mathbb{E}_{\boldsymbol{h}\sim\mathcal{N}(0,\boldsymbol{Q}_s)}\big[\partial_i\partial_j F(\boldsymbol{h})\big], \qquad \frac{d^2}{ds^2}\kappa(\boldsymbol{Q}_s) = \frac{1}{4}\sum_{i,j,k,l}\Delta_{ij}\Delta_{kl}\,\mathbb{E}_{\boldsymbol{h}\sim\mathcal{N}(0,\boldsymbol{Q}_s)}\big[\partial_i\partial_j\partial_k\partial_l F(\boldsymbol{h})\big],$$

where the derivatives of F are understood in the distributional sense. Taylor's theorem with integral remainder in s gives

$$\kappa(\boldsymbol{Q}(\boldsymbol{\varsigma})) = \kappa(\bar{\boldsymbol{Q}}) + \frac{1}{2}\sum_{i,j}\Delta_{ij}\,\mathbb{E}_{\boldsymbol{h}\sim\mathcal{N}(0,\bar{\boldsymbol{Q}})}\big[\partial_i\partial_j F(\boldsymbol{h})\big] + \int_0^1 (1-s)\,\frac{d^2}{ds^2}\kappa(\boldsymbol{Q}_s)\,ds.$$

The first-order term is linear in $\boldsymbol{\Delta}$ with deterministic coefficients, and $\mathbb{E}_{\boldsymbol{\varsigma}}[\boldsymbol{\Delta}] = 0$, so it vanishes after taking the expectation. By Eq. (C.1) we obtain

$$|k_1^*(\boldsymbol{x},\boldsymbol{x}') - k_1(\boldsymbol{x},\boldsymbol{x}')| \le \frac{1}{2}\,\mathbb{E}_{\boldsymbol{\varsigma}}\Big[\sup_{s\in[0,1]}\Big|\frac{d^2}{ds^2}\kappa(\boldsymbol{Q}_s)\Big|\Big].$$

To bound the second derivative, we apply the same argument as in Lemma 4.4 and taking $n = 4$ and $\boldsymbol{U}_i \in \{\boldsymbol{e}_1, \boldsymbol{e}_2\}$, and using $\sum_{i,j}|\Delta_{ij}| \le 2\|\boldsymbol{\Delta}\|_F$, we get

$$\Big|\frac{d^2}{ds^2}\kappa(\boldsymbol{Q}_s)\Big| \le C_4\,\|\boldsymbol{\Delta}\|_F^2\,\max\{1, \|\boldsymbol{Q}_s^{-1}\|_{\mathrm{op}}\}^8\big(1 + \|\boldsymbol{Q}_s\|_{\mathrm{op}}^3\big).$$

where $C_4 > 0$ depends on L_σ and $\sigma(0)$. Let $K_{\mathcal{S}} := \max\{1, 8R^2/(3c^4)\}$. Since $\lambda_{\max}(\boldsymbol{\Sigma}) \le \mathrm{Tr}(\boldsymbol{\Sigma}) \le 2R^2$ on $B_{R,c}$, we have $\|\boldsymbol{Q}_s\|_{\mathrm{op}} \le \max\{\|\bar{\boldsymbol{Q}}\|_{\mathrm{op}}, \|\boldsymbol{Q}(\boldsymbol{\varsigma})\|_{\mathrm{op}}\} \le 2R^2 + \frac{s_0^2}{m}\big(\mathbb{E}\|\boldsymbol{p}\|^2 + \|\boldsymbol{p}\|^2\big)$, and by Eq. (C.2) and the Cauchy–Schwarz inequality,

$$|k_1^*(\boldsymbol{x},\boldsymbol{x}') - k_1(\boldsymbol{x},\boldsymbol{x}')| \le \frac{C_4 K_{\mathcal{S}}^8}{2}\big(\mathbb{E}_{\boldsymbol{\varsigma}}\|\boldsymbol{\Delta}\|_F^4\big)^{1/2}\Big(\mathbb{E}_{\boldsymbol{\varsigma}}\Big[\Big(1 + \big(2R^2 + \tfrac{s_0^2}{m}(\mathbb{E}\|\boldsymbol{p}\|^2 + \|\boldsymbol{p}\|^2)\big)^3\Big)^2\Big]\Big)^{1/2}. \tag{C.3}$$

Moments of $\boldsymbol{p}$. Write $\boldsymbol{x}^{(1)} = \boldsymbol{x}$, $\boldsymbol{x}^{(2)} = \boldsymbol{x}'$, and decompose

$$p_j = \langle\boldsymbol{\varsigma}, \boldsymbol{x}^{(j)}\rangle = n\langle\boldsymbol{w}^*, \boldsymbol{x}^{(j)}\rangle + R^{(j)}, \qquad R_j := \sum_{i=1}^n\big(y_i\langle\boldsymbol{x}_i, \boldsymbol{x}^{(j)}\rangle - \langle\boldsymbol{w}^*, \boldsymbol{x}^{(j)}\rangle\big).$$

Splitting $y_i = g(\langle\boldsymbol{w}^*, \boldsymbol{x}_i\rangle) + \varepsilon_i$, we analyze the term $g(\langle\boldsymbol{w}^*, \boldsymbol{x}_i\rangle)\langle\boldsymbol{x}_i, \boldsymbol{x}^{(j)}\rangle$ and show it is sub-Gaussian. For fixed $\boldsymbol{x}_j$ and $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$ we have $\langle\boldsymbol{x}_i, \boldsymbol{x}^{(j)}\rangle \sim \mathcal{N}(0, \|\boldsymbol{x}^{(j)}\|^2)$ and its sub-Gaussian with $\|\langle\boldsymbol{x}_i, \boldsymbol{x}^{(j)}\rangle\|_{\psi_2} = \Theta(\|\boldsymbol{x}^{(j)}\|)$. Furthermore g is uniformly bounded by Assumption 2.1, therefore pointwise

$$|g(\langle\boldsymbol{w}^*, \boldsymbol{x}_i\rangle)\langle\boldsymbol{x}_i, \boldsymbol{x}^{(j)}\rangle| \le C_g|\langle\boldsymbol{x}_i, \boldsymbol{x}^{(j)}\rangle|$$

and

$$\|g(\langle\boldsymbol{w}^*, \boldsymbol{x}_i\rangle)\langle\boldsymbol{x}_i, \boldsymbol{x}^{(j)}\rangle\|_{\psi_2} = \mathcal{O}(\|\boldsymbol{x}^{(j)}\|)\,,$$

hence it's also sub-Gaussian. For the noise term ε_i we have

$$\|\varepsilon_i\langle\boldsymbol{x}_i, \boldsymbol{x}^{(j)}\rangle\|_{\psi_1} = \mathcal{O}(\|\varepsilon_i\|_{\psi_2}\|\langle\boldsymbol{x}_i, \boldsymbol{x}^{(j)}\rangle\|_{\psi_2}) = \mathcal{O}(\|\varepsilon_i\|_{\psi_2}\|\boldsymbol{x}^{(j)}\|)\,,$$

and thus sub-exponential. Combining both bounds we have

$$\|y_i\langle\boldsymbol{x}_i, \boldsymbol{x}^{(j)}\rangle\|_{\psi_1} = \mathcal{O}(\|\boldsymbol{x}^{(j)}\|)\,.$$

Hence, since

$$\mathbb{E}[y_i\langle\boldsymbol{x}_i, \boldsymbol{x}^{(j)}\rangle] = \langle\mathbb{E}[y_i\boldsymbol{x}_i], \boldsymbol{x}^{(j)}\rangle = c_1\langle\boldsymbol{w}^*, \boldsymbol{x}^{(j)}\rangle\,,$$

each summand of R_j is centered and sub-exponential with ψ_1-norm $\mathcal{O}(\|\boldsymbol{x}^{(j)}\|)$, and by Bernstein's inequality, for every fixed $q \ge 1$,

$$\|R_j\|_{L^q} \lesssim_q \sqrt{n}\,\|\boldsymbol{x}^{(j)}\|.$$

Define $\Lambda(\boldsymbol{x}, \boldsymbol{x}') := 1 + |\langle\boldsymbol{w}^*, \boldsymbol{x}\rangle| + |\langle\boldsymbol{w}^*, \boldsymbol{x}'\rangle|$. On $B_{R,c}$ we have $\|\boldsymbol{x}_j\| \le R\sqrt{d}$, so, with $n \asymp d$, for $p_j = n\langle\boldsymbol{w}^*, \boldsymbol{x}^{(j)}\rangle + R_j$ we have

$$\begin{aligned}\|p_j\|_{L^q} &\le n|\langle\boldsymbol{w}^*, \boldsymbol{x}^{(j)}\rangle| + \|R_j\|_{L^q}\\ &\lesssim d|\langle\boldsymbol{w}^*, \boldsymbol{x}^{(j)}\rangle| + d\end{aligned}$$

$$\lesssim d\Lambda(\boldsymbol{x}, \boldsymbol{x}')\,.$$

Moreover, we analyze $p_j p_k - \mathbb{E}[p_j p_k]$. The first term can be expanded into

$$p_j p_k = n^2 \langle \boldsymbol{w}^*, \boldsymbol{x}^{(j)} \rangle \langle \boldsymbol{w}^*, \boldsymbol{x}^{(k)} \rangle + R_j R_k + n\langle \boldsymbol{w}^*, \boldsymbol{x}^{(j)} \rangle R_k + n\langle \boldsymbol{w}^*, \boldsymbol{x}^{(k)} \rangle R_j$$

and since $\mathbb{E}[R_j] = \mathbb{E}[R_k] = 0$, we have

$$\mathbb{E}[p_j p_k] = n^2 \langle \boldsymbol{w}^*, \boldsymbol{x}^{(j)} \rangle \langle \boldsymbol{w}^*, \boldsymbol{x}^{(k)} \rangle + \mathbb{E}[R_j R_k]$$

therefore

$$p_j p_k - \mathbb{E}[p_j p_k] = n\langle \boldsymbol{w}^*, \boldsymbol{x}^{(j)} \rangle R_k + n\langle \boldsymbol{w}^*, \boldsymbol{x}^{(k)} \rangle R_j + \big(R_j R_k - \mathbb{E}[R_j R_k]\big)\,.$$

We now estimate each term separately. First,

$$\|n\langle \boldsymbol{w}^*, \boldsymbol{x}^{(j)} \rangle R_k\|_{L^q} \le n|\langle \boldsymbol{w}^*, \boldsymbol{x}^{(j)} \rangle| \|R_k\|_{L^q} \lesssim_q d^2 |\langle \boldsymbol{w}^*, \boldsymbol{x}^{(j)} \rangle|$$

and the same holds for $\|n\langle \boldsymbol{w}^*, \boldsymbol{x}^{(k)} \rangle R_j\|_{L_p}$. Furthermore, by Holder's inequality we have

$$\|R_j R_k\|_{L^q} \le \|R_j\|_{L^{2q}} \|R_k\|_{L^{2q}} \lesssim_q d^2$$

and

$$|\mathbb{E}[R_j R_k]| \le \|R_j\|_{L^2} \|R_k\|_{L^2} \lesssim d^2$$

so

$$\|R_j R_k - \mathbb{E}[R_j R_k]\|_{L^q} \lesssim_q d^2\,.$$

Combining all estimates we have

$$\|p_j p_k - \mathbb{E}[p_j p_k]\|_{L^q} \lesssim_q d^2(1 + |\langle \boldsymbol{w}^*, \boldsymbol{x}^{(j)} \rangle| + |\langle \boldsymbol{w}^*, \boldsymbol{x}^{(k)} \rangle|)$$

Recall that

$$\boldsymbol{\Delta} = \frac{s_0^2}{m}(\boldsymbol{p}\boldsymbol{p}^\top - \mathbb{E}[\boldsymbol{p}\boldsymbol{p}^\top])\,,$$

so entrywise

$$\boldsymbol{\Delta}_{jk} = \frac{s_0^2}{m}(p_j p_k - \mathbb{E}[p_j p_k])\,,$$

and since $s_0^2/m = \mu_1^2 \eta^2/(n^2 m^2) = \Theta(\eta^2/d^4)$, we conclude

$$\|\boldsymbol{\Delta}_{jk}\|_{L^q} \lesssim_q \frac{\eta^2}{d^2}\,\Lambda(\boldsymbol{x}, \boldsymbol{x}'),$$

Finally, since $\|\boldsymbol{p}\|^2 = p_1^2 + p_2^2$, by the L^q estimates from earlier, we have

$$\Big\| \|\boldsymbol{p}\|^2 \Big\|_{L^q} \le \|p_1\|_{L^{2q}}^2 + \|p_2\|_{L^{2q}}^2 \lesssim_q d^2 \Lambda(\boldsymbol{x}, \boldsymbol{x}')$$

therefore

$$\Big\| \frac{s_0^2}{m} \|\boldsymbol{p}\|^2 \Big\|_{L^q} \lesssim_q \frac{\eta^2}{d^2}\,\Lambda(\boldsymbol{x}, \boldsymbol{x}')^2.$$

Plugging these into Eq. (C.3), and using $\eta \le d$ and $\Lambda \ge 1$, we obtain, for all $(\boldsymbol{x}, \boldsymbol{x}') \in S$,

$$|k_1^*(\boldsymbol{x}, \boldsymbol{x}') - k_1(\boldsymbol{x}, \boldsymbol{x}')| \le C_{\mathcal{S}}\,\frac{\eta^4}{d^4}\,\Lambda(\boldsymbol{x}, \boldsymbol{x}')^8, \tag{C.4}$$

where $C_{\mathcal{S}} > 0$ depends only on R, c, $\sigma(0)$, L_σ, $\|g\|_\infty$, $\|\varepsilon\|_{\psi_2}$, μ_1, α and β, and not on d. We note that this argument only uses finite moments of the labels, so no tail assumption on the law of $\boldsymbol{v}$ is required.

Bound of the difference over S: Therefore, since $\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle, \langle \boldsymbol{w}^*, \boldsymbol{x}'\rangle \sim \mathcal{N}(0, \|\boldsymbol{w}^*\|^2)$ with $\|\boldsymbol{w}^*\| = 1$, we have $\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[\Lambda(\boldsymbol{x},\boldsymbol{x}')^{16}] = \mathcal{O}(1)$, and Eq. (C.4) gives

$$\iint_S |k_1(\boldsymbol{x},\boldsymbol{x}') - k_1^*(\boldsymbol{x},\boldsymbol{x}')|^2 \, d\rho(\boldsymbol{x}) \, d\rho(\boldsymbol{x}') \le C_S^2 \frac{\eta^8}{d^8} \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}\left[\Lambda(\boldsymbol{x},\boldsymbol{x}')^{16}\right] = \mathcal{O}\Big(\frac{\eta^8}{d^8}\Big),$$

Bound of the difference outside S and inside $B_{R,c} \times B_{R,c}$: Now, we bound the difference in the set $\{|\tau| > 1/2\}$, where the probability is negligible. First, we establish a crude bound for the full set using the Lipschitz property. Note that

$$\mathbb{E}[\langle \boldsymbol{w}, \boldsymbol{x}\rangle^2] = \frac{1}{d}\boldsymbol{x}^\top \boldsymbol{\Gamma} \boldsymbol{x} \le \lambda_{\max}(\boldsymbol{\Gamma}) \frac{\|\boldsymbol{x}\|^2}{d} \, .$$

From $|\sigma(t)| \le L|t| + |\sigma(0)|$ and Cauchy-Schwarz we have

$$|k(\boldsymbol{x},\boldsymbol{x}')| \le \sqrt{\mathbb{E}\left[(L|\langle \boldsymbol{w}, \boldsymbol{x}\rangle| + |\sigma(0)|)\right]^2} \sqrt{\mathbb{E}\left[(L|\langle \boldsymbol{w}, \boldsymbol{x}'\rangle| + |\sigma(0)|)\right]^2}$$

For each term we have $\mathbb{E}\left[(L|\langle \boldsymbol{w}, \boldsymbol{x}\rangle| + |\sigma(0)|)\right]^2 \le 2\sigma(0)^2 + 2L^2\mathbb{E}[\langle \boldsymbol{w}, \boldsymbol{x}\rangle^2]$, and using $\sqrt{ab} \le \frac{1}{2}(a+b)$ we have

$$|k(\boldsymbol{x},\boldsymbol{x}')| \le 2\sigma(0)^2 + L^2(\mathbb{E}[\langle \boldsymbol{w}, \boldsymbol{x}\rangle^2] + \mathbb{E}[\langle \boldsymbol{w}, \boldsymbol{x}'\rangle^2]) \, .$$

Then the difference inherits the bound by

$$|k_1(\boldsymbol{x},\boldsymbol{x}') - k_1^*(\boldsymbol{x},\boldsymbol{x}')|^2 \le 2|k_1|^2 + 2|k_1^*|^2 \le 4\left(2\sigma(0)^2 + 2L^2(\mathbb{E}[\langle \boldsymbol{w}, \boldsymbol{x}\rangle^2] + \mathbb{E}[\langle \boldsymbol{w}, \boldsymbol{x}'\rangle^2])\right)^2 \, .$$

Using the bound we have for the variance, and the fact that both points are inside the bounded region ($\|\boldsymbol{x}\|^2 \le R^2 d$) we have

$$\mathbb{E}[\langle \boldsymbol{w}, \boldsymbol{x}\rangle^2] \le \lambda_{\max}(\boldsymbol{\Gamma}) \frac{\|\boldsymbol{x}\|^2}{d} \le R^2 \lambda_{\max}(\boldsymbol{\Gamma})$$

and, the same bound holds for $\mathbb{E}[\langle \boldsymbol{w}, \boldsymbol{x}'\rangle^2]$, thus since $\lambda_{\max}(\boldsymbol{\Gamma}) \le A + B\|\boldsymbol{w}^*\|^2 = \mathcal{O}\left(\frac{\eta^2}{d}\right)$, there exists a constant $K > 0$, independent of d, such that

$$\sup_{(\boldsymbol{x},\boldsymbol{x}') \in B_{R,c} \times B_{R,c}} |k_1 - k_1^*|^2 \le 4\left(2\sigma(0)^2 + L^2\lambda_{\max}(\boldsymbol{\Gamma})R^2\right)^2 := M_1(1 + \eta^4/d^2) \, .$$

Now, conditioning on $\boldsymbol{x}'$, we write $\boldsymbol{u} = \frac{\boldsymbol{x}'}{\|\boldsymbol{x}'\|}$ and inside the bounded set $\|\boldsymbol{x}\| \ge c\sqrt{d}$, so $|\tau| > t$ forces $|\langle \boldsymbol{x}, \boldsymbol{u}\rangle| > tc\sqrt{d}$ and the Gaussian tail decay gives the bound

$$\mathbb{P}\Big(\{|\tau| > 1/2\} \cap \{\boldsymbol{x} \in B_{R,c}\}\Big) \le 2\exp\left(-\frac{c^2 d}{8}\right) \, .$$

Therefore, we have

$$\iint_S |k_1(\boldsymbol{x},\boldsymbol{x}') - k_1^*(\boldsymbol{x},\boldsymbol{x}')|^2 \mathrm{d}\rho(\boldsymbol{x}) \mathrm{d}\rho(\boldsymbol{x}') = \mathcal{O}\left(\frac{\eta^8}{d^8}\right) \, ,$$

and

$$\begin{aligned}
\iint_{(B_{R,c} \times B_{R,c}) \cap S^c} |k_1(\boldsymbol{x},\boldsymbol{x}') - k_1^*(\boldsymbol{x},\boldsymbol{x}')|^2 \mathrm{d}\rho(\boldsymbol{x}) \mathrm{d}\rho(\boldsymbol{x}') &= \int \left(\int |k_1(\boldsymbol{x},\boldsymbol{x}') - k_1^*(\boldsymbol{x},\boldsymbol{x}')|^2 \mathrm{d}\rho(\boldsymbol{x})\right) \mathrm{d}\rho(\boldsymbol{x}') \\
&\le K \int \mathbb{P}\Big(\{|\tau| > 1/2\} \cap \{\boldsymbol{x} \in B_{R,c}\}\Big) \mathrm{d}\rho(\boldsymbol{x}') \\
&\le 2\left(1 + \frac{\eta^4}{d^2}\right) \exp\left(-\frac{c^2 d}{8}\right)
\end{aligned}$$

which is $o\left(\frac{\eta^8}{d^8}\right)$ in the asymptotic regime. Combining both together gives

$$\iint_{B_{R,c}\times B_{R,c}} |k_1(\boldsymbol{x},\boldsymbol{x}') - k_1^*(\boldsymbol{x},\boldsymbol{x}')|^2 \mathrm{d}\rho(\boldsymbol{x})\mathrm{d}\rho(\boldsymbol{x}') = \mathcal{O}\left(\frac{\eta^8}{d^8}\right).$$

Bounding the integral over the tails: Since the integration over the bounded set is handled, we analyze the integral over $(B_{R,c}\times B_{R,c})^c$. Because the activation function σ is L_σ-Lipschitz, we have the inequality $|\sigma(t)| \le |\sigma(0)| + L_\sigma|t|$. Consequently, for both kernels, there exist absolute constants $C_1, C_2 > 0$ such that the diagonal grows at most quadratically with the input norm

$$k(\boldsymbol{x},\boldsymbol{x}) \le C_1 + C_2\mathbb{E}_{\boldsymbol{w}}[\langle \boldsymbol{w},\boldsymbol{x}\rangle^2].$$

Because the weights in both kernels share the same covariance matrix $\frac{1}{d}\boldsymbol{\Gamma}$, we have that

$$\mathbb{E}_{\boldsymbol{w}}[\langle \boldsymbol{w},\boldsymbol{x}\rangle^2] = \frac{1}{d}\boldsymbol{x}^\top\boldsymbol{\Gamma}\boldsymbol{x} \le \frac{\lambda_{\max}(\boldsymbol{\Gamma})}{d}\|\boldsymbol{x}\|^2,$$

in both cases. Since $\lambda_{\max}(\boldsymbol{\Gamma}) \le A + B\|\boldsymbol{w}^*\|^2 = \mathcal{O}\left(\frac{\eta^2}{d}\right)$ we have

$$\mathbb{E}_{\boldsymbol{w}}[\langle \boldsymbol{w},\boldsymbol{x}\rangle^2] \le C'\frac{\eta^2\|\boldsymbol{x}\|^2}{d^2}$$

for some absolute constant $C' > 0$. Hence, we can find $C_2' > 0$ such that

$$k(\boldsymbol{x},\boldsymbol{x}) \le C_1 + C_2'\frac{\eta^2\|\boldsymbol{x}\|^2}{d^2}.$$

By the Cauchy-Schwarz inequality, absorbing all constants into $C > 0$, the off-diagonal terms are bounded by

$$|k(\boldsymbol{x},\boldsymbol{x}')| \le \sqrt{k(\boldsymbol{x},\boldsymbol{x})k(\boldsymbol{x}',\boldsymbol{x}')} \le \sqrt{\left(1 + C\frac{\eta^2\|\boldsymbol{x}\|^2}{d^2}\right)\left(1 + C\frac{\eta^2\|\boldsymbol{x}'\|^2}{d^2}\right)}$$

and using that $\sqrt{ab} \le (a+b)/2$, absorbing necessary constants into C again, we have

$$|k(\boldsymbol{x},\boldsymbol{x}')| \le C\left(1 + \frac{\eta^2\|\boldsymbol{x}\|^2}{d^2} + \frac{\eta^2\|\boldsymbol{x}'\|^2}{d^2}\right)$$

and using the algebraic identity $(a-b)^2 \le 2a^2 + 2b^2$, we get

$$|k_1^*(\boldsymbol{x},\boldsymbol{x}') - k_1(\boldsymbol{x},\boldsymbol{x}')|^2 \le 2|k_1^*(\boldsymbol{x},\boldsymbol{x}')|^2 + 2|k_1(\boldsymbol{x},\boldsymbol{x}')|^2$$

and applying our bound gives

$$|k_1^*(\boldsymbol{x},\boldsymbol{x}') - k_1(\boldsymbol{x},\boldsymbol{x}')|^2 \le 4C\left(1 + \frac{\eta^2\|\boldsymbol{x}\|^2}{d^2} + \frac{\eta^2\|\boldsymbol{x}'\|^2}{d^2}\right)^2.$$

Finally, using the identity $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, we have

$$|k_1^*(\boldsymbol{x},\boldsymbol{x}') - k_1(\boldsymbol{x},\boldsymbol{x}')|^2 \le M\left(1 + \frac{\eta^4\|\boldsymbol{x}\|^4}{d^4} + \frac{\eta^4\|\boldsymbol{x}'\|^4}{d^4}\right),$$

where M is a constant depending on L_σ, $\sigma(0)$ and C. Therefore, the integral over the unbounded set is bounded by

$$\iint_{(B_{R,c}\times B_{R,c})^c} |k_1^* - k_1|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \le \iint_{(B_{R,c}\times B_{R,c})^c} M\left(1 + \frac{\eta^4\|\boldsymbol{x}\|^4}{d^4} + \frac{\eta^4\|\boldsymbol{x}'\|^4}{d^4}\right)\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}').$$

Noting that

$$\Big(B_{R,c}\times B_{R,c}\Big)^c \subset \Big\{\Big\{\|\boldsymbol{x}\|>R\sqrt{d}\Big\}\times\mathbb{R}^d\Big\}\cup\Big\{\Big\{\|\boldsymbol{x}\|<c\sqrt{d}\Big\}\times\mathbb{R}^d\Big\} \\ \cup\Big\{\mathbb{R}^d\times\Big\{\|\boldsymbol{x}'\|>R\sqrt{d}\Big\}\Big\}\cup\Big\{\mathbb{R}^d\times\Big\{\|\boldsymbol{x}'\|<c\sqrt{d}\Big\}\Big\}\,,$$

since $\rho_{\mathcal{X}}$ is a probability measure, by symmetry, using the union bound we have that

$$\iint_{(B_{R,c}\times B_{R,c})^c} 1\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \le 2\mathbb{P}\left(\|\boldsymbol{x}\|>R\sqrt{d}\right)+2\mathbb{P}\left(\|\boldsymbol{x}\|<c\sqrt{d}\right)\,.$$

If $\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{I}_d)$, by standard Gaussian concentration we have the following probability bound

$$\mathbb{P}\Big(\|\boldsymbol{x}\|>R\sqrt{d}\Big)=\mathbb{P}\Big(\|\boldsymbol{x}\|>\sqrt{d}+(R-1)\sqrt{d}\Big)\le\exp\left(-\frac{(R-1)^2d}{2}\right)\,.$$

Furthermore, the probability mass of the excluded inner ball is bounded by

$$\mathbb{P}(\|\boldsymbol{x}\|<c\sqrt{d})=\mathbb{P}(\|\boldsymbol{x}\|^2<c^2d)\le\left[c^2e^{1-c^2}\right]^{d/2}\,.$$

For every fixed $0<c<1$, we have

$$c^2e^{1-c^2}<1\,.$$

Defining $y:=c^2e^{1-c^2}$, the bound becomes

$$\mathbb{P}(\|\boldsymbol{x}\|<c\sqrt{d})\le y^{d/2}.$$

Since $y<1$, we have $\ln y<0$, and therefore

$$y^{d/2}=\exp\left(\frac{d}{2}\ln y\right)=\exp(-c'd)\,,$$

where

$$c'\coloneqq-\frac{1}{2}\ln y=\frac{1}{2}\big(c^2-1-2\ln c\big)>0\,.$$

Give this, it suffices to estimate the term

$$\iint_{\left\{\{\|\boldsymbol{x}\|>R\sqrt{d}\}\times\mathbb{R}^d\right\}}\frac{\eta^4}{d^4}\|\boldsymbol{x}\|^4\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}')\,,$$

and by symmetry the other will follow exactly the same. Since the integral does not depend on $\boldsymbol{x}'$, we have

$$\iint_{\left\{\{\|\boldsymbol{x}\|>R\sqrt{d}\}\times\mathbb{R}^d\right\}}\frac{\eta^4}{d^4}\|\boldsymbol{x}\|^4\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}')=\int_{\{\|\boldsymbol{x}\|>R\sqrt{d}\}}\frac{\eta^4}{d^4}\|\boldsymbol{x}\|^4\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\,.$$

By Cauchy Schwarz we have that

$$\int_{\{\|\boldsymbol{x}\|>R\sqrt{d}\}}\frac{\eta^4}{d^4}\|\boldsymbol{x}\|^4\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\le\frac{\eta^4}{d^4}\sqrt{\mathbb{P}\Big(\|\boldsymbol{x}\|>R\sqrt{d}\Big)}\sqrt{\mathbb{E}_{\boldsymbol{x}}[\|\boldsymbol{x}\|^8]}$$

and since $\|\boldsymbol{x}\|^2$ follows a χ^2_d distribution, we have $\mathbb{E}[\|\boldsymbol{x}\|^8]=\mathcal{O}(d^4)$. Substituting this back and using the concentration of the measure again, we get

$$\int_{\{\|\boldsymbol{x}\|>R\sqrt{d}\}}\frac{\eta^4}{d^4}\|\boldsymbol{x}\|^4\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})=\mathcal{O}\left[\frac{\eta^4}{d^2}\exp\left(-\frac{(R-1)^2d}{4}\right)\right]\,.$$

Analogously, we have

$$\int_{\{\|\boldsymbol{x}\|<c\sqrt{d}\}}\frac{\eta^4}{d^4}\|\boldsymbol{x}\|^4\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\le\frac{\eta^4}{d^4}\sqrt{\mathbb{P}\Big(\|\boldsymbol{x}\|<c\sqrt{d}\Big)}\sqrt{\mathbb{E}_{\boldsymbol{x}}[\|\boldsymbol{x}\|^8]}=\mathcal{O}\left[\frac{\eta^4}{d^2}\exp(-c'd/2)\right]\,.$$

Hence, collecting everything, the integral over $\left(B_{R,c} \times B_{R,c}\right)^c$ is bounded by

$$\iint_{(B_{R,c}\times B_{R,c})^c} |k_1^*-k_1|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') = \mathcal{O}\left[\frac{\eta^4}{d^2}\exp\left(-\frac{(R-1)^2 d}{4}\right)\right] + \mathcal{O}\left[\frac{\eta^4}{d^2}\exp(-c'd/2)\right] .$$

Since $\eta = \Theta(d^\zeta)$, with $\zeta \in [1/2, 1)$, this gives the bound

$$\iint_{(B_{R,c}\times B_{R,c})^c} |k_1^* - k_1|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') = \mathcal{O}\left(d^{4\zeta-2}e^{-Ed}\right) .$$

for the constant $E = \min\left\{\frac{(R-1)^2}{4}, \frac{c'}{2}\right\} > 0$, which does not depend on d. Because of the exponential decay, this is strictly smaller than the bound inside $B_{R,c} \times B_{R,c}$.

Final bound on the norm: From the previous discussions we know

$$\iint_{B_{R,c}\times B_{R,c}} |k_1^* - k_1|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') = \mathcal{O}\left(\frac{\eta^4}{d^4}\right)^2$$

and

$$\iint_{(B_{R,c}\times B_{R,c})^c} |k_1^* - k_1|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') = \mathcal{O}\left(d^{4\zeta-2}e^{-Ed}\right) .$$

Writing $\|k_1^* - k_1\|_{L^2(\rho_{\mathcal{X}})\times L^2(\rho_{\mathcal{X}})}$ as the integral of interest and separating into the integration of both sets

$$\iint |k_1^*-k_1|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') = \iint_{B_{R,c}\times B_{R,c}} |k_1^*-k_1|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') + \iint_{(B_{R,c}\times B_{R,c})^c} |k_1^*-k_1|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}')$$

we can combine the bounds over both sets to obtain

$$\|k_1^* - k_1\|_{L^2(\rho_{\mathcal{X}})\times L^2(\rho_{\mathcal{X}})} = \mathcal{O}\left(\frac{\eta^4}{d^4}\right) ,$$

and since

$$\|T_1^* - T_1\|_{\text{op}} \leq \|T_1^* - T_1\|_{\text{HS}} = \|k_1^* - k_1\|_{L^2(\rho_{\mathcal{X}})\times L^2(\rho_{\mathcal{X}})} ,$$

the same bound shows the operators are close in norm as well.

Proof of the distribution shift identity We consider the integral operators $T_0 : L^2(\rho_{\mathcal{X}}) \to L^2(\rho_{\mathcal{X}})$ and $T_1 : L^2(\rho_{\mathcal{X}}) \to L^2(\rho_{\mathcal{X}})$

$$(T_0 f)(\boldsymbol{x}) = \int_{\mathcal{X}} k_0(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \qquad (T_1 f)(\boldsymbol{x}) = \int_{\mathcal{X}} k_1(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}'),$$

and because T_0 and T_1 are trace-class and compact, Mercer's theorem guarantees that both kernels can be diagonalized. For k_0 we have

$$k_0(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{\infty} \xi_i \varphi_i(\boldsymbol{x}) \varphi_i(\boldsymbol{x}')$$

where $\{\varphi_i\}_{i=1}^{\infty}$ is an orthonormal system and $\{\xi_i\}_{i=1}^{\infty}$ is a family of eigenvalues associated with these basis such that $\xi_1 \geq \xi_2 \geq ... > 0$. For k_1 we have a similar result and

$$k_1(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{\infty} \chi_i \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{x}') ,$$

where $\{\psi_i\}_{i=1}^{\infty}$ is an orthonormal system and $\{\chi_i\}_{i=1}^{\infty}$ are the associated non-increasing eigenvalues. In particular, if the kernels are universal, we have that $\overline{\text{span}\{\varphi_i : i \geq 1\}} = \overline{\text{span}\{\psi_i : i \geq 1\}} = L_2(\mathcal{X}, \rho_{\mathcal{X}})$.

Consider the matrix from Eq. (3.3) (without the $\frac{1}{d}$ scaling), written in short form

$$\mathbf{\Gamma} = A\boldsymbol{I}_d + B\boldsymbol{w}^*(\boldsymbol{w}^*)^\top,$$

with $A, B > 0$ for big enough d.

In particular, given $\boldsymbol{w} \sim \mathcal{N}\left(0, \frac{1}{d}\boldsymbol{I}\right)$ and $\langle \boldsymbol{w}, \boldsymbol{x}\rangle$, we know that for $\boldsymbol{z} \sim \mathcal{N}\left(0, \frac{1}{d}\mathbf{\Gamma}\right)$ we can write

$$\langle \boldsymbol{z}, \boldsymbol{x}\rangle = \langle \boldsymbol{w}, \mathbf{\Gamma}^{1/2}\boldsymbol{x}\rangle$$

and the kernels are related by

$$\begin{aligned} k_1(\boldsymbol{x}, \boldsymbol{x}') &= \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}\left(0,\frac{1}{d}\mathbf{\Gamma}\right)}[\sigma(\langle \boldsymbol{z}, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{z}, \boldsymbol{x}'\rangle)] \\ &= \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}\left(0,\frac{1}{d}\boldsymbol{I}_d\right)}[\sigma(\langle \boldsymbol{w}, \mathbf{\Gamma}^{1/2}\boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{w}, \mathbf{\Gamma}^{1/2}\boldsymbol{x}\rangle)] \\ &= k_0(\mathbf{\Gamma}^{1/2}\boldsymbol{x}, \mathbf{\Gamma}^{1/2}\boldsymbol{x}'). \end{aligned}$$

Let us define the measure ν as the pushforward measure $\nu = (\mathbf{\Gamma}^{\frac{1}{2}})_{\#}V$

$$\nu(V) = (\rho_{\mathcal{X}} \circ \mathbf{\Gamma}^{-1/2})(V) = \rho_{\mathcal{X}}(\mathbf{\Gamma}^{-1/2}V),$$

for every measurable set V of $(\mathbf{\Gamma}^{1/2}\mathcal{X})$. From now on, we denote the input space transformed by the $\mathbf{\Gamma}^{1/2}$ matrix as $(\mathbf{\Gamma}^{1/2}\mathcal{X}) := \mathcal{Z} \subset \mathbb{R}^d$.

Consider now, the spaces $L_2(\mathcal{X}, \rho_{\mathcal{X}})$ and $L_2(\mathcal{Z}, \nu)$, and $T_0^\nu : \mathcal{H}_0 \to \mathcal{H}_0$ the integral operator of k_0 w.r.t ν

$$(T_0^\nu f)(\boldsymbol{z}) = \int_{\mathcal{Z}} k_0(\boldsymbol{z}, \boldsymbol{z}')f(\boldsymbol{z}')\mathrm{d}\nu(\boldsymbol{z}'),\ \forall \boldsymbol{z} \in \mathcal{Z}$$

Due to Mercer's decomposition we have that k_0 can be diagonalized into a family of eigenvalues $\{\omega_i\}_{i\in N}$ and eigenfunctions $\{\boldsymbol{e}_i\}_{i\in N}$ which are orthonormal in $L_2(\mathcal{Z}, \nu)$ and

$$k_0(\boldsymbol{z}, \boldsymbol{z}') = \sum_{i\in\mathbb{N}} \omega_i \boldsymbol{e}_i(\boldsymbol{z})\boldsymbol{e}_i(\boldsymbol{z}') \ \text{ w.r.t. } \nu,$$

where $\{\omega_i\}$ is not necessarily the same family as $\{\xi_i\}$, nor $\{\boldsymbol{e}_i\}$ are the same eigenfunctions as $\{\boldsymbol{\varphi}_i\}$.

We define $\boldsymbol{h}_i(\cdot) = \boldsymbol{e}_i \circ \mathbf{\Gamma}^{1/2}(\cdot)$, the integral operator T_1 can be written as

$$\begin{aligned} (T_1 f)(\boldsymbol{x}) &= \int_{\mathcal{X}} k_1(\boldsymbol{x}, \boldsymbol{x}')f(\boldsymbol{x}')\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \\ &= \int_{\mathcal{X}} k_0(\mathbf{\Gamma}^{1/2}\boldsymbol{x}, \mathbf{\Gamma}^{1/2}\boldsymbol{x}')f(\boldsymbol{x}')\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \\ &= \int_{\mathcal{Z}} k_0(\boldsymbol{z}, \boldsymbol{z}')(f \circ \mathbf{\Gamma}^{-1/2})(\boldsymbol{z}')\mathrm{d}\nu(\boldsymbol{z}'), \end{aligned}$$

and if we plug in $f = \boldsymbol{h}_i$ we get

$$\begin{aligned} (T_1 f)(\boldsymbol{x}) &= \int_{\mathcal{Z}} k_0(\boldsymbol{z}, \boldsymbol{z}')(\boldsymbol{h}_i \circ \mathbf{\Gamma}^{-1/2})(\boldsymbol{z}')\mathrm{d}\nu(\boldsymbol{z}') \\ &= \int_{\mathcal{Z}} k_0(\boldsymbol{z}, \boldsymbol{z}')(\boldsymbol{e}_i \circ \mathbf{\Gamma}^{1/2} \circ \mathbf{\Gamma}^{-1/2})(\boldsymbol{z}')\mathrm{d}\nu(\boldsymbol{z}') \\ &= \int_{\mathcal{Z}} k_0(\boldsymbol{z}, \boldsymbol{z}')\boldsymbol{e}_i(\boldsymbol{z}')\mathrm{d}\nu(\boldsymbol{z}') \\ &= \omega_i \boldsymbol{e}_i(\boldsymbol{z}) = \omega_i \boldsymbol{e}_i(\mathbf{\Gamma}^{1/2}\boldsymbol{x}), \end{aligned}$$

so $\{\boldsymbol{e}_i \circ \mathbf{\Gamma}^{1/2}\}$ are eigenfunctions of T_1 with the same eigenvalues from k_0 w.r.t. ν. Also,

$$\int_{\mathcal{X}} \boldsymbol{e}_i(\mathbf{\Gamma}^{1/2}\boldsymbol{x})\boldsymbol{e}_j(\mathbf{\Gamma}^{1/2}\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}) = \int_{\mathcal{Z}} \boldsymbol{e}_i(\boldsymbol{z})\boldsymbol{e}_j(z)\mathrm{d}\nu(\boldsymbol{z}) = \delta_{i,j},$$

which implies $\{\boldsymbol{e}_i \circ \boldsymbol{\Gamma}^{1/2}\}$ are orthonormal in $L_2(\mathcal{X}, \rho_{\mathcal{X}})$. Therefore, we can also diagonalize k_1 such that

$$k_1(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i \in \mathbb{N}} \omega_i \boldsymbol{e}_i(\boldsymbol{\Gamma}^{1/2}\boldsymbol{x}) \boldsymbol{e}_i(\boldsymbol{\Gamma}^{1/2}\boldsymbol{x}') \text{ w.r.t } \rho_{\mathcal{X}}.$$

□

C.2 Proof of Theorem 4.1

Proof. First, we denote $\boldsymbol{\Lambda} := \boldsymbol{I}_d + \frac{\gamma_2}{\gamma_1}\boldsymbol{u}\boldsymbol{u}^\top$. Then by Sherman–Morrison, for $\boldsymbol{\Sigma} = \gamma_1 \boldsymbol{\Lambda} \boldsymbol{I}_d$, we have:

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\gamma_1}\boldsymbol{I}_d - \frac{\gamma_2}{\gamma_1(\gamma_1+\gamma_2)}\boldsymbol{u}\boldsymbol{u}^\top,$$

and, in particular, the determinants are given by

$$\det(\boldsymbol{\Sigma}) = \det(\boldsymbol{\Lambda}) \det(\gamma_1 \boldsymbol{I}_d) = \frac{\gamma_1+\gamma_2}{\gamma_1} \det(\gamma_1 \boldsymbol{I}_d).$$

Therefore, we write the expectation of G using the explicit Gaussian density under covariance $\boldsymbol{\Sigma}$ and use the Taylor expansion of the exponential function to obtain

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[G(\boldsymbol{w})] &= \int G(\boldsymbol{w}) \frac{1}{(2\pi)^{d/2}\sqrt{\det \boldsymbol{\Sigma}}} \exp\left(-\frac{\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{w}}{2}\right) \mathrm{d}\boldsymbol{w} \\
&= \int \sqrt{\frac{\gamma_1}{\gamma_1+\gamma_2}} \exp\left(\frac{\gamma_2}{2\gamma_1(\gamma_1+\gamma_2)}\langle \boldsymbol{u}, \boldsymbol{w}\rangle^2\right) \frac{G(\boldsymbol{w})}{(2\pi)^{d/2}\sqrt{\det \gamma_1 \boldsymbol{I}_d}} e^{-\frac{\|\boldsymbol{w}\|^2}{2\gamma_1}} \mathrm{d}\boldsymbol{w} \\
&= \sqrt{\frac{\gamma_1}{\gamma_1+\gamma_2}} \int \sum_{k=0}^{\infty} \frac{1}{k!}\left(\frac{\gamma_2}{2\gamma_1(\gamma_1+\gamma_2)}\right)^k \langle \boldsymbol{u}, \boldsymbol{w}\rangle^{2k} G(\boldsymbol{w}) \mathrm{d}\rho_{\gamma_1 \boldsymbol{I}_d}(\boldsymbol{w}) \\
&= \sqrt{\frac{\gamma_1}{\gamma_1+\gamma_2}} \sum_{k=0}^{\infty} \frac{1}{k!}\left(\frac{\gamma_2}{2\gamma_1(\gamma_1+\gamma_2)}\right)^k \int \langle \boldsymbol{u}, \boldsymbol{w}\rangle^{2k} G(\boldsymbol{w}) \mathrm{d}\rho_{\gamma_1 \boldsymbol{I}_d}(\boldsymbol{w}) \\
&= \sqrt{\frac{\gamma_1}{\gamma_1+\gamma_2}} \sum_{k=0}^{\infty} \frac{1}{k!}\left(\frac{\gamma_2}{2\gamma_1(\gamma_1+\gamma_2)}\right)^k \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\gamma_1\boldsymbol{I}_d)}[\langle \boldsymbol{u}, \boldsymbol{w}\rangle^{2k} G(\boldsymbol{w})],
\end{aligned}$$

where we used the relationship between the determinants in the second passage.

Next, we justify the application of Fubini's theorem. If we denote $\phi_{\gamma_1}(\boldsymbol{w}) = e^{-\frac{\|\boldsymbol{w}\|^2}{2\gamma_1}}$ and define

$$f(k, \boldsymbol{w}) = \frac{1}{k!}\left(\frac{\gamma_2}{2\gamma_1(\gamma_1+\gamma_2)}\right)^k \langle \boldsymbol{u}, \boldsymbol{w}\rangle^{2k} G(\boldsymbol{w}) \phi_{\gamma_1}(\boldsymbol{w}),$$

and we check its sufficient condition: absolute integrability

$$\int \sum_{k\geq 0} |f(k, \boldsymbol{w})| \mathrm{d}\boldsymbol{w} < \infty.$$

Looking at this sum we can see

$$\begin{aligned}
\sum_{k\geq 0} |f(k, \boldsymbol{w})| \mathrm{d}\boldsymbol{w} &= \sum_{k\geq 0} \frac{1}{k!} \left| \left(\frac{\gamma_2}{2\gamma_1(\gamma_1+\gamma_2)}\right)^k \langle \boldsymbol{u}, \boldsymbol{w}\rangle^{2k} G(\boldsymbol{w}) \phi_{\gamma_1}(\boldsymbol{w}) \right| \\
&\leq \sum_{k\geq 0} \frac{1}{k!} \left| \left(\frac{\gamma_2}{2\gamma_1(\gamma_1+\gamma_2)}\right) \langle \boldsymbol{u}, \boldsymbol{w}\rangle^{2} \right|^k |G(\boldsymbol{w})| |\phi_{\gamma_1}(\boldsymbol{w})| \\
&= \exp\left(\left| \frac{\gamma_2}{2\gamma_1(\gamma_1+\gamma_2)} \right| \langle \boldsymbol{u}, \boldsymbol{w}\rangle^2 \right) |G(\boldsymbol{w})| \phi_{\gamma_1}(\boldsymbol{w})
\end{aligned}$$

$$= \exp\left(\left|\frac{\gamma_2}{2\gamma_1(\gamma_1+\gamma_2)}\right|\langle \boldsymbol{u}, \boldsymbol{w}\rangle^2 - \frac{\|\boldsymbol{w}\|^2}{2\gamma_1}\right)|G(\boldsymbol{w})|.$$

If this exponent is negative, the Gaussian density is able to suppress the polynomial growth of G under the integral. We always have $\langle \boldsymbol{u}, \boldsymbol{w}\rangle \leq \|\boldsymbol{w}\|$ because of the unit vector $\boldsymbol{u}$, which leads to

$$\exp\left(\left|\frac{\gamma_2}{2\gamma_1(\gamma_1+\gamma_2)}\right| - \frac{1}{2\gamma_1}\right) = \exp\left(\frac{|\gamma_2| - (\gamma_1+\gamma_2)}{2\gamma_1(\gamma_1+\gamma_2)}\right).$$

Due to the assumption that $\gamma_1 + \gamma_2 > |\gamma_2|$, the exponent is always negative and consequently

$$\int \sum_{k\geq 0} |f(k, \boldsymbol{w})| \mathrm{d}\boldsymbol{w} < \infty.$$

We now study the terms

$$\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\gamma_1 \boldsymbol{I}_d)}[\langle \boldsymbol{u}, \boldsymbol{w}\rangle^{2k} G(\boldsymbol{w})].$$

By using Stein's Lemma exhaustively we can obtain

$$\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\gamma_1 \boldsymbol{I}_d)}[\langle \boldsymbol{u}, \boldsymbol{w}\rangle^{2k} G(\boldsymbol{w})] = \sum_{n=0}^{k} \gamma_1^{k+n} \binom{2k}{2n} (2k-2n-1)!! \mathbb{E}[D_{\boldsymbol{u}}^{(2n)} G(\boldsymbol{w})],$$

and the original expression is given by

$$\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[G(\boldsymbol{w})] = \sqrt{\frac{\gamma_1}{\gamma_1+\gamma_2}} \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{\gamma_2}{2\gamma_1(\gamma_1+\gamma_2)}\right)^k \left[\sum_{n=0}^{k} \gamma_1^{k+n} \binom{2k}{2n} (2k-2n-1)!! \mathbb{E}[D_{\boldsymbol{u}}^{(2n)} G(\boldsymbol{w})]\right].$$

Since the exchange of the integral and the outer series was justified previously, and for each fixed k the sum over n is finite, the resulting double series is absolutely summable under the same domination bound; hence we may interchange the order of summation.

$$\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[G(\boldsymbol{w})] = \sqrt{\frac{\gamma_1}{\gamma_1+\gamma_2}} \sum_{n=0}^{\infty} \mathbb{E}[D_{\boldsymbol{u}}^{(2n)} G(\boldsymbol{w})] \left[\sum_{k=n}^{\infty} \frac{\gamma_1^n}{k!} \left(\frac{\gamma_2}{2(\gamma_1+\gamma_2)}\right)^k \binom{2k}{2n} (2k-2n-1)!!\right].$$

Now, due to Lemma B.1 we have the closed form expression for each infinite sum given a fixed i, and if we choose $y = \frac{\gamma_2}{2(\gamma_1+\gamma_2)}$:

$$\begin{aligned}\sum_{k=n}^{\infty} \frac{\gamma_1^n}{k!} \left(\frac{\gamma_2}{2\gamma_1(\gamma_1+\gamma_2)}\right)^k \binom{2k}{2n} (2k-2n-1)!! &= \frac{\gamma_1^n}{n!} \left(\frac{\gamma_2}{2(\gamma_1+\gamma_2)}\right)^n \left(\frac{\gamma_1+\gamma_2}{\gamma_1}\right)^{(n+1/2)} \\ &= \frac{1}{n!} \left(\frac{\gamma_2}{2}\right)^n \sqrt{\frac{\gamma_1+\gamma_2}{\gamma_1}}.\end{aligned}$$

Thus, substituting this back into the expansion gives

$$\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[G(\boldsymbol{w})] = \sum_{i=0}^{\infty} \frac{1}{n!} \left(\frac{\gamma_2}{2}\right)^n \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\gamma_1 \boldsymbol{I}_d)}[D_{\boldsymbol{u}}^{(2n)} G(\boldsymbol{w})]$$

and the proof is complete. □

C.3 Proof of Lemma 4.4

Proof. For $t \in [0, 1]$, we define the parameterized family of covariance matrices

$$\boldsymbol{\Gamma}(t) = A\boldsymbol{I}_d + tB\boldsymbol{w}^*(\boldsymbol{w}^*)^\top,$$

and, if we let ϕ be the Gaussian density function induced by $\mathcal{N}(0, \frac{1}{d}\mathbf{\Gamma}(t))$, we define the scalar function $F : [0, 1] \to \mathbb{R}$ given by

$$F(t) = \int G(\boldsymbol{w})\phi(\boldsymbol{w}, t)\mathrm{d}\boldsymbol{w}.$$

Since $\phi \in C^\infty$, for every $n \geq 0$, the n-th derivative of F is well-defined through the distributional property

$$F^{(n)}(t) = \int G(\boldsymbol{w})\frac{\mathrm{d}^n}{\mathrm{d}t^n}[\phi(\boldsymbol{w}, t)]\mathrm{d}\boldsymbol{w},$$

and $F \in C^\infty$. Expanding the Taylor series of F around $t = 0$ and evaluating at 1, Taylor's Theorem guarantees the exact identity

$$F(1) = F(0) + F'(0) + \int_0^1 (1 - t)F''(t)\mathrm{d}t\,.$$

By Price's Theorem (Price, 1958; McMahon, 1964), we have the formal identity

$$F^{(n)}(t) = \left(\frac{B}{2d}\right)^n \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{1}{d}\mathbf{\Gamma}(t))}[D^{(2n)}_{\boldsymbol{w}^*}G(\boldsymbol{w})],$$

therefore

$$F''(t) = \left(\frac{B}{2d}\right)^2 \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{1}{d}\mathbf{\Gamma}(\xi))}[D^{(4)}_{\boldsymbol{w}^*}G(\boldsymbol{w})],$$

and the desired identity holds

$$R(\boldsymbol{x}, \boldsymbol{x}') = \left(\frac{B}{2d}\right)^2 \int_0^1 (1 - t)\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{1}{d}\mathbf{\Gamma}(\xi))}[D^{(4)}_{\boldsymbol{w}^*}G(\boldsymbol{w})]\,.$$

To bound this quantity, if we show the fourth derivative under the expectation is bounded independently of the dimension, the multiplying factor will ensure the asymptotic profile of the result.

For any $\xi \in [0, 1]$ we consider the matrix $\mathbf{\Gamma}(\xi)$ and define the quantities

$$h_1 = \langle \boldsymbol{w}, \boldsymbol{x}\rangle \quad h_2 = \langle \boldsymbol{w}, \boldsymbol{x}'\rangle$$

which are joint Gaussian variables such that $(h_1, h_2) := \boldsymbol{h} \sim \mathcal{N}(0, \boldsymbol{Q}_\xi)$ where

$$\boldsymbol{Q}_\xi = \frac{1}{d}\begin{bmatrix} \boldsymbol{x}^\top\mathbf{\Gamma}(\xi)\boldsymbol{x} & \boldsymbol{x}^\top\mathbf{\Gamma}(\xi)\boldsymbol{x}' \\ \boldsymbol{x}^\top\mathbf{\Gamma}(\xi)\boldsymbol{x}' & (\boldsymbol{x}')^\top\mathbf{\Gamma}(\xi)\boldsymbol{x}' \end{bmatrix}.$$

If $\phi_{\mathbf{\Gamma}(\xi)}(\boldsymbol{w})$ denotes the Gaussian density from the measure $\mathcal{N}(0, \mathbf{\Gamma}(\xi))$, let $\phi(h_1, h_2)$ denote the probability density function of this 2D Gaussian.

Then we can write

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{1}{d}\mathbf{\Gamma}(t))}[D^{(4)}_{\boldsymbol{w}^*}G(\boldsymbol{w})] &= \int D^{(4)}_{\boldsymbol{w}^*}G(\boldsymbol{w})\phi_{\mathbf{\Gamma}(\xi)}(\boldsymbol{w})\mathrm{d}\boldsymbol{w} \\
&= \int D^{(4)}_{\boldsymbol{w}^*}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{w}, \boldsymbol{x}'\rangle)]\phi_{\mathbf{\Gamma}(\xi)}(\boldsymbol{w})\mathrm{d}\boldsymbol{w} \\
&= \iint D^{(4)}_{\boldsymbol{w}^*}[\sigma(h_1)\sigma(h_2)]\phi(h_1, h_2)\mathrm{d}h_1\mathrm{d}h_2.
\end{aligned}$$

Next, by the chain rule, applying the directional derivative $D_{\boldsymbol{w}^*} = \langle \boldsymbol{w}^*, \nabla_{\boldsymbol{w}}\rangle$ yields the two dimensional operator over h_1 and h_2

$$D_{\boldsymbol{w}^*} = \left(\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle\frac{\partial}{\partial h_1} + \langle \boldsymbol{w}^*, \boldsymbol{x}'\rangle\frac{\partial}{\partial h_2}\right) := \langle \boldsymbol{U}, \nabla_{\boldsymbol{h}}\rangle,$$

where we define

$$\boldsymbol{U} = [\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle, \langle \boldsymbol{w}^*, \boldsymbol{x}'\rangle]^\top \in \mathbb{R}^2.$$

Since σ is only Lipschitz, we have no information about its higher order derivatives. Remember that, for any locally integrable f and any smooth compactly supported test function φ,

$$\langle D_{w^*} f, \varphi\rangle = -\langle f, D_{w^*}\varphi\rangle.$$

Iterating,

$$\langle D^n_{w^*} f, \varphi\rangle = (-1)^n \langle f, D^n_{w^*}\varphi\rangle.$$

Hence, we compute the expectation as a 2D integral and, integrating by parts, we transfer the derivative operator to the density function.

$$\iint \left[D^4_{\boldsymbol{w}^*}\Big(\sigma(h_1)\sigma(h_2)\Big)\right]\phi(h_1,h_2)\,\mathrm{d}h_1\mathrm{d}h_2 = \iint \sigma(h_1)\sigma(h_2)\left[D^4_{\boldsymbol{w}^*}\phi(h_1,h_2)\right]\,\mathrm{d}h_1\mathrm{d}h_2.$$

Now, the closed form of the density is given by

$$\phi(h_1,h_2) = \phi(\boldsymbol{h}) = \frac{1}{2\pi\sqrt{\det \boldsymbol{Q}(\xi)}}\exp\left(-\frac{1}{2}\boldsymbol{h}^T\boldsymbol{Q}_\xi^{-1}\boldsymbol{h}\right)$$

and differentiating this with our notation gives

$$\nabla_{\boldsymbol{h}}\phi(h_1,h_2) = \nabla_{\boldsymbol{h}}\phi(\boldsymbol{h}) = \phi(\boldsymbol{h})\nabla_{\boldsymbol{h}}\left(-\frac{1}{2}\boldsymbol{h}^T Q_\xi^{-1}\boldsymbol{h}\right) = -\phi(\boldsymbol{h})\boldsymbol{Q}_\xi^{-1}\boldsymbol{h}$$

which implies the directional derivative is given by

$$D_{\boldsymbol{w}^*}\phi(h_1,h_2) = -\langle \boldsymbol{U}, \boldsymbol{Q}_\xi^{-1}\boldsymbol{h}\rangle\phi(\boldsymbol{h}).$$

Therefore, iterating through this 4 times we obtain

$$D^4_{\boldsymbol{w}^*}\phi(h_1,h_2) = \left(\langle \boldsymbol{U}, \boldsymbol{Q}_\xi^{-1}\boldsymbol{h}\rangle^4 - 6\langle \boldsymbol{U}, \boldsymbol{Q}_\xi^{-1}\boldsymbol{h}\rangle^2\langle \boldsymbol{U}, \boldsymbol{Q}_\xi^{-1}\boldsymbol{U}\rangle + 3\langle \boldsymbol{U}, \boldsymbol{Q}_\xi^{-1}\boldsymbol{U}\rangle^2\right)\phi(\boldsymbol{h}),$$

and, if we denote the polynomial multiplying ϕ by $\mathrm{P}_4(\boldsymbol{U},\boldsymbol{h},\boldsymbol{Q}_\xi)$, we can see that

$$|\mathrm{P}_4(\boldsymbol{U},\boldsymbol{h},\boldsymbol{Q}_\xi)| \le \|\boldsymbol{U}\|^4\left(\|\boldsymbol{Q}_\xi^{-1}\|^4_{\mathrm{op}}\|\boldsymbol{h}\|^4 + 6\|\boldsymbol{Q}_\xi^{-1}\|^3_{\mathrm{op}}\|\boldsymbol{h}\|^2 + 3\|\boldsymbol{Q}_\xi^{-1}\|^2_{\mathrm{op}}\right).$$

Using that σ is L_σ-Lipschitz we have,

$$|\sigma(t)| \le M_{L_\sigma}(1+|t|)$$

where $M_{L_\sigma} = \max(|\sigma(0)|, L_\sigma)$, therefore the product is bounded by

$$|\sigma(h_1)\sigma(h_2)| \le M^2_{L_\sigma}(1+|h_1|)(1+|h_2|) \le M^2_{L_\sigma}(2+\|\boldsymbol{h}\|^2).$$

Thus, the absolute value of R can be bounded with

$$\begin{aligned}
|R(\boldsymbol{x},\boldsymbol{x}')| &\le \frac{B^2}{8d^2}\iint |\sigma(h_1)\sigma(h_2)||\mathrm{P}_4(\boldsymbol{U},\boldsymbol{h},\boldsymbol{Q}_\xi)|\phi(h_1,h_2)\,\mathrm{d}h_1\mathrm{d}h_2 \\
&\le \frac{B^2}{8d^2}\iint M^2_{L_\sigma}(2+\|\boldsymbol{h}\|^2)|\mathrm{P}_4(\boldsymbol{U},\boldsymbol{h},\boldsymbol{Q}_\xi)|\phi(h_1,h_2)\,\mathrm{d}h_1\mathrm{d}h_2 \\
&\le \frac{B^2}{8d^2}P^*(\boldsymbol{U},\boldsymbol{Q}_\xi),
\end{aligned}$$

where we define

$$\begin{aligned}
P^*(\boldsymbol{U},\boldsymbol{Q}_\xi) &= M_L^2\|\boldsymbol{U}\|^4\cdot\mathbb{E}_{\boldsymbol{h}}\left[(2+\|\boldsymbol{h}\|^2)(\|\boldsymbol{Q}_\xi^{-1}\|^4_{\mathrm{op}}\|\boldsymbol{h}\|^4 + 6\|\boldsymbol{Q}_\xi^{-1}\|^3_{\mathrm{op}}\|\boldsymbol{h}\|^2 + 3\|\boldsymbol{Q}_\xi^{-1}\|^2_{\mathrm{op}})\right] \\
&= M_L^2\|\boldsymbol{U}\|^4\cdot\mathbb{E}_{\boldsymbol{h}}\Big[\|\boldsymbol{Q}_\xi^{-1}\|^4_{\mathrm{op}}\|\boldsymbol{h}\|^6 + (2\|\boldsymbol{Q}_\xi^{-1}\|^4_{\mathrm{op}} + 6\|\boldsymbol{Q}_\xi^{-1}\|^3_{\mathrm{op}})\|\boldsymbol{h}\|^4
\end{aligned}$$

$$+ (3\|\boldsymbol{Q}_\xi^{-1}\|_{\text{op}}^2 + 12\|\boldsymbol{Q}_\xi^{-1}\|_{\text{op}}^3)\|\boldsymbol{h}\|^2 + 6\|\boldsymbol{Q}_\xi^{-1}\|_{\text{op}}^2\Big].$$

Next, we note that since $\boldsymbol{h} \sim \mathcal{N}(0, \boldsymbol{Q}_\xi)$ we know

$$\mathbb{E}[\|\boldsymbol{h}\|^{2p}] \leq (2p-1)!! \cdot (2\|\boldsymbol{Q}_\xi\|_{op})^p,$$

therefore distributing the expectation operator and bounding every moment of $\|\boldsymbol{h}\|$ we have

$$\begin{aligned} P^*(\boldsymbol{U}, \boldsymbol{Q}_\xi) \leq M_{L_\sigma}^2 \|\boldsymbol{U}\|^4 \cdot \Big[&(120\|\boldsymbol{Q}_\xi^{-1}\|_{\text{op}}^4\|\boldsymbol{Q}_\xi\|_{\text{op}}^3 + 12(2\|\boldsymbol{Q}_\xi^{-1}\|_{\text{op}}^4\| + 6\|\boldsymbol{Q}_\xi^{-1}\|_{\text{op}}^3)\|\boldsymbol{Q}_\xi\|_{\text{op}}^2 \\ &+ 2(2\|\boldsymbol{Q}_\xi^{-1}\|_{\text{op}}^2 + 6\|\boldsymbol{Q}_\xi^{-1}\|_{\text{op}}^3)\|\boldsymbol{Q}_\xi\|_{\text{op}} + 4\|\boldsymbol{Q}_\xi^{-1}\|_{\text{op}}^2\Big]. \end{aligned} \tag{C.5}$$

To understand the operator norm of T_R we bound its HS norm. Since the absolute value of R is bounded by P^*, we have

$$\|T_R\|_{\text{HS}}^2 \leq \frac{B^4}{64d^4}\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[P^*(\boldsymbol{U}, \boldsymbol{Q}_\xi)^2]\,.$$

To bound this value, we first note that, since $\boldsymbol{w}^*$ is a unit vector,

$$\|\boldsymbol{U}\|^2 = \langle \boldsymbol{w}^*, \boldsymbol{x}\rangle^2 + \langle \boldsymbol{w}^*, \boldsymbol{x}'\rangle^2 := z_1^2 + z_2^2$$

where $z_1, z_2 \sim \mathcal{N}(0,1)$ are under i.i.d. Gaussian input distribution and therefore

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[\|\boldsymbol{U}\|^k] = \mathbb{E}[(z_1^2 + z_2^2)^{k/2}] = \mathcal{O}(1),$$

because the expectation simplifies to a sum of one dimensional Gaussian moments, which are independent of the dimension d.

Next, we bound the moments of operator norm $\boldsymbol{Q}_\xi$ and its inverse. If we define $\widetilde{\boldsymbol{X}} = [\boldsymbol{x}, \boldsymbol{x}'] \in \mathbb{R}^{d\times 2}$ and let $\widetilde{\boldsymbol{W}} = \widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{X}}$, we have that

$$\boldsymbol{Q}_\xi = \frac{1}{d}\widetilde{\boldsymbol{X}}^\top\boldsymbol{\Gamma}(\xi)\widetilde{\boldsymbol{X}}$$

and since $\boldsymbol{\Gamma}(\xi)$ is PSD and $A + \xi B > 0$, the eigenvalues of this matrix must be strictly positive. Let $\gamma_{\min} = A$ denote the minimum eigenvalue of $\boldsymbol{\Gamma}(\xi)$. Then, for the inverse matrix we have

$$\boldsymbol{Q}_\xi^{-1} \preceq \frac{d}{\gamma_{\min}}\widetilde{\boldsymbol{W}}^{-1}.$$

which implies the following bounds on the moments

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}\left[\|\boldsymbol{Q}_\xi^{-1}\|_{\text{op}}^k\right] \leq \left(\frac{d}{\gamma_{\min}}\right)^k \mathbb{E}_{\boldsymbol{W}}\left[\|\widetilde{\boldsymbol{W}}^{-1}\|_{\text{op}}^k\right].$$

Because $\widetilde{\boldsymbol{X}}$ is a $d \times 2$ matrix of independent $\mathcal{N}(0,1)$ entries, the inner product matrix follows a Wishart distribution: $\widetilde{\boldsymbol{W}} \sim \mathcal{W}_2(\boldsymbol{I}_2, d)$ and, consequently, its inverse follows an Inverse-Wishart distribution $\widetilde{\boldsymbol{W}}^{-1} \sim \mathcal{W}_2^{-1}(\boldsymbol{I}_2, d)$.

For an $n \times n$ Inverse-Wishart matrix with d degrees of freedom, the k-th moments are strictly finite and integrable as long as the degrees of freedom exceed the matrix size, i.e. $d > n + 2k - 1$. In our case, $\widetilde{\boldsymbol{W}}$ is 2×2, therefore we have

$$\mathbb{E}_{\widetilde{\boldsymbol{W}}}\left[\|\widetilde{\boldsymbol{W}}^{-1}\|_{\text{op}}^k\right] = \mathcal{O}\left(\frac{1}{d^k}\right)$$

as long as $d > 2k + 1$.

As a consequence, since A does not depend on d, for a fixed k, as the dimension grows, the operator norm is bounded independently of d, i.e.

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}\left[\|\boldsymbol{Q}_\xi^{-1}\|_{\text{op}}^k\right] = \mathcal{O}(1)\,.$$

To bound the non-inverse moments, we decompose the quadratic form directly

$$\boldsymbol{Q}_\xi = \frac{1}{d}\widetilde{\boldsymbol{X}}^\top(A\boldsymbol{I}_d + \xi B\boldsymbol{w}^*(\boldsymbol{w}^*)^\top)\widetilde{\boldsymbol{X}} = \frac{A}{d}\widetilde{\boldsymbol{W}} + \frac{\xi B}{d}\boldsymbol{z}\boldsymbol{z}^\top$$

where $\boldsymbol{z} = \widetilde{\boldsymbol{X}}^\top\boldsymbol{w}^* \in \mathbb{R}^2$. Because the columns of $\widetilde{\boldsymbol{X}}$ are standard Gaussian and $\boldsymbol{w}^*$ is a deterministic unit vector, the projection $\boldsymbol{z}$ is distributed as a standard 2-dimensional Gaussian, $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_2)$.

Applying the triangle inequality to the operator norm we obtain

$$\|\boldsymbol{Q}_\xi\|_{\text{op}} \le \frac{A}{d}\|\widetilde{\boldsymbol{W}}\|_{\text{op}} + \frac{B}{d}\|\boldsymbol{z}\boldsymbol{z}^\top\|_{\text{op}} = \frac{A}{d}\|\widetilde{\boldsymbol{W}}\|_{\text{op}} + \frac{B}{d}\|\boldsymbol{z}\|_2^2\,.$$

Raising the norm to the k-th power and applying the inequality $(a+b)^k \le 2^{k-1}(a^k + b^k)$ yields

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[\|\boldsymbol{Q}_\xi\|_{\text{op}}^k] \le 2^{k-1}\left(\frac{A^k}{d^k}\mathbb{E}[\|\widetilde{\boldsymbol{W}}\|_{\text{op}}^k] + \left(\frac{B}{d}\right)^k\mathbb{E}[\|\boldsymbol{z}\|_2^{2k}]\right)$$

Given $B = \Theta(\eta^2/d)$, with $\eta = \Theta(d^\zeta)$ and $\zeta \in [1/2, 1)$, we have $B = \Theta(d^{2\zeta-1})$ and the coefficient scales as $\frac{B}{d} = \Theta(d^{2\zeta-2})$. Because $\zeta < 1$ by assumption, it follows that $2\zeta - 2 < 0$ and $\frac{B}{d} = o_d(1)$. Furthermore, because $\|\boldsymbol{z}\|_2^2$ follows a chi-squared distribution with 2 degrees of freedom, its moments $\mathbb{E}[\|\boldsymbol{z}\|_2^{2k}]$ are constants dependent only on k, bounded by $\mathcal{O}(1)$. Hence, we also have

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}\left[\|\boldsymbol{Q}_\xi\|_{\text{op}}^k\right] = \mathcal{O}(1)\,.$$

Furthermore, for fixed exponents $k_1, k_2 \ge 0$, by Cauchy-Schwarz we are able to obtain

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}\left[\|\boldsymbol{Q}_\xi^{-1}\|_{\text{op}}^{k_1}\|\boldsymbol{Q}_\xi\|_{\text{op}}^{k_2}\right] = \mathcal{O}(1).$$

Lastly, notice that if we square the bound in C.5, we obtain a combination of powers of $\|\boldsymbol{U}\|$, $\|\boldsymbol{Q}_\xi^{-1}\|_{\text{op}}$ and $\|\boldsymbol{Q}_\xi\|_{\text{op}}$. Applying Holder's inequality to decouple the expectations on each of the terms of the inequality and collecting all the bounds we have on the moments of these quantities we get

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[P^*(\boldsymbol{U}, \boldsymbol{Q}_\xi)^2] = \mathcal{O}(1).$$

Finally, since all the bounds hold uniformly over $\xi \in [0, 1]$, we have the following result for the HS norm

$$\begin{aligned}
\|T_R\|_{\text{HS}}^2 &= \iint |R(\boldsymbol{x}, \boldsymbol{x}')|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \\
&\le \frac{B^4}{64d^4}\int_0^1 (1-t)\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[P^*(\boldsymbol{U}, \boldsymbol{Q}_t)^2]\mathrm{d}t \\
&= \mathcal{O}\left(\frac{B^4}{d^4}\right)
\end{aligned}$$

which implies

$$\|T_R\|_{\text{op}} \le \|T_R\|_{\text{HS}} = \mathcal{O}\left(\frac{B^2}{d^2}\right)$$

and the proof is completed. □

C.4 Proof of Lemma 5.1

Proof. We write the ReLU kernel as its closed-form

$$k_0(\boldsymbol{x}, \boldsymbol{x}') = \frac{\|\boldsymbol{x}\|\|\boldsymbol{x}'\|}{2\pi d}\left[\gamma(\pi - \arccos(\gamma)) + \sqrt{1-\gamma^2}\right] := \frac{rr'}{2\pi d}J_1(\gamma)$$

where $r = \|\boldsymbol{x}\|$ and $r' = \|\boldsymbol{x}'\|$ and J_1 is the standard first-order arc cosine kernel.

In this context, the function $J_1(\gamma)$ is a dot-product kernel on the unit sphere and, by the Funk-Hencke's theorem, J_1 can be expanded into the spherical harmonics basis. Therefore, if we denote $\boldsymbol{\omega} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ and $\boldsymbol{\omega}' = \frac{\boldsymbol{x}'}{\|\boldsymbol{x}'\|}$, we have

$$k_0(\boldsymbol{x}, \boldsymbol{x}') = \frac{rr'}{2\pi d} \sum_{k=0}^{\infty} \lambda_k \sum_m Y_{k,m}(\boldsymbol{\omega}) Y_{k,m}(\boldsymbol{\omega}').$$

By moving the r and r' inside the sum, we get the exact Mercer expansion over the Gaussian space

$$k_0(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{2\pi} \sum_{k=0}^{\infty} \sum_m \lambda_k \left(\frac{r}{\sqrt{d}} Y_{k,m}(\boldsymbol{\omega}) \right) \left(\frac{r'}{\sqrt{d}} Y_{k,m}(\boldsymbol{\omega}') \right).$$

Note that these functions are indeed normalized and pairwise orthogonal since the inner product of two elements of this family can always be separated into an integral acting only on the radii and an integral acting only on the angles. Because the spherical harmonics are pairwise orthogonal, the angular integral will enforce the orthogonality conditions.

Therefore, since this expansion fully constructs the kernel, the only eigenfunctions with non-zero eigenvalues are of the form $\frac{r}{\sqrt{d}} Y_{k,m}(\boldsymbol{\omega})$. □

C.5 Proof of Lemma 5.2

Proof. For this proof we write

$$\begin{aligned} T_1 f(\boldsymbol{x}) &= A T_0 f(\boldsymbol{x}) + \frac{B}{2d} T_S f(\boldsymbol{x}) + T_R f(\boldsymbol{x}) \\ &= A T_0 f(\boldsymbol{x}) + \frac{B}{2d} \left(T_{(1*)} f(\boldsymbol{x}) + T_{(2*)} f(\boldsymbol{x}) \right) + T_R f(\boldsymbol{x}) \end{aligned}$$

and, as we have seen by Lemma 4.4, $\|T_R f\|_{L^2(\rho_{\mathcal{X}})} = o_d\left(\frac{|B|}{d}\right)$. Thus, we will study the terms $T_{(1*)} f$ and $T_{(2*)} f$.

Analysis of the first term: We start by analyzing

$$T_{(1*)} f(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{w}^* \rangle \int \frac{(\pi - \theta_{\boldsymbol{x},\boldsymbol{x}'})}{\pi} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}'),$$

and focus on the term under the integral. Note that f and $\langle \cdot, \boldsymbol{w}^* \rangle$ are normalized functions such that $\|f\|_{L^2(\rho_{\mathcal{X}})} = \|\langle \cdot, \boldsymbol{w}^* \rangle\|_{L^2(\rho_{\mathcal{X}})} = 1$.

If we consider the set

$$\mathcal{A} = \left\{ \boldsymbol{x}' \in \mathbb{R}^d : \left| \theta_{\boldsymbol{x},\boldsymbol{x}'} - \frac{\pi}{2} \right| < \epsilon \right\}, \epsilon \in \left(0, \frac{\pi}{2}\right),$$

we split the integral over the entire space into the sum of integrals over $\mathcal{A}$ and $\mathcal{A}^c$.

By Lemma B.2, if we let $g = \langle \cdot, \boldsymbol{w}^* \rangle f$, for a given $\epsilon > 0$, the concentration of the Gaussian measure implies

$$\left| \int_{\mathcal{A}^c} \frac{(\pi - \theta_{\boldsymbol{x},\boldsymbol{x}'})}{\pi} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \right| \leq \sqrt{2} e^{-c_1 d \epsilon^2/2} \|g\|_{L^2(\rho_{\mathcal{X}})}$$

for an absolute constant $c_1 > 0$, provided $g \in L^2(\rho_{\mathcal{X}})$. We can see that

$$\int |\langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}')|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') = \int |\|\boldsymbol{x}'\|^2 (\cos \theta_{\boldsymbol{x}',\boldsymbol{w}^*})^2 f(\boldsymbol{x}')|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \leq \int \|\boldsymbol{x}'\|^2 |f(\boldsymbol{x}')|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}').$$

Now, due to Lemma 5.1, the orthonormal eigenbasis is of the separated form

$$\psi_{k,m}(\boldsymbol{x}) = \frac{\|\boldsymbol{x}\|}{\sqrt{d}} Y_{k,m}(\boldsymbol{\omega}),$$

a purely radial function governing the magnitude and $Y_{k,m}(\boldsymbol{\omega})$ is a spherical harmonic of degree k, governing the direction $\boldsymbol{\omega} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$.

Although the ReLU kernel is not universal on the Gaussian function space, by assumption we can write f as a linear combination of elements of the eigenbasis of the kernel

$$f(\boldsymbol{x}) = \sum_{k,m\geq 0} c_{k,m}\psi_{k,m}(\boldsymbol{x})$$

therefore

$$\begin{aligned}|f(\boldsymbol{x})|^2 &= \left(\sum_{k,m} c_{k,m}\frac{\|\boldsymbol{x}\|}{\sqrt{d}}Y_{k,m}(\boldsymbol{\omega})\right)\left(\sum_{k',m'}\overline{c_{k',m'}}\frac{\|\boldsymbol{x}\|}{\sqrt{d}}\overline{Y_{k',m'}(\boldsymbol{\omega})}\right)\\ &= \frac{\|\boldsymbol{x}\|^2}{d}\sum_{k,m}\sum_{k',m'} c_{k,m}\overline{c_{k',m'}}Y_{k,m}(\boldsymbol{\omega})\overline{Y_{k',m'}(\boldsymbol{\omega})}\end{aligned}$$

Decomposing f into this form inside of the integral yields the separation

$$\begin{aligned}\int \|\boldsymbol{x}'\|^2|f(\boldsymbol{x}')|^2\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') &= \sum_{k,m}\sum_{k',m'} c_{k,m}c_{k',m'}\left(\frac{1}{d}\int_{\mathbb{R}}|r^2|^2\phi(r)dr\right)\left(\int_{\Omega}Y_{k,m}(\omega)\overline{Y_{k',m'}(\omega)}\mathrm{d}\Omega\right),\\ &= \sum_{k,m}|c_{k,m}|^2\left(\frac{1}{d}\int_{\mathbb{R}}|r^2|^2\phi(r)dr\right)\\ &= \left(\frac{1}{d}\int_{\mathbb{R}}|r^2|^2\phi(r)dr\right)\end{aligned}$$

where we denote by $\phi(r)$ the Gaussian density associated with the radial decomposition of the norm and we use Parseval's identity to simplify

$$\sum_{k,m}|c_{k,m}|^2 = \|f\|_{L^2(\rho_{\mathcal{X}})} = 1.$$

The second term of the product is simply the norm of the spherical harmonics, which evaluates to 1 since it is a normalized function. The first term is the integral of a fourth Gaussian moment so we solve analytically

$$\frac{1}{d}\int r^4\phi(r)\mathrm{d}r = \frac{1}{d}\mathbb{E}[\|\boldsymbol{x}\|^4] = \frac{d(d+2)}{d} = d+2.$$

Therefore, if we define

$$E_1^{(1)}(\boldsymbol{x}) := \langle\boldsymbol{x},\boldsymbol{w}^*\rangle\int_{\mathcal{A}^c}\frac{(\pi-\theta_{\boldsymbol{x},\boldsymbol{x}'})}{\pi}\langle\boldsymbol{x}',\boldsymbol{w}^*\rangle f(\boldsymbol{x}')\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}').$$

we have

$$\left|E_1^{(1)}(\boldsymbol{x})\right| = \mathcal{O}(de^{-d\epsilon^2})|\langle\boldsymbol{x},\boldsymbol{w}^*\rangle|$$

and we can see the L^2 norm respects

$$\|E_1^{(1)}\|_{L^2(\rho_{\mathcal{X}})} = \mathcal{O}(de^{-d\epsilon^2}).$$

Moreover, inside of $\mathcal{A}$ we write

$$\begin{aligned}&\langle\boldsymbol{x},\boldsymbol{w}^*\rangle\int_{\mathcal{A}}\frac{(\pi-\theta_{\boldsymbol{x},\boldsymbol{x}'})}{\pi}\langle\boldsymbol{x}',\boldsymbol{w}^*\rangle f(\boldsymbol{x}')\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}')\\ &\quad= \frac{\langle\boldsymbol{x},\boldsymbol{w}^*\rangle}{2}\left[\int_{\mathcal{A}}\langle\boldsymbol{x}',\boldsymbol{w}^*\rangle f(\boldsymbol{x}')\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') - \int_{\mathcal{A}}\frac{\theta_{\boldsymbol{x},\boldsymbol{x}'}-\pi/2}{\pi}\langle\boldsymbol{x}',\boldsymbol{w}^*\rangle f(\boldsymbol{x}')\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}')\right]\end{aligned}$$

$$:= \frac{\langle \boldsymbol{x}, \boldsymbol{w}^* \rangle}{2} \int_{\mathcal{A}} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') + E_2^{(1)}(\boldsymbol{x}).$$

Since inside of $\mathcal{A}$ the difference between the angles is bounded by ϵ, using Cauchy-Schwarz we can bound $E_2^{(1)}(\boldsymbol{x})$ by

$$\begin{aligned}|E_2(\boldsymbol{x})^{(1)}| &= \left| \langle \boldsymbol{x}, \boldsymbol{w}^* \rangle \int_{\mathcal{A}} \frac{\theta_{\boldsymbol{x},\boldsymbol{x}'} - \pi/2}{\pi} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \right| \leq \epsilon |\langle \boldsymbol{x}, \boldsymbol{w}^* \rangle| \int_{\mathcal{A}} |\langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}')| \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \\ &\leq \epsilon |\langle \boldsymbol{x}, \boldsymbol{w}^* \rangle| \|\langle \cdot, \boldsymbol{w}^* \rangle\|_{L^2(\rho_{\mathcal{X}})} \|f\|_{L^2(\rho_{\mathcal{X}})} \leq \epsilon |\langle \boldsymbol{x}, \boldsymbol{w}^* \rangle|,\end{aligned}$$

because $\langle \cdot, \boldsymbol{w}^* \rangle$ and f are normalized in $L^2(\rho_{\mathcal{X}})$. Again, this implies the L^2 norm respects

$$\|E_2^{(1)}\|_{L^2(\rho_{\mathcal{X}})} = \mathcal{O}(\epsilon).$$

This gives the following expression for the first term

$$\langle \boldsymbol{x}, \boldsymbol{w}^* \rangle \int \frac{(\pi - \theta_{\boldsymbol{x},\boldsymbol{x}'})}{\pi} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') = \frac{\langle \boldsymbol{x}, \boldsymbol{w}^* \rangle}{2} \int_{\mathcal{A}} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') + E_1^{(1)}(\boldsymbol{x}) + E_2^{(1)}(\boldsymbol{x})$$

We can calculate the integral over $\mathcal{A}$ by writing

$$\int_{\mathcal{A}} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') = \int \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') - \int_{\mathcal{A}^c} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}')$$

and by using the same argument as before, the integral over $\mathcal{A}^c$ induces a function $E_3^{(1)}$ and

$$\begin{aligned}\langle \boldsymbol{x}, \boldsymbol{w}^* \rangle \int_{\mathcal{A}} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') &= \langle \boldsymbol{x}, \boldsymbol{w}^* \rangle \int \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') + E_3^{(1)}(\boldsymbol{x}) \\ &= \langle \boldsymbol{x}, \boldsymbol{w}^* \rangle \langle \langle \cdot, \boldsymbol{w}^* \rangle, f \rangle + E_3^{(1)}(\boldsymbol{x}).\end{aligned}$$

such that $\|E_3^{(1)}\|_{L^2(\rho_{\mathcal{X}})} = \mathcal{O}(de^{-d\epsilon^2})$.

Finally, the first term can be written as

$$\begin{aligned}T_{(1*)} f(\boldsymbol{x}) &= \langle \boldsymbol{x}, \boldsymbol{w}^* \rangle \int \frac{(\pi - \theta_{\boldsymbol{x},\boldsymbol{x}'})}{\pi} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \\ &= \frac{\langle \boldsymbol{x}, \boldsymbol{w}^* \rangle}{2} \langle \langle \cdot, \boldsymbol{w}^* \rangle, f \rangle + E_1^{(1)}(\boldsymbol{x}) + E_2^{(1)}(\boldsymbol{x}) + E_3^{(1)}(\boldsymbol{x}) \\ &:= T_S^{(1*)}(\boldsymbol{x}) + E_1^{(1)}(\boldsymbol{x}) + E_2^{(1)}(\boldsymbol{x}) + E_3^{(1)}(\boldsymbol{x})\end{aligned}$$

and choosing $\epsilon = d^{-1/3}$ we have the bound on the L^2 difference by noticing

$$\begin{aligned}\|T_{(1*)} f - T_S^{(1*)} f\|_{L^2(\rho_{\mathcal{X}})} &\leq \left(\|E_1^{(1)}\|_{L^2(\rho_{\mathcal{X}})} + \|E_2^{(1)}\|_{L^2(\rho_{\mathcal{X}})} + \|E_3^{(1)}\|_{L^2(\rho_{\mathcal{X}})} \right) \\ &\leq \left(\mathcal{O}(de^{-d^{1/3}}) + \mathcal{O}(d^{-1/3}) + \mathcal{O}(de^{-d^{1/3}}) \right) \\ &= o_d(1).\end{aligned}$$

Analysis of the second term: Now we turn to the second term

$$T_{(2*)} f(\boldsymbol{x}) = \int \frac{\sin \theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi} \left(\frac{\|\boldsymbol{x}'\|}{\|\boldsymbol{x}\|} \langle \boldsymbol{x}, \boldsymbol{w}^* \rangle^2 + \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{x}'\|} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle^2 \right) f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}'),$$

and apply the same treatment.

Considering the same set $\mathcal{A}$, we split the integral over the entire space into $\mathcal{A}$ and $\mathcal{A}^c$, and first we show the following function is negligible

$$E_1^{(2)}(\boldsymbol{x}) := \frac{\langle \boldsymbol{x}, \boldsymbol{w}^* \rangle^2}{\|\boldsymbol{x}\|} \int_{\mathcal{A}^c} \frac{\sin \theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi} \|\boldsymbol{x}'\| f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') + \|\boldsymbol{x}\| \int_{\mathcal{A}^c} \frac{\sin \theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi} \frac{\langle \boldsymbol{x}', \boldsymbol{w}^* \rangle^2}{\|\boldsymbol{x}'\|} f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}')$$

Looking at the integral in the second term, again by Lemma B.2, if we define $g = \frac{\langle \cdot, \boldsymbol{w}^* \rangle^2}{\|\cdot\|} f$ we have that

$$\left| \int_{\mathcal{A}^c} \frac{\sin \theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi} \frac{\langle \boldsymbol{x}', \boldsymbol{w}^* \rangle^2}{\|\boldsymbol{x}'\|} f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \right| \leq \sqrt{2} e^{-c_2 d\epsilon^2/2} \|g\|_{L^2(\rho_{\mathcal{X}})},$$

for an absolute constant $c_2 > 0$, as long as $g \in L^2(\rho_{\mathcal{X}})$. Notice that, outside of a set with zero measure, we have

$$\langle \boldsymbol{x}', \boldsymbol{w}^* \rangle = \|\boldsymbol{x}'\| \cos \theta_{\boldsymbol{x}', \boldsymbol{w}^*} \implies \frac{\langle \boldsymbol{x}', \boldsymbol{w}^* \rangle^2}{\|\boldsymbol{x}'\|} = \cos \theta_{\boldsymbol{x}', \boldsymbol{w}^*} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle$$

therefore calculating the $L^2(\rho_{\mathcal{X}})$ norm of g gives:

$$\int \left| \frac{\langle \boldsymbol{x}', \boldsymbol{w}^* \rangle^2}{\|\boldsymbol{x}'\|} f(\boldsymbol{x}') \right|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') = \int \left| \cos \theta_{\boldsymbol{x}, \boldsymbol{w}^*} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}') \right|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \leq \int \left| \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle f(\boldsymbol{x}') \right|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}').$$

Recall that in the analysis of the first term we showed that $\|\cdot\| f \in L^2(\rho_{\mathcal{X}})$, and since this function dominates g, we must have $g \in L^2(\rho_{\mathcal{X}})$ and the concentration bound holds.

Similarly, if we let $g' = \|\cdot\| f$, we already showed $g' \in L^2(\rho_{\mathcal{X}})$, and using the lemma again implies

$$\left| \int_{\mathcal{A}^c} \frac{\sin \theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi} \|\boldsymbol{x}'\| f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \right| \leq \sqrt{2} e^{-c_2 d\epsilon^2/2} \|g'\|_{L^2(\rho_{\mathcal{X}})}.$$

Thus, outside of a set of measure zero, we have the following bound

$$|E_1^{(2)}(\boldsymbol{x})| \leq \mathcal{O}(d e^{-d\epsilon^2}) \left(\|\boldsymbol{x}\| + \frac{\langle \boldsymbol{x}, \boldsymbol{w}^* \rangle^2}{\|\boldsymbol{x}\|} \right)$$

which implies the $L^2(\rho_{\mathcal{X}})$ norm is bounded by

$$\|E_1^{(2)}\|_{L^2(\rho_{\mathcal{X}})} \leq \mathcal{O}(d e^{-d\epsilon^2}) \left(\left\| \|\cdot\| \right\|_{L^2} + \left\| \frac{\langle \cdot, \boldsymbol{w}^* \rangle^2}{\|\cdot\|} \right\|_{L^2} \right) = \mathcal{O}(d^{3/2} e^{-d\epsilon^2}).$$

Next, inside of $\mathcal{A}$, we have

$$\begin{aligned} &\int_{\mathcal{A}} \frac{\sin \theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi} \left(\frac{\|\boldsymbol{x}'\|}{\|\boldsymbol{x}\|} \langle \boldsymbol{x}, \boldsymbol{w}^* \rangle^2 + \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{x}'\|} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle^2 \right) f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \\ &:= \frac{1}{2\pi} \int_{\mathcal{A}} \left(\frac{\|\boldsymbol{x}'\|}{\|\boldsymbol{x}\|} \langle \boldsymbol{x}, \boldsymbol{w}^* \rangle^2 + \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{x}'\|} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle^2 \right) f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') + E_2^{(2)}(\boldsymbol{x}). \end{aligned}$$

where we define

$$E_2^{(2)}(\boldsymbol{x}) = \int_{\mathcal{A}} \left[\frac{\sin \theta_{\boldsymbol{x},\boldsymbol{x}'} - 1}{2\pi} \right] \left(\frac{\|\boldsymbol{x}'\|}{\|\boldsymbol{x}\|} \langle \boldsymbol{x}, \boldsymbol{w}^* \rangle^2 + \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{x}'\|} \langle \boldsymbol{x}', \boldsymbol{w}^* \rangle^2 \right) f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}').$$

Note that

$$|\sin \theta - 1| = 1 - \sin \theta = 1 - \cos(\theta - \pi/2),$$

and if we expand $\cos x$ using Taylor's theorem around 0, noting that $\cos'' x = -\cos x$, we obtain

$$\cos x = 1 + \frac{\cos''(c)}{2} x^2 \implies 1 - \cos x = \frac{\cos(c)}{2} x^2$$

for some $c \in [0, 2\pi]$. Since $\cos\theta \leq 1$ for all $\theta \in [0, 2\pi]$ we have the identity

$$1 - \cos x \leq \frac{x^2}{2}.$$

Thus, letting $x = \theta - \pi/2$ we get

$$|\sin\theta - 1| \leq \frac{1}{2}\left(\theta - \frac{\pi}{2}\right)^2, \forall \theta \in [0, 2\pi],$$

and, since the difference between the angles is bounded by ϵ inside of $\mathcal{A}$, using Cauchy-Schwarz again we can obtain the following bound for $E_2^{(2)}$:

$$\begin{aligned}
|E_2^{(2)}(\boldsymbol{x})| &\leq \frac{\epsilon^2}{4\pi}\sqrt{\int\left(\frac{\|\boldsymbol{x}'\|}{\|\boldsymbol{x}\|}\langle \boldsymbol{x}, \boldsymbol{w}^*\rangle^2 + \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{x}'\|}\langle \boldsymbol{x}', \boldsymbol{w}^*\rangle^2\right)^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}')}\sqrt{\int |f(\boldsymbol{x}')|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}')} \\
&\leq \frac{\epsilon^2}{4\pi}\left(\Big\|\|\cdot\|\Big\|_{L^2(\rho_{\mathcal{X}})}\frac{\langle \boldsymbol{x}, \boldsymbol{w}^*\rangle^2}{\|\boldsymbol{x}\|} + \|\boldsymbol{x}\|\left\|\frac{\langle \cdot, \boldsymbol{w}^*\rangle^2}{\|\cdot\|}\right\|_{L^2(\rho_{\mathcal{X}})}\right)
\end{aligned}$$

almost everywhere.

Taking the $L^2(\rho_{\mathcal{X}})$ norm gives

$$\|E_2^{(2)}(\boldsymbol{x})\|_{L^2(\rho_{\mathcal{X}})} \leq \frac{2\epsilon^2}{4\pi}\Big\|\|\cdot\|\Big\|_{L^2(\rho_{\mathcal{X}})}\left\|\frac{\langle \cdot, \boldsymbol{w}^*\rangle^2}{\|\cdot\|}\right\|_{L^2(\rho_{\mathcal{X}})} \leq \frac{2\epsilon^2\sqrt{d}}{4\pi}.$$

Lastly, we are left with

$$T_{(2*)}f(\boldsymbol{x}) = \frac{1}{2\pi}\int_{\mathcal{A}}\left(\frac{\|\boldsymbol{x}'\|}{\|\boldsymbol{x}\|}\langle \boldsymbol{x}, \boldsymbol{w}^*\rangle^2 + \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{x}'\|}\langle \boldsymbol{x}', \boldsymbol{w}^*\rangle^2\right) f(\boldsymbol{x}')\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') + E_1^{(2)}(\boldsymbol{x}) + E_2^{(2)}(\boldsymbol{x}).$$

Following the same argument as before we can write the integral over $\mathcal{A}$ as the integral over the whole space plus a term $E_3^{(2)}(\boldsymbol{x})$ such that

$$|E_3^{(2)}(\boldsymbol{x})| \leq \mathcal{O}(de^{-d\epsilon^2})\left(\|\boldsymbol{x}\| + \frac{\langle \boldsymbol{x}, \boldsymbol{w}^*\rangle^2}{\|\boldsymbol{x}\|}\right)$$

almost everywhere, and thus $\|E^{(3)}\|_{L^2(\rho_{\mathcal{X}})} = \mathcal{O}(d^{3/2}e^{-d\epsilon^2})$.

Therefore, we can write

$$\begin{aligned}
T_{(2*)}f(\boldsymbol{x}) &= \frac{1}{2\pi}\int\left(\frac{\|\boldsymbol{x}'\|}{\|\boldsymbol{x}\|}\langle \boldsymbol{x}, \boldsymbol{w}^*\rangle^2 + \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{x}'\|}\langle \boldsymbol{x}', \boldsymbol{w}^*\rangle^2\right) f(\boldsymbol{x}')\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') + E_1^{(2)}(\boldsymbol{x}) + E_2^{(2)}(\boldsymbol{x}) + E_3^{(2)}(\boldsymbol{x}) \\
&= \frac{1}{2\pi}\frac{\langle \boldsymbol{x}, \boldsymbol{w}^*\rangle^2}{\|\boldsymbol{x}\|}\langle \|\cdot\|, f\rangle + \frac{1}{2\pi}\|\boldsymbol{x}\|\left\langle \frac{\langle \cdot, \boldsymbol{w}^*\rangle^2}{\|\cdot\|}, f\right\rangle + E_1^{(2)}(\boldsymbol{x}) + E_2^{(2)}(\boldsymbol{x}) + E_3^{(2)}(\boldsymbol{x}) \\
&:= T_S^{(2*)}(\boldsymbol{x}) + E_1^{(2)}(\boldsymbol{x}) + E_2^{(2)}(\boldsymbol{x}) + E_3^{(2)}(\boldsymbol{x}).
\end{aligned}$$

Finally, since we chose $\epsilon = d^{-1/3}$, we have

$$\begin{aligned}
\|T_{(2*)}f - T_S^{(2*)}f\|_{L^2(\rho_{\mathcal{X}})} &\leq \left(\|E_1^{(2)}\|_{L^2(\rho_{\mathcal{X}})} + \|E_2^{(2)}\|_{L^2(\rho_{\mathcal{X}})} + \|E_3^{(2)}\|_{L^2(\rho_{\mathcal{X}})}\right) \\
&\leq \left(\mathcal{O}(d^{3/2}e^{-d^{1/3}}) + \mathcal{O}(d^{-1/6}) + \mathcal{O}(d^{3/2}e^{-d^{1/3}})\right) \\
&= o_d(1).
\end{aligned}$$

□

C.6 Proof of Theorem 5.3

Proof. We show the linear eigenfunction $\psi = \langle \cdot, \boldsymbol{v} \rangle$, $\boldsymbol{v} \in \mathbb{R}^d$, continues to be an eigenfunction of T_1 if $\boldsymbol{v} = \boldsymbol{w}^*$ or $\boldsymbol{v} \perp \boldsymbol{w}^*$.

We use Fubini's Theorem to write

$$T_1\psi(\boldsymbol{x}) = \int k_1(\boldsymbol{x}, \boldsymbol{x}')\psi(\boldsymbol{x}')\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') = \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{1}{d}\boldsymbol{\Gamma})}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\mathbb{E}_{\boldsymbol{x}'\sim\mathcal{N}(0,\boldsymbol{I}_d)}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}'\rangle)\psi(\boldsymbol{x}')]]$$

and analytically solve these integrals. Looking at the inner expectation, and applying Stein's Lemma, we have

$$\mathbb{E}_{\boldsymbol{x}'\sim\mathcal{N}(0,\boldsymbol{I}_d)}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}'\rangle)\psi(\boldsymbol{x}')] = \mathbb{E}_{\boldsymbol{x}'\sim\mathcal{N}(0,\boldsymbol{I}_d)}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}'\rangle)\langle \boldsymbol{x}', \boldsymbol{v}\rangle] = \mathbb{E}_{\boldsymbol{x}'\sim\mathcal{N}(0,\boldsymbol{I}_d)}[\langle D_{\boldsymbol{x}'}\left\{\sigma(\langle \boldsymbol{w}, \boldsymbol{x}'\rangle)\right\}, \boldsymbol{v}\rangle)].$$

Now, since $\sigma'(t) = \mathbf{1}_{\{t\geq 0\}}$ almost everywhere, we have

$$\mathbb{E}_{\boldsymbol{x}'\sim\mathcal{N}(0,\boldsymbol{I}_d)}[D_{\boldsymbol{x}'}\left\{\sigma(\langle \boldsymbol{w}, \boldsymbol{x}'\rangle)\right\}] = \mathbb{E}_{\boldsymbol{x}'\sim\mathcal{N}(0,\boldsymbol{I}_d)}[\langle \boldsymbol{w}, \boldsymbol{v}\rangle \mathbf{1}_{\{\langle \boldsymbol{w}, \boldsymbol{x}'\rangle\geq 0\}}] = \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{v}\rangle.$$

Going back to the outer integral

$$T_1\psi(\boldsymbol{x}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{1}{d}\boldsymbol{\Gamma})}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\langle \boldsymbol{w}, \boldsymbol{v}\rangle],$$

and using Stein's Lemma and the same argument again

$$\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{1}{d}\boldsymbol{\Gamma})}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\langle \boldsymbol{w}, \boldsymbol{v}\rangle] = \frac{1}{d}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{1}{d}\boldsymbol{\Gamma})}[\langle \boldsymbol{\Gamma}\nabla_{\boldsymbol{w}}\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle), \boldsymbol{v}\rangle] = \frac{1}{2d}\langle \boldsymbol{\Gamma}\boldsymbol{x}, \boldsymbol{v}\rangle,$$

therefore

$$T_1\psi(\boldsymbol{x}) = \frac{1}{4d}\langle \boldsymbol{\Gamma}\boldsymbol{x}, \boldsymbol{v}\rangle.$$

Note that the same calculation shows that

$$T_0\psi(\boldsymbol{x}) = \frac{1}{4d}\langle \boldsymbol{x}, \boldsymbol{v}\rangle.$$

As a consequence, if we let $\boldsymbol{v} = \boldsymbol{w}^*$, we have

$$\frac{1}{4d}\langle \boldsymbol{\Gamma}\boldsymbol{x}, \boldsymbol{v}\rangle = \frac{1}{4d}\langle \boldsymbol{\Gamma}\boldsymbol{x}, \boldsymbol{w}^*\rangle = \left(\frac{A}{4d} + \frac{B}{4d}\right)\langle \boldsymbol{x}, \boldsymbol{w}^*\rangle,$$

And, on the other hand, if $\boldsymbol{v} \perp \boldsymbol{w}^*$

$$\frac{1}{4d}\langle \boldsymbol{\Gamma}\boldsymbol{x}, \boldsymbol{v}\rangle = \frac{A}{4d}\langle \boldsymbol{x}, \boldsymbol{v}\rangle.$$

□

C.7 Proof of Theorem 5.5

Proof. We have from Lemma 5.2 that, for a given f that can be represented by the eigenfunctions of the kernel, the action of operator S can be approximated by

$$T_S f = T_S^{(1*)} f + T_S^{(2*)} f + E.$$

If we recall the closed form of the first term

$$T_S^{(1*)} f(\boldsymbol{x}) = \frac{\langle \boldsymbol{x}, \boldsymbol{w}^*\rangle}{2}\langle \langle \cdot, \boldsymbol{w}^*\rangle, f\rangle,$$

we note the inner product term implies that any function that is orthogonal to the linear function must not interact with this term.

Also note that Theorem 5.3 shows the linear function is a true eigenfunction of T_1, thus if a different function happens to be the top eigenfunction of the operator, then they must be orthogonal to each other.

Therefore, to understand the action

$$T_1 f(\boldsymbol{x}) = AT_0 f(\boldsymbol{x}) + \frac{B}{2d} T_S f(\boldsymbol{x}) + T_R f(\boldsymbol{x})$$

on any function other than the linear function it suffices to solve the eigenvalue problem for $AT_0 + \frac{B}{2d} T_S^{(2*)}$.

Solving the eigenvalue problem for $AT_0 + \frac{B}{2d} T_S^{(2*)}$: Recall that

$$T_S^{(2*)} f(\boldsymbol{x}) = \frac{1}{2\pi} \left(\|\boldsymbol{x}\| \left\langle \frac{\langle \cdot, \boldsymbol{w}^* \rangle^2}{\|\cdot\|}, f \right\rangle + \frac{\langle \boldsymbol{x}, \boldsymbol{w}^* \rangle^2}{\|\boldsymbol{x}\|} \langle \|\cdot\|, f \rangle \right)$$

and if we define $f_1(\boldsymbol{x}) := \|\boldsymbol{x}\|$ and $f_2(\boldsymbol{x}) := \frac{\langle \boldsymbol{x}, \boldsymbol{w}^* \rangle^2}{\|\boldsymbol{x}\|}$ we have

$$T_S^{(2*)} f = \frac{1}{2\pi} \Big(f_1 \langle f_2, f \rangle + f_2 \langle f_1, f \rangle \Big),$$

therefore the eigenfunctions must lie in $\text{span}\{f_1, f_2\}$. To bridge the gap between both operators, we recall the eigenbasis of the isotropic operator given by Lemma 5.1, and we translate our functions f_1 and f_2 to that basis.

First, we let $r = \|\boldsymbol{x}\|$ and $\boldsymbol{\omega} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ and consider the following elements of the spherical harmonics basis

$$Y_0(\boldsymbol{\omega}) = 1, \quad \widehat{Y}_2(\boldsymbol{\omega}, \boldsymbol{w}^*) = \langle \boldsymbol{\omega}, \boldsymbol{w}^* \rangle^2 - \frac{1}{d}.$$

Note the notation $\widehat{Y}_2$ to indicate this function is not normalized in $L^2(\rho_{\mathcal{X}})$. With these we can write

$$f_1(\boldsymbol{x}) = \|\boldsymbol{x}\| = r Y_0(\boldsymbol{\omega}), \quad f_2(\boldsymbol{x}) = \frac{\langle \boldsymbol{x}, \boldsymbol{w}^* \rangle^2}{\|\boldsymbol{x}\|} = r \langle \boldsymbol{\omega}, \boldsymbol{w}^* \rangle^2,$$

and writing the inner product as a combination of both functions yields

$$\langle \boldsymbol{\omega}, \boldsymbol{w}^* \rangle^2 = \widehat{Y}_2(\boldsymbol{\omega}, \boldsymbol{w}^*) + \frac{1}{d} Y_0(\boldsymbol{\omega}),$$

$$f_1(\boldsymbol{x}) = r Y_0(\boldsymbol{\omega}), \quad f_2(\boldsymbol{x}) = r \widehat{Y}_2(\boldsymbol{\omega}, \boldsymbol{w}^*) + \frac{r}{d} Y_0(\boldsymbol{\omega}).$$

Next, we solve the eigenvalue problem of the combined operator $AT_0 + \frac{B}{2d} T_S^{(2*)}$. Let φ be a non trivial eigenfunction of this operator. Since T_0 is fully diagonalized by rY_0 and $r\widehat{Y}_2$ and since the action of $T_S^{(2*)}$ is restricted to $\text{span}\{rY_0, r\widehat{Y}_2\}$, we know φ must take the form

$$\varphi = c_0 [r Y_0(\boldsymbol{\omega})] + c_2 [r \widehat{Y}_2(\boldsymbol{\omega}, \boldsymbol{w}^*)]$$

for some coefficients c_0 and c_2.

If we apply

$$T_S^{(2*)} \varphi = \frac{\langle f_2, \varphi \rangle}{2\pi} f_1 + \frac{\langle f_1, \varphi \rangle}{2\pi} f_2,$$

using the identities $f_1 = rY_0(\omega)$ and $f_2 = r\widehat{Y}_2(\omega, \boldsymbol{w}^*) + \frac{r}{d} Y_0(\omega)$, we obtain

$$T_S^{(2*)} \varphi = \left\langle r\widehat{Y}_2 + \frac{r}{d} Y_0, c_0 [rY_0] + c_2 [r\widehat{Y}_2] \right\rangle \frac{rY_0}{2\pi} + \left\langle rY_0, c_0 [rY_0] + c_2 [r\widehat{Y}_2] \right\rangle \frac{1}{2\pi} \left(r\widehat{Y}_2 + \frac{r}{d} Y_0 \right)$$

$$= \frac{rY_0}{2\pi}\left(\frac{c_0}{d}\langle rY_0, rY_0\rangle + c_2\langle r\widehat{Y}_2, r\widehat{Y}_2\rangle\right) + \frac{1}{2\pi}\left(r\widehat{Y}_2 + \frac{r}{d}Y_0\right)c_0\langle rY_0, rY_0\rangle$$
$$= \frac{rY_0}{2\pi}\left(\frac{2c_0}{d}\langle rY_0, rY_0\rangle + c_2\langle r\widehat{Y}_2, r\widehat{Y}_2\rangle\right) + \frac{r\widehat{Y}_2}{2\pi}\Big(c_0\langle rY_0, rY_0\rangle\Big).$$

This translates to the matrix operator on the vector of coefficients

$$T_S^{(2*)}\varphi = \frac{1}{2\pi}\begin{bmatrix}\frac{2}{d}\langle rY_0, rY_0\rangle & \langle r\widehat{Y}_2, r\widehat{Y}_2\rangle \\ \langle rY_0, rY_0\rangle & 0\end{bmatrix}\begin{pmatrix}c_0\\c_2\end{pmatrix}\cdot\begin{pmatrix}rY_0\\r\widehat{Y}_2\end{pmatrix}$$

and since $\langle rY_0, rY_0\rangle = d$ and $\langle r\widehat{Y}_2, r\widehat{Y}_2\rangle = \frac{2(d-1)}{d(d+2)}$ we have

$$T_S^{(2*)}\varphi = \frac{1}{2\pi}\begin{bmatrix}2 & \frac{2(d-1)}{d(d+2)} \\ d & 0\end{bmatrix}\begin{pmatrix}c_0\\c_2\end{pmatrix}\cdot\begin{pmatrix}rY_0\\r\widehat{Y}_2\end{pmatrix}.$$

Furthermore, because these are non normalized orthogonal eigenfunctions of T_0, they perfectly diagonalize the operator. Thus, if $\lambda_{\max}(T_0)$ and $\lambda_2(T_0)$ are the eigenvalues for rY_0 and rY_2 respectively, we have the following

$$AT_0\varphi = \begin{bmatrix}A\lambda_{\max}(T_0) & 0 \\ 0 & A\lambda_2(T_0)\end{bmatrix}\begin{pmatrix}c_0\\c_2\end{pmatrix}\cdot\begin{pmatrix}rY_0\\r\widehat{Y}_2\end{pmatrix},$$

and the combined action is simply the combination of these matrices

$$\left(AT_0 + \frac{B}{2d}T_S^{(2*)}\right)\varphi = \begin{bmatrix}A\lambda_{\max}(T_0) + \frac{B}{2\pi d} & \frac{B(d-1)}{2\pi d^2(d+2)} \\ \frac{B}{4\pi} & A\lambda_2(T_0)\end{bmatrix}\begin{pmatrix}c_0\\c_2\end{pmatrix}\cdot\begin{pmatrix}rY_0\\r\widehat{Y}_2\end{pmatrix}.$$

Note that λ_2 represents the eigenvalue corresponding to a single spherical harmonic of degree 2. We know the dimension of the degree-2 subspace on $\mathbb{S}^{d-1}$ is given by the degeneracy formula $N(d,2) = \frac{(d-1)(d+2)}{2}$, thus since the total macroscopic energy is distributed uniformly across the orthogonal basis in the subspace, the individual eigenvalue scales as $\lambda_2 = \mathcal{O}(d^{-2})$.

We draw attention to the $\mathcal{O}(B)$ coefficient on the lower left entry of the matrix which implies a strong coupling between these functions under the combined operator.

Again, solving for the roots of the characteristic polynomial, we find the following two eigenvalues

$$\widetilde{\lambda}_\pm = \frac{1}{2}\left[\left(A\lambda_{\max}(T_0) + \frac{B}{2\pi d} + A\lambda_2(T_0)\right) \pm \sqrt{\left(A\lambda_{\max}(T_0) + \frac{B}{2\pi d} - A\lambda_2(T_0)\right)^2 + \frac{B^2(d-1)}{2\pi^2 d^2(d+2)}}\right],$$

or, expanding the square root,

$$\widetilde{\lambda}_+ = A\lambda_{\max}(T_0) + \frac{B}{2\pi d} + \frac{B^2}{4\pi d^2}\left(1 + o(1)\right), \qquad \widetilde{\lambda}_- = A\lambda_2(T_0) - \frac{B^2}{4\pi d^2}\left(1 + o(1)\right).$$

Furthermore, we know that any eigenfunction with this eigenvalue must satisfy

$$\begin{bmatrix}A\lambda_{\max}(T_0) + \frac{B}{2\pi d} & \frac{B(d-1)}{2\pi d^2(d+2)} \\ \frac{B}{4\pi} & A\lambda_2(T_0)\end{bmatrix}\begin{pmatrix}c_0\\c_2\end{pmatrix} = \lambda_\pm\begin{pmatrix}c_0\\c_2\end{pmatrix}$$

which implies the relationship

$$\frac{B}{4\pi}c_0 = (\lambda_\pm - A\lambda_2(T_0))c_2.$$

Since the eigenfunctions can be arbitrarily scaled, we choose $c_0 = \frac{1}{\sqrt{d}}$ to reconstruct the isotropic eigenfunction $c_0\|\boldsymbol{x}\|Y_0 = \frac{\|\boldsymbol{x}\|}{\sqrt{d}}Y_0$, and solving for c_2 gives

$$c_2 = \frac{B}{4\pi\sqrt{d}(\widetilde{\lambda}_\pm - A\lambda_2(T_0))}$$

which lets us define the new eigenfunctions as

$$\widehat{\Psi}_\pm(\boldsymbol{x}) = \frac{\|\boldsymbol{x}\|}{\sqrt{d}}Y_0(\boldsymbol{\omega}) + \left[\frac{B}{4\pi(\widetilde{\lambda}_\pm - A\lambda_2(T_0))}\right]\frac{\|\boldsymbol{x}\|}{\sqrt{d}}\widehat{Y}_2(\boldsymbol{\omega}, \boldsymbol{w}^*)\,.$$

To ensure this function is normalized, we calculate

$$\|\widehat{\Psi}_\pm\|^2_{L^2(\rho_\mathcal{X})} = 1 + \frac{B^2}{16\pi^2(\widetilde{\lambda}_\pm - A\lambda_2(T_0))^2}\frac{2d-2}{d^2(d+2)} := Z_\pm$$

and therefore the normalized approximate eigenfunctions are given by

$$\widetilde{\Psi}_\pm(\boldsymbol{x}) = \frac{1}{\sqrt{Z_\pm}}\left(\frac{\|\boldsymbol{x}\|}{\sqrt{d}}Y_0(\boldsymbol{\omega}) + \left[\frac{B}{4\pi(\widetilde{\lambda}_\pm - A\lambda_2(T_0))}\right]\frac{\|\boldsymbol{x}\|}{\sqrt{d}}\widehat{Y}_2(\boldsymbol{\omega}, \boldsymbol{w}^*)\right)\,.$$

Lastly, to explicitly show the magnitude of the quadratic feature, if Y_2 is the normalized zonal harmonics, we write $\widehat{Y}_2 = \|\widehat{Y}_2\|_{L^2(\rho_\mathcal{X})}Y_2$ with $\|\widehat{Y}_2\|_{L^2(\rho_\mathcal{X})} = \frac{1}{d}\sqrt{\frac{2d-2}{d+2}}$. Therefore, we have the final form for the eigenfunctions:

$$\widetilde{\Psi}_\pm(\boldsymbol{x}) = \frac{1}{\sqrt{Z_\pm}}\frac{\|\boldsymbol{x}\|}{\sqrt{d}}\left[Y_0(\boldsymbol{\omega}) + \tau_\pm Y_2(\boldsymbol{\omega}, \boldsymbol{w}^*)\right],$$

where we define the alignment magnitudes by the quantity

$$\tau_\pm := \frac{1}{d}\sqrt{\frac{2d-2}{d+2}}\left[\frac{B}{4\pi(\widetilde{\lambda}_\pm - A\lambda_2(T_0))}\right].$$

From now on we study the approximate eigenfunction $\widetilde{\Psi}_+$ with associated approximate eigenvalue λ_+, since we clearly have $\widetilde{\lambda}_+ > \widetilde{\lambda}_-$.

Separation of $\widetilde{\lambda}_+$ from the remaining spectrum.: By Theorem 5.3, the linear subspace $\mathcal{E}_{\text{lin}}$ is invariant under T_1, and its largest eigenvalue satisfies

$$\lambda_{\text{lin},1}(T_1) = O(d^{-1}).$$

We therefore consider the orthogonal complement of the linear sector. By Lemma 5.2,

$$T_S = T_S^{(1*)} + T_S^{(2*)} + E.$$

The operator $T_S^{(1*)}$ depends only on the projection onto $\langle\cdot, \boldsymbol{w}^*\rangle$, and hence vanishes on $\mathcal{E}_{\text{lin}}^\perp$. Moreover,

$$T_S^{(2*)}f = \frac{1}{2\pi}\left(f_1\langle f_2, f\rangle + f_2\langle f_1, f\rangle\right),$$

where

$$f_1(\boldsymbol{x}) = \|\boldsymbol{x}\|, \qquad f_2(\boldsymbol{x}) = \frac{\langle \boldsymbol{x}, \boldsymbol{w}^*\rangle^2}{\|\boldsymbol{x}\|}.$$

Thus $T_S^{(2*)}$ is supported on

$$\mathcal{V}_{\text{eff}} = \text{span}\{f_1, f_2\}.$$

If we define

$$\mathcal{H}_{\mathrm{rem}} = (\mathcal{E}_{\mathrm{lin}} \oplus \mathcal{V}_{\mathrm{eff}})^{\perp}.$$

for every $f \in \mathcal{H}_{\mathrm{rem}}$, we have

$$T_S^{(1*)} f = T_S^{(2*)} f = 0.$$

Hence, using the uniform error estimate in Lemma 5.2 and the remainder bound of Lemma 4.4,

$$T_1 f = AT_0 f + \frac{B}{2d} E f + T_R f,$$

with

$$\|Ef\| = o_d(1)\|f\|, \qquad \|T_R f\| = o_d\left(\frac{|B|}{d}\right)\|f\|.$$

Therefore, on $\mathcal{H}_{\mathrm{rem}}$,

$$\langle f, T_1 f\rangle = A\langle f, T_0 f\rangle + o_d\left(\frac{|B|}{d}\right)\|f\|^2.$$

Since $\mathcal{V}_{\mathrm{eff}}$ contains the degree-0 radial mode and the target-aligned degree-2 mode, the largest eigenvalue of T_0 on $\mathcal{H}_{\mathrm{rem}}$ is bounded by the individual degree-2 eigenvalue. Since k_0 is a dot-product kernel in the angular variable, by the Funk–Hecke theorem all $N(d,k)$ eigenfunctions $\psi_{k,m}$ of degree k share the same eigenvalue $\lambda_k(T_0)$. Hence, by the trace identity,

$$N(d,k)\,\lambda_k(T_0) \le \sum_{j\ge 0} N(d,j)\,\lambda_j(T_0) = \mathrm{Tr}(T_0) = \int k_0(\boldsymbol{x},\boldsymbol{x})\,d\rho_X(\boldsymbol{x}) = \frac{\mathbb{E}\|\boldsymbol{x}\|^2}{2d} = \frac{1}{2}.$$

Since $N(d,k)$ is nondecreasing in k for $d \ge 3$ and $N(d,2) = \frac{(d-1)(d+2)}{2}$, we obtain, for every $k \ge 2$,

$$\lambda^{(k)}(T_0) \le \frac{1}{2N(d,2)} = \frac{1}{(d-1)(d+2)} = \mathcal{O}(d^{-2}).$$

Finally, H_{rem} is orthogonal to the linear functions and to $\|\boldsymbol{x}\|Y_0(\boldsymbol{\omega})$, hence contained in the closed span of $\{\psi_{k,m} : k \ge 2\}$, and therefore

$$\sup_{f\in H_{\mathrm{rem}},\, \|f\|_{L^2(\rho_X)}=1} \langle f, T_0 f\rangle \le \frac{1}{(d-1)(d+2)}.$$

Thus, by this argument, Lemma 5.1 and the degree-2 multiplicity,

$$\lambda_{\mathrm{rem}}(T_0) = O(d^{-2}).$$

Thus

$$\sup_{\|f\|=1,\ f\in\mathcal{H}_{\mathrm{rem}}} \langle f, T_1 f\rangle = O(d^{-2}) + o_d\left(\frac{|B|}{d}\right) = o_d(1),$$

since $\frac{|B|}{d} = \Theta(d^{2\zeta-2})$ and $\zeta < 1$.

On the effective space $\mathcal{V}_{\mathrm{eff}}$, the preceding 2×2 calculation gives

$$\widetilde{\lambda}_+ = A\lambda_{\max}(T_0) + \frac{B}{2\pi d} + o_d\left(\frac{|B|}{d}\right) = \Theta(1),$$

while

$$\widetilde{\lambda}_- = O\left(\lambda_{\deg 2}(T_0) + \frac{B^2}{d^3}\right) = o_d(1).$$

Hence $\widetilde{\lambda}_+$ is separated by a constant gap from the linear sector, the effective lower mode, and $\mathcal{H}_{\mathrm{rem}}$. Finally, from

$$(T_1 - \widetilde{\lambda}_+ I)\widetilde{\Psi}_+ = e, \qquad \|e\| = o_d\left(\frac{|B|}{d}\right),$$

the spectral theorem implies

$$\operatorname{dist}(\widetilde{\lambda}_+, \sigma(T_1)) \leq \|e\| = o_d\left(\frac{|B|}{d}\right).$$

Since $\widetilde{\lambda}_+ = \Theta(1)$ while every other spectral sector is $o_d(1)$, this eigenvalue must be close to the unique top eigenvalue. Therefore

$$\lambda_{\max}(T_1) = \widetilde{\lambda}_+ + o_d\left(\frac{|B|}{d}\right),$$

and in particular

$$\sup\{\sigma(T_1) \setminus \lambda_{\max}(T_1)\} = o_d(1).$$

The approximate eigenfunction and the true eigenfunction are close in norm: Finally, to show $\widetilde{\Psi}_+$ is close to the original top eigenfunction, we denote by Ψ the top eigenfunction and use the expansion to write

$$T_1\widetilde{\Psi}_+ = \left(AT_0 + \frac{B}{2d}T_{S_{(1*)}} + \frac{B}{2d}T_S^{(2*)}\right)\widetilde{\Psi} + e = \widetilde{\lambda}_+\widetilde{\Psi}_+ + e$$

with $\|e\| = o_d\left(\frac{|B|}{d}\right)$ and we write this as

$$(T_1 - \widetilde{\lambda}_+ I)\widetilde{\Psi}_+ = e.$$

Now, consider the projection onto the top eigenspace of T_1 defined by

$$P_{\text{top}} f = \langle \Psi, f\rangle \Psi,$$

and its orthogonal complement $P_\perp = I - P_{\text{top}}$. Expanding the action of the combined operator using these projections we get

$$(T_1 - \widetilde{\lambda}_+ I)\widetilde{\Psi}_+ = (T_1 - \widetilde{\lambda}_+ I)(P_{\text{top}}\widetilde{\Psi}_+ + P_\perp\widetilde{\Psi}_+) = e$$

and since $T_1(P_{\text{top}} f) = \lambda_{\max}(T_1)P_{\text{top}} f$ this simplifies to

$$[\lambda_{\max}(T_1) - \widetilde{\lambda}_+]P_{\text{top}}\widetilde{\Psi}_+ + (T_1 - \widetilde{\lambda}_+ I)P_\perp\widetilde{\Psi}_+ = e.$$

Since we have only orthogonal objects, the norm of this expression is equal to

$$|\lambda_{\max}(T_1) - \lambda_+|^2\|P_{\text{top}}\widetilde{\Psi}_+\|^2_{L^2(\rho_{\mathcal{X}})} + \|(T_1 - \widetilde{\lambda}_+ I)P_\perp\widetilde{\Psi}_+\|^2_{L^2(\rho_{\mathcal{X}})} = o_d\left(\frac{B^2}{d^2}\right)$$

which immediately implies

$$\|(T_1 - \widetilde{\lambda}_+ I)P_\perp\widetilde{\Psi}_+\|^2_{L^2(\rho_{\mathcal{X}})} = o_d\left(\frac{B^2}{d^2}\right),$$

and we will use this to bound $\|P_\perp\widetilde{\Psi}\|$. If we define the approximate spectral gap by $\widetilde{\delta} = \inf_{\mu\in\sigma(T_1)\setminus\{\lambda_{\max}(T_1)\}}|\mu - \widetilde{\lambda}_+|$, we have

$$\|(T_1 - \widetilde{\lambda}_+ I)P_\perp\widetilde{\Psi}_+\|_{L^2(\rho_{\mathcal{X}})} \geq \widetilde{\delta}\|P_\perp\widetilde{\Psi}_+\|_{L^2(\rho_{\mathcal{X}})}$$

and

$$\widetilde{\delta}\|P_\perp\widetilde{\Psi}_+\|_{L^2(\rho_{\mathcal{X}})} = o_d\left(\frac{|B|}{d}\right).$$

From the previous discussion, we established that

$$\sup\{\sigma(T_1) \setminus \lambda_{\max}(T_1)\} = o_d(1), \quad \lambda_{\max}(T_1) = \widetilde{\lambda}_+ + o_d\left(\frac{|B|}{d}\right) = \Theta(1),$$

therefore the true spectral gap is lower bounded uniformly, i.e.

$$\delta := \lambda_{\max}(T_1) - \sup\{\sigma(T_1) \setminus \lambda_{\max}(T_1)\}$$

satisfies

$$\delta = \Omega(1)\,.$$

For any $\lambda \in \sigma(T_1) \setminus \lambda_{\max}(T_1)$, by the triangle inequality we have

$$\begin{aligned}|\lambda - \widetilde{\lambda}_+| &\geq |\lambda - \lambda_{\max}(T_1)| - |\lambda_{\max}(T_1) - \widetilde{\lambda}_+| \\ &\geq \delta - |\lambda_{\max}(T_1) - \widetilde{\lambda}_+|\end{aligned}$$

therefore

$$\widetilde{\delta} \geq \delta - o_d\left(\frac{|B|}{d}\right),$$

and we know $\widetilde{\delta}$ must bounded away from zero independently of d. Thus, going back to bounding the norm of $P_\perp \widetilde{\Psi}_+$, we obtain

$$\|P_\perp \widetilde{\Psi}_+\|_{L^2(\rho_{\mathcal{X}})} = \frac{1}{\widetilde{\delta}} o_d\left(\frac{|B|}{d}\right) = o_d\left(\frac{|B|}{d}\right).$$

To bound the difference under the norm, we note that $\|\widetilde{\Psi}\| = 1$ and write

$$\|\widetilde{\Psi}_+\|^2_{L^2(\rho_{\mathcal{X}})} = |\langle \widetilde{\Psi}_+, \Psi\rangle|^2 + \|P_\perp \widetilde{\Psi}_+\|^2_{L^2(\rho_{\mathcal{X}})}$$

$$|\langle \widetilde{\Psi}_+, \Psi\rangle|^2 = 1 - o_d\left(\frac{B^2}{d^2}\right)$$

Finally, combining all together we have

$$\begin{aligned}\|\widetilde{\Psi}_+ - \Psi\|^2_{L^2(\rho_{\mathcal{X}})} &= \|\widetilde{\Psi}_+\|^2_{L^2(\rho_{\mathcal{X}})} - 2\langle \widetilde{\Psi}_+, \Psi\rangle + \|\Psi\|^2_{L^2(\rho_{\mathcal{X}})} \\ &= 2 - 2\left[1 - o_d\left(\frac{B^2}{d^2}\right)\right] \\ &= o_d\left(\frac{B^2}{d^2}\right)\end{aligned}$$

and the proof is completed. □

D Experimental details

In this section, we provide comprehensive details regarding the experimental setup, hyperparameter configurations, and the mathematical framework used for the results presented in the main text. All experiments follow the two-stage protocol of Eq. (2.3). Inputs are drawn as $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, the target direction $\boldsymbol{w}^*$ is drawn uniformly on the unit sphere, and labels are $y = g(\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle) + \varepsilon$. Unless stated otherwise, the link is

$$g(z) = \tanh(2z) + 2\big(e^{-1/2} - \cos z\big),$$

which is bounded, centered and Lipschitz with $c_1 = \mathbb{E}[g'(z)] \approx 0.73$, so Assumption 2.1 holds, and the noise is Gaussian with standard deviation $\sigma = 0.3$. We refer to this setting as the *original scenario*. Kernel parameters A and B follow from Eq. (3.3) for the step size and sample sizes of the run.

D.1 Different models

Throughout the experiments we consider the following models.

1. **Base ReLU kernel** (k_0): the arc-cosine kernel at initialization.

2. **Idealized kernel** (k_1): the idealized post update kernel with weights governed by the distribution $\mathcal{N}(0, \frac{1}{d}\mathbf{\Gamma})$ with $\mathbf{\Gamma}$ given by Eq. (3.3). We compute this kernel by the infinite width identity in Eq. (5.1)

$$k_1(\boldsymbol{x}, \boldsymbol{x}') = \frac{\sqrt{(\boldsymbol{x}^\top \mathbf{\Gamma} \boldsymbol{x})(\boldsymbol{x}'^\top \mathbf{\Gamma} \boldsymbol{x}')}}{2\pi d}\Big[\gamma_{\mathbf{\Gamma}}\big(\pi - \arccos(\gamma_{\mathbf{\Gamma}})\big) + \sqrt{1-\gamma_{\mathbf{\Gamma}}^2}\Big],$$

with $\gamma_{\mathbf{\Gamma}} := \boldsymbol{x}^\top \mathbf{\Gamma} \boldsymbol{x}' / \sqrt{(\boldsymbol{x}^\top \mathbf{\Gamma} \boldsymbol{x})(\boldsymbol{x}'^\top \mathbf{\Gamma} \boldsymbol{x}')}$. We also emphasize that this object has direct access to the target $\boldsymbol{w}^*$ through the spiked covariance.

3. **Single-dataset kernel** ($k_{\mathcal{D}}$): the kernel of the network after the rank-one part of the step for the realized first-stage sample $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, in the infinite-width limit, is the ReLU kernel where the weights depend solely on the Gaussian distribution spiked by the training sample: with $\mu_1 = \mathbb{E}[\sigma'(z)] = \frac{1}{2}$ for the ReLU. Unlike k_1, it uses only the data, not $\boldsymbol{w}^*$.

4. **ReLU network**: the two-layer network of Section 2 without biases,

$$f(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \, \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x} \rangle), \qquad \boldsymbol{w}_j \sim \mathcal{N}(0, \boldsymbol{I}_d/d), \qquad a \sim \mathcal{N}(0, \boldsymbol{I}_m/m).$$

The first layer takes one full-batch gradient step Eq. (2.3), $\boldsymbol{W}_1 = \boldsymbol{W}_0 - \eta \nabla_W L(\boldsymbol{W}_0)$ with $L(W) = \frac{1}{2n}\sum_{i=1}^n (f(\boldsymbol{x}_i) - y_i)^2$ and a fixed. The second layer is then fitted by ridge regression on the features $\sigma(\langle \boldsymbol{w}_j, \boldsymbol{x} \rangle)/\sqrt{m}$, which is kernel ridge regression with the finite-width kernel

$$k_{\boldsymbol{W}}(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{m} \sum_{j=1}^{m} \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x} \rangle) \, \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}' \rangle) \qquad \text{at } \boldsymbol{W} = \boldsymbol{W}_1.$$

5. **Random features at initialization**: the same, with $\boldsymbol{W} = \boldsymbol{W}_0$ (same width, no gradient step).

6. **Linear ridge**: kernel ridge regression with $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle / d$.

Kernels. For each kernel k_i above, we calculate the kernel matrix $\boldsymbol{K}_i(\boldsymbol{Z}) \in \mathbb{R}^{N \times N}$ whose entries correspond to

$$[\boldsymbol{K}_i(\boldsymbol{Z})]_{k,l} = k_i(\boldsymbol{z}_k, \boldsymbol{z}_l).$$

D.2 Metrics and significance

The alignment and end-to-end experiments were executed over 10 independent trials, the end-to-end stress tests over 5 and the alignment stress tests over 10. Shaded regions in the figures indicate $\pm$ one standard deviation across trials. For every experiment we set random seeds for all sources of randomness (data sampling, network initialization, cross-validation folds), listed below.

D.3 Alignment experiment

For the results in Fig. 1 we set the seed 54643. For the alignment analysis, we vary the dimension $d \in \{50, 100, 200, 400, 800, 1600, 3200\}$, sample $N = d$ points $\boldsymbol{Z} \in \mathbb{R}^{N \times d}$, and consider four levels of the spike $B \in \{5d^{3/10}, 5d^{5/10}, 5d^{7/10}, 5d^{9/10}\}$, one per panel. The alignment reported is calculated as

$$\langle \boldsymbol{v}_i^{\text{top}}, Y_2(\boldsymbol{Z}) \rangle^2,$$

where $\boldsymbol{v}_i^{\text{top}}$ is the lead eigenvector of the kernel matrix $\boldsymbol{K}_i(\boldsymbol{Z})$ and $\boldsymbol{Y}_2(\boldsymbol{Z}) \in \mathbb{R}^N$ is the normalized vector of

$$\widehat{Y}_2(\boldsymbol{\omega}_k, \boldsymbol{w}^*) = \langle \boldsymbol{\omega}_k, \boldsymbol{w}^* \rangle^2 - \frac{1}{d}, \qquad \boldsymbol{\omega}_k = \frac{\boldsymbol{z}_k}{\|\boldsymbol{z}_k\|}.$$

The alignment is squared because the sign of an eigenvector is arbitrary.

Each panel of Fig. 1 corresponds to one value of B and also shows the network after one gradient step, with width $m = 400$, trained on $n = 2d$ samples of the original scenario. Its step size η is

obtained by solving Eq. (3.3) for η, so that the idealized kernel of the network has exactly the B of the panel; the corresponding A follows from Eq. (3.3). For this step we report the alignment of the network's kernel matrix $K_{\boldsymbol{W}_1}(\boldsymbol{Z})$, of k_1 with the A and B of the step, and of $k_{\mathcal{D}}$ for the same first-stage sample. The network uses a separate random stream (seed $10^4(t+1)+d$ for trial t), so that k_0 and k_1 are computed on exactly the same draws.

D.4 End-to-end experiment

For the results in Fig. 2 we set the seed $40000+t$ for trial t, for both the data and the network initialization. We work on the fixed dimension $d=300$ with the original scenario. The network has width $m=400$ and its first layer takes one gradient step on $n=8d=2400$ samples with step size $\eta=3d^{\zeta}$, $\zeta\in\{5/8,3/4,7/8\}$. The second stage uses $N\in\{75,150,300,600,1200,2400\}$ fresh samples Z, independent of the first stage, and all models are fitted on the same second-stage sample.

We use Kernel Ridge Regression (KRR) with every model of Appendix D.1 to obtain the solutions $\widehat{a}_i=(\boldsymbol{K}_i(\boldsymbol{Z})+\lambda\boldsymbol{I}_N)^{-1}y$, with regularization parameter λ. The regularization is chosen through 5-fold cross-validation, for each sample size N and each model, from 28 logarithmically spaced values between $10^{-8}\kappa_i$ and $10\,\kappa_i$, where $\kappa_i=\operatorname{tr}(\boldsymbol{K}_i(\boldsymbol{Z}))/N$ is the scale of the kernel, with random state 38182. We sample a test set $\widetilde{\boldsymbol{Z}}$ from the same distribution as Z, consisting of $M=2000$ independent samples, and construct the test kernel matrices $\boldsymbol{K}_i(\widetilde{\boldsymbol{Z}},\boldsymbol{Z})\in\mathbb{R}^{M\times N}$ whose entries are given by

$$[\boldsymbol{K}_i(\widetilde{\boldsymbol{Z}},\boldsymbol{Z})]_{k,l}=k_i(\widetilde{\boldsymbol{z}_k},\boldsymbol{z}_l),$$

where $\widetilde{\boldsymbol{z}_k}$ is the k-th sample from $\widetilde{\boldsymbol{Z}}$ and $\boldsymbol{z}_l$ is the l-th sample from $\boldsymbol{Z}$. Then, we construct the predictors

$$\widehat{f}_i(\boldsymbol{z})=\sum_{j=1}^{N}(\widehat{a}_i)_j\,k_i(\boldsymbol{z},\boldsymbol{z}_j).$$

To obtain the metric from the figure, we calculate the normalized excess risk

$$\frac{\frac{1}{M}\sum_{k=1}^{M}\big(\widehat{f}_i(\widetilde{\boldsymbol{z}_k})-f^*(\widetilde{\boldsymbol{z}_k})\big)^2}{\mathbb{E}[f^*(\boldsymbol{x})^2]},\qquad f^*(\boldsymbol{x})=g(\langle\boldsymbol{w}^*,\boldsymbol{x}\rangle),$$

where $\mathbb{E}[f^*(\boldsymbol{x})^2]=\mathbb{E}[g(z)^2]\approx 1.43$ is computed by Gauss–Hermite quadrature. A value of 1 corresponds to the zero predictor. The dotted line in the figures is the risk $1-c_1^2/\mathbb{E}[g(z)^2]$ of the best linear predictor.

D.5 Stress tests

The stress tests change one factor at a time from the original scenario, which is shown as the first panel of every stress figure. Table 1 lists the scenarios.

factor	scenarios
bounded link	$\tanh(2z)$; $\sin(2z)$; $\cos(2z)$ ($c_1=0$, violates Assumption 2.1)
unbounded link	$3z+2z^2+4\sin(2z)$; $\max(z,0)$; e^z
noise	Gaussian with $\sigma\in\{0,1\}$; Student-t_3 with standard deviation 1
input distribution	Rademacher $\{\pm1\}^d$; uniform on $\sqrt{d}\,\mathbb{S}^{d-1}$; elliptical Student-$t_5$ with identity covariance; $\mathcal{N}(0,\Sigma)$ with log-uniform spectrum in $[1/4,4]$ and random eigenbasis
dimension	$d\in\{50,100,200,400,800\}$ (end-to-end only)

Table 1: Scenarios of the stress tests; every other setting is that of the original scenario.

Every link is centered and rescaled to $\mathbb{E}[g(z)^2]\approx 1.43$, the value of the original link, so that only its shape changes. For the anisotropic inputs the index is standardized, $\boldsymbol{z}=\langle\boldsymbol{w}^*,\boldsymbol{x}\rangle/\mathrm{sd}(\langle\boldsymbol{w}^*,\boldsymbol{x}\rangle)$. The constants A,B of k_1 are always computed from the Gaussian formula of Eq. (3.3), even for non-Gaussian inputs, while $k_{\mathcal{D}}$ uses the realized data. When $\langle\boldsymbol{w}^*,\boldsymbol{x}\rangle$ is not Gaussian, $\mathbb{E}[f^*(\boldsymbol{x})^2]$ is estimated by Monte Carlo with $2\cdot10^5$ samples, and the best-linear-predictor line is not drawn.

End-to-end stress test. Same protocol as Appendix D.4, with $n = 8d$, $m = 4d/3$ and $\eta = 3d^{\zeta}$ at every dimension; Figs. 3–6 show $\zeta = 7/8$. Seeds as in the end-to-end experiment.

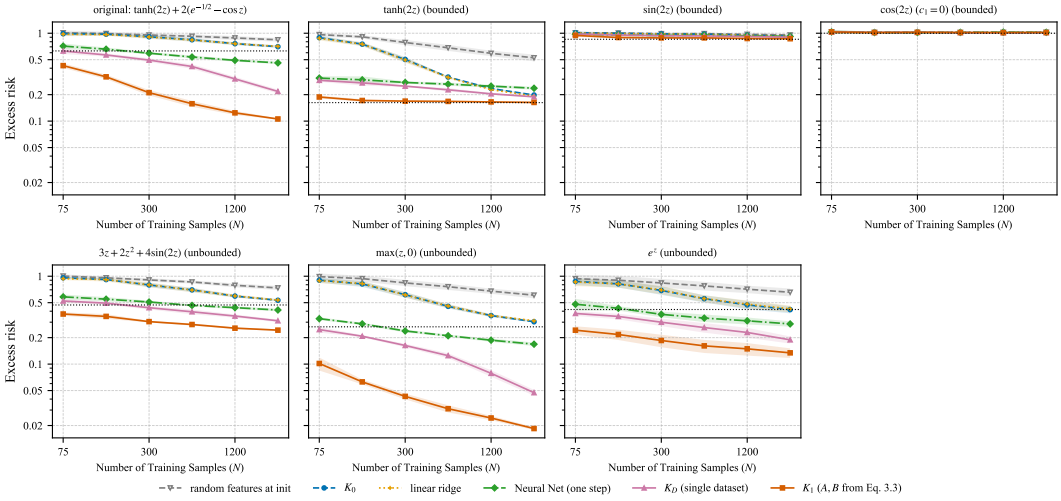


Figure 3: Stress test comparing variations of link functions. The functions are allowed to violate Assumption 2.1 and are marked as bounded versus unbounded respectively.

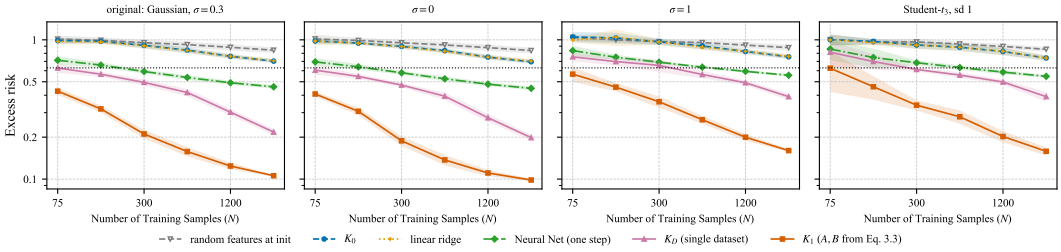


Figure 4: Stress test comparing different levels of noise.

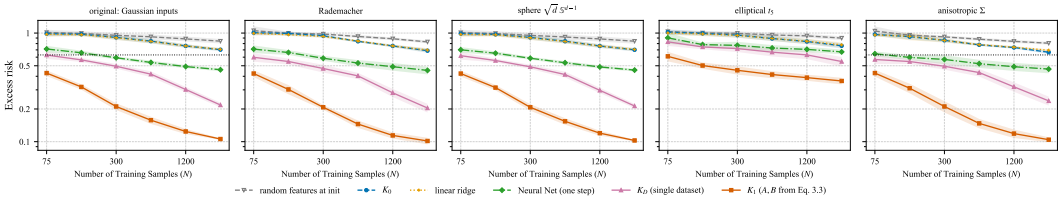


Figure 5: Stress test comparing different input distribution. Distributions are also allowed to violated Assumption 2.1.

Alignment stress test. Same measurement as Appendix D.3 at $d = 300$ with $N = d$ points, for k_0, the network ($m = 400$, $n = 8d$), k_D, and k_1 with the constants of Eq. (3.3) for the scenario's link and noise. The step size for each level p is the one that gives $B = 5d^{p/10}$ in the original scenario; it is not retuned, so in the other scenarios B follows Eq. (3.3). Since $\mathbb{E}\langle\omega, w^*\rangle^2 \neq 1/d$ for anisotropic inputs, $Y_2(Z)$ is centered by its sample mean instead of $1/d$, which for rotation-invariant inputs differs from the original definition by $O(N^{-1/2})$. Seed $50000 + 10^4 t + d$ for trial t.

D.6 Implementation, Software Stack and Hardware

A repository cointaining the code for the experiments can be found in `github.com/joaolobo-git/rkhs-feature-learning`.

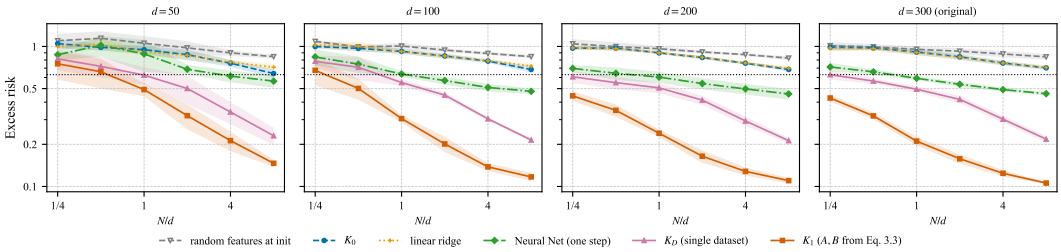


Figure 6: Stress test comparing different ambient dimensions.

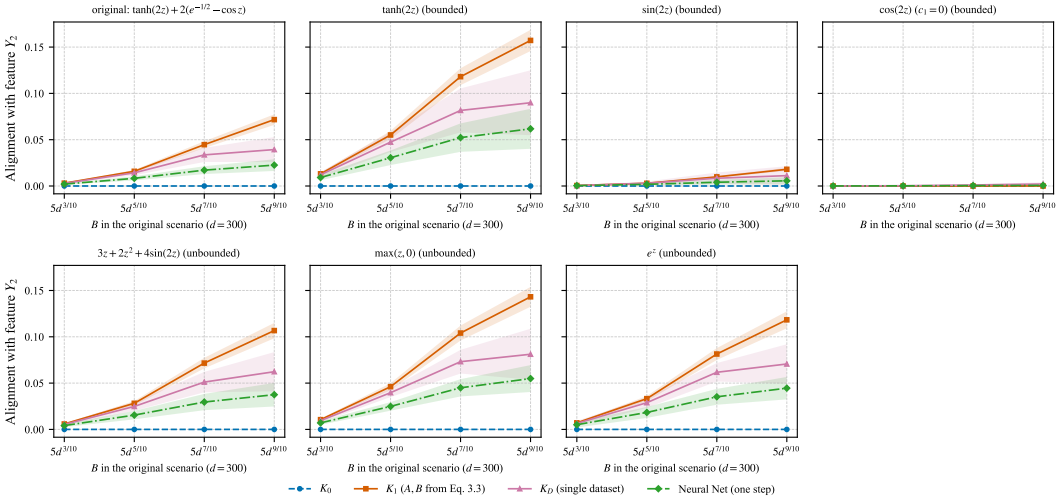


Figure 7: Stress test comparing variations of link functions. The functions are allowed to violate Assumption 2.1 and are marked as bounded versus unbounded respectively.

The experimental framework was implemented in Python (v3.10.22). We use PyTorch (v2.11.0) (Paszke et al., 2019) for the network and its gradient step, computed by automatic differentiation in double precision. We utilized the NumPy (v1.26.4) (Harris et al., 2020) and SciPy (v1.15.3) (Virtanen et al., 2020) libraries for numerical linear algebra operations, specifically for the eigendecomposition (`scipy.linalg.eigh`) of the kernel matrices, which gives both the lead eigenvectors and the cross-validated ridge solutions, and for the KRR implementation (`numpy.linalg.solve`). Lastly, for K-fold validation we used sklearn's (v1.4.1.post1) `model_selection.KFold` object (Pedregosa et al., 2011). All experiments were conducted on Apple M4 chip and 16 GB of unified memory, on CPU only during the course of a few hours.

NeurIPS Paper Checklist

1. **Claims**

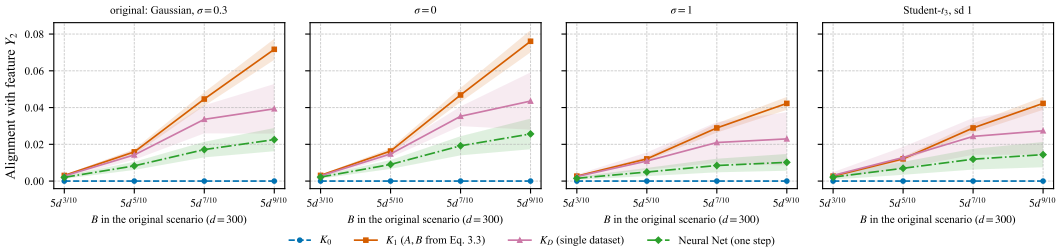


Figure 8: Stress test comparing different levels of noise.

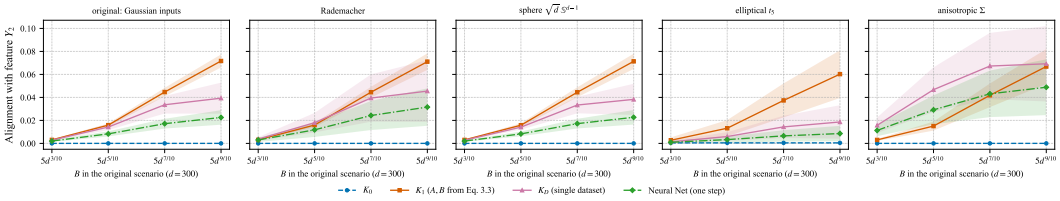


Figure 9: Stress test comparing different input distribution. Distributions are also allowed to violated Assumption 2.1.

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The claims and main results accurately reflect the contributions and transparently discuss the scope, highlighting when the implications hold.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper discusses the limitations of the theoretical framework and setup. While we analyze the scope and implications of this, we explicitly acknowledge the constraints of the specific assumptions used in the results.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Every result presented completely states the assumptions necessary and is accompanied by a complete (correct) proof.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We detail all parameters used to obtain the experimental results. Code will be provided with the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code used to execute the experiments will be provided with the paper and distributed via a code sharing platform (GitHub).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all parameters and details used in the experiments. Full details are available in the code accompanied by the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We illustrate the variance over all experiments with shaded areas around the experimental curves.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments may be reproduced in a personal laptop.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The theoretical results conform with the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Because the work is purely theoretical it does not have direct broader societal impacts to discuss.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Experiments are performed on random synthetic data.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: Experiments are performed on random synthetic data.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Experiments are performed on random synthetic data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects or crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The development of the research does not involve LLMs as a core research method, being used only for editorial purposes.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.