HiLD 2026: 4th Workshop on High-dimensional Learning Dynamics 1–59

How does feature learning change the function space evolution?

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

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 precisely characterize how the function space spanned by the features of a two-layer neural network evolves during gradient descent training. We prove that, in the high-dimensional proportional regime, after a large gradient step the post-update feature distribution is well approximated by a target-dependent spiked Gaussian covariance. This induces a data-adaptive kernel that reshapes the function space and modifies its spectral structure. Our analysis reveals that feature learning can be interpreted as a distributional transformation in either parameter space or input space, equivalently as the introduction of a target-dependent kernel. In particular, 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 precise 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

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

© .

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

In this work, we provide a precise characterization of the function-space evolution induced by a single large gradient descent step in a two-layer network trained on a Gaussian single-index model.

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 setting 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). 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)

*(**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 $\zeta \in [1/2, 1)$.

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 $\boldsymbol{a}^0 \sim \mathcal{N}(0, \frac{1}{m}\boldsymbol{I}_m)$. 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 $\mu_1 := \mathbb{E}_{z\sim\mathcal{N}(0,1)}[g'(z)] \neq 0$.

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). Defining $\bar{\boldsymbol{A}} := \frac{\mu_1\eta}{\sqrt{m}}\frac{\boldsymbol{X}^\top\boldsymbol{y}\boldsymbol{a}^\top}{n}$, the first-layer weight matrix admits the description $\boldsymbol{W}_t = \boldsymbol{W}_0 + \bar{\boldsymbol{A}} + \boldsymbol{E}_t$ where $\boldsymbol{E}_t$ collects asymptotically negligible fluctuations, satisfying $\|\boldsymbol{E}_t\|_{\mathrm{op}} = \widetilde{\mathcal{O}}(1/\sqrt{d})$. It holds in the regime $\eta = \Theta(\sqrt{d})$ with any fixed training step $t \in \mathbb{N}$ (Ba et al., 2022; Wang et al., 2024). Besides, at the first step ($t = 1$), this formulation still holds for an intermediate step-size $\eta = \Theta(d^\zeta)$ with $\zeta \in [1/2, 1)$ in (Moniri et al., 2023) as well as a large step-size $\eta = \Theta(d)$ (Cui et al., 2024; Dandi et al., 2025).

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.

As discussed in Section 2, after one gradient step under the step-size $\eta = \Theta(d^\zeta)$ with $\zeta \in [1/2, 1)$, we know that $(\boldsymbol{W}_1^\top\boldsymbol{x})_j = \langle\boldsymbol{w}_j^0, \boldsymbol{x}\rangle + \langle\boldsymbol{v}_j, \boldsymbol{x}\rangle$ for vectors $\boldsymbol{v}_j \in \mathbb{R}^d$ deriving from the deterministic update $\bar{\boldsymbol{A}} = \frac{\mu_1\eta}{\sqrt{m}}\frac{\boldsymbol{X}^\top\boldsymbol{y}\boldsymbol{a}^\top}{n}$. Hence the new feature map $\phi(\boldsymbol{x}, \boldsymbol{w} + \boldsymbol{v})$ induces the following spiked conjugate kernel, where μ is the Gaussian measure:

$$k_1^*(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathcal{W}} \phi(\boldsymbol{x}, \boldsymbol{w} + \boldsymbol{v})\, \phi\left(\boldsymbol{x}', \boldsymbol{w} + \boldsymbol{v}\right) \mathrm{d}\mu(\boldsymbol{w})\,. \tag{3.1}$$

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.

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 all weights are independent, $\boldsymbol{z}$ has zero mean, and, conditioned on the dataset (or equivalently conditioned on $\boldsymbol{\varsigma}$) we have $\boldsymbol{z} \mid \boldsymbol{\varsigma} \sim \mathcal{N}\Big(0,\ \frac{1}{d}\boldsymbol{I}_d + \frac{\mu_1^2 \eta^2}{n^2 m^2}\boldsymbol{\varsigma}\boldsymbol{\varsigma}^\top\Big)$. Therefore, conditionally on $\boldsymbol{\varsigma}$, $\boldsymbol{z}$ is a Gaussian vector with a spiked covariance along the random direction $\boldsymbol{\varsigma}$. Without it, the distribution of $\boldsymbol{z}$ is complex with covariance matrix given by

$$\frac{1}{d}\boldsymbol{\Gamma} := \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \underbrace{\left(1 + \frac{\eta^2 \lambda_{\alpha,\beta}}{d^2}\mathbb{E}[f^*(\boldsymbol{x})^2] + \frac{\eta^2 \lambda_{\alpha,\beta}}{d^2}\sigma_\varepsilon^2\right)}_{:=A} \frac{\boldsymbol{I}_d}{d} + \underbrace{\left[\frac{\eta^2}{d}\frac{\mu_1^4}{\alpha\beta^2}\left(\alpha - \frac{1}{d} + \frac{2s}{\mu_1^2 d}\right)\right]}_{:=B} \frac{\boldsymbol{w}^*(\boldsymbol{w}^*)^\top}{d}\,, \tag{3.3}$$

where $\lambda_{\alpha,\beta} := \frac{\mu_1^2}{\alpha\beta^2}$, α, β are constants as defined in Eq. (2.2), μ_1 is the first Hermite coefficient of g 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 to Appendix A.1. See the discussion on when $s \gtreqless 0$ and some examples in Appendix A.2. Note that B has three terms and α will dominate in the high-dimensional asymptotic regime, thus we can ensure $\boldsymbol{\Gamma}$ to be positive definite in the asymptotic regime.

Feature learning as an additional target-dependent kernel: To tractably analyze this we present our main theorem to approximate the true kernel. The proof of this can be found in Appendix C.1.

Theorem 3.1 *Denote $\boldsymbol{\varsigma} := \sum_{i=1}^n y_i \boldsymbol{x}_i$ and $\boldsymbol{z} := \boldsymbol{w} + \frac{\mu_1 \eta}{n\sqrt{m}} a \boldsymbol{\varsigma}$ 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

$$\|T_1^* - T_1\|_{HS} = \|k_1^* - k_1\|_{L^2(\rho_{\mathcal{X}})\times L^2(\rho_{\mathcal{X}})} = \mathcal{O}\left(\frac{\eta^4 \ln^3 d}{d^5}\right).$$

Moreover, the spectral decomposition of k_0 with respect to $\nu := (\rho_{\mathcal{X}} \circ \boldsymbol{\Gamma}^{-1/2})$ is the same as the one for k_1 with respect to $\rho_{\mathcal{X}}$.

Thanks to Theorem 3.1, we study k_1, to investigate its spectrum and dynamics. As a preliminary step, Theorem D.1 establishes that T_1 does not prematurely deactivate features by forcing eigenvalues to zero. The theorem and its proof are provided in Appendix D.1. Next, we use Theorem D.2 (see proof in Appendix D.2) to isolate the effect of the spike onto the kernel. With this, we can perform a first order expansion of the kernel k_1 to obtain

$$k_1(\boldsymbol{x}, \boldsymbol{x}') = k_0^{(A)}(\boldsymbol{x}, \boldsymbol{x}') + \frac{B}{2d}S(\boldsymbol{x}, \boldsymbol{x}') + R(\boldsymbol{x}, \boldsymbol{x}')\,, \tag{3.5}$$

where $k_0^{(A)}(\boldsymbol{x},\boldsymbol{x}') := \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{A}{d}\boldsymbol{I}_d)}[\sigma(\langle\boldsymbol{w},\boldsymbol{x}\rangle)\sigma(\langle\boldsymbol{w},\boldsymbol{x}'\rangle)]$, $S := \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{A}{d}\boldsymbol{I}_d)}[D^{(2)}_{\boldsymbol{w}^*}\{\sigma(\langle\boldsymbol{w},\boldsymbol{x}\rangle)\sigma(\langle\boldsymbol{w},\boldsymbol{x}'\rangle)\}]$ and R is a remainder term. In Appendix D.3, we show how to perform this expansion for the kernel and give the definition of the remainder as well as explicitly calculate the first term S. Furthermore, in Appendix D.4, we prove Lemma D.3 to ensure R is dominated by the first terms when $\eta = \Theta(d^\zeta)$ with $\zeta < 1$.

According to Moniri et al. (2023), the feature matrix receives an increasing number of non-linear "spikes" as $\zeta \to 1$. In our context, every higher order derivative term introduces non-linear projections onto $\boldsymbol{w}^*$, so we expect a similar effect on the kernel as a function of ζ. Even though the importance of these terms grows, S remains the dominant term in every scenario where $\zeta < 1$.

The ReLU case: We apply the tools from the previous discussion to analyze the concrete case of the ReLU activation. When $\sigma(t) = \max(0,t)$, we have $k_0^{(A)} = Ak_0$ and the integral operator of S can be approximated by an operator with explicit actions in the function space as detailed by Lemma D.5 with proof in Appendix D.6. Using this and the original eigenfunctions of T_0 (see Lemma D.4 and proof in Appendix D.5), we are able to characterize the top eigenfunctions of T_1 with Theorems 3.2 and 3.3. Their proofs are detailed in Appendix C.2 and Appendix C.3, respectively.

Theorem 3.2 (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 $\lambda_{\psi_\perp}(T_1) = A\lambda_{\psi_\perp}(T_0)$.*

Theorem 3.3 (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*

$$\widetilde{\lambda} = A\lambda_{\max}(T_0) + \frac{B}{2\pi d} + o_d\left(\frac{|B|}{d}\right), \quad \widetilde{\Psi}(\boldsymbol{x}) = \frac{1}{\sqrt{N}}\frac{\|\boldsymbol{x}\|}{\sqrt{d}}\Big[Y_0(\boldsymbol{\omega}) + \tau Y_2(\boldsymbol{\omega},\boldsymbol{w}^*)\Big], \tag{3.6}$$

where N 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 $|\lambda_{\max}(T_1) - \widetilde{\lambda}| = o_d\left(\frac{|B|}{d}\right)$.

4. Conclusion

In this work we studied how deterministic first-layer weight updates in a two-layer network reshape the induced function space. Using a data-dependent kernel, we showed that such updates approximately correspond to a distribution shift of the original kernel, mixing eigenfunctions and selectively rescaling eigenvalues to favor functions aligned with the target, and the strenght of this transformation is governed by the step size. We also perform numerical experiments for validation, which can be found in Appendix E.

References

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, pages 37932–37946, 2022.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(1):629–681, 2017a.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(1):714–751, 2017b.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2933–2943, 2019.

Hugo Cui, Luca Pesce, Yatin Dandi, Florent Krzakala, Yue Lu, Lenka Zdeborova, and Bruno Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9662–9695. PMLR, 21–27 Jul 2024.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, 25(349):1–65, 2024.

Yatin Dandi, Luca Pesce, Hugo Cui, Florent Krzakala, Yue Lu, and Bruno Loureiro. A random matrix theory perspective on the spectrum of learned features and asymptotic generalization capabilities. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 2224–2232. PMLR, 03–05 May 2025.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

Zhenyu Liao and Romain Couillet. On the spectrum of random features maps of high dimensional data. In *International Conference on Machine Learning*, pages 3063–3071, 2018.

E. McMahon. An extension of price's theorem (corresp.). *IEEE Transactions on Information Theory*, 10(2):168–168, 1964. doi: 10.1109/TIT.1964.1053656.

Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. *arXiv preprint arXiv:2310.07891*, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

R. Price. A useful theorem for nonlinear devices having gaussian inputs. *IRE Transactions on Information Theory*, 4(2):69–72, 1958. doi: 10.1109/TIT.1958.1057444.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2007.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer, 2001.

Taiji Suzuki. 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*, 2019.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay

Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Zhichao Wang, Denny Wu, and Zhou Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. In *Conference on Learning Theory*, pages 4891–4957. PMLR, 2024.

Contents

A Properties of the Spiked Covariance Matrix **10**
A.1 Derivation of the second moment matrix 10
A.2 Characterization of the coefficient s in the new covariance matrix 11

B Technical Lemmas and their proofs **13**
B.1 Characteristic-function comparison for activation products 13
B.2 Closed form of the infinite sums 16
B.3 Concentration for Gaussian integrals 18

C Proofs of the main results **20**
C.1 Proof of Theorem 3.1 20
C.2 Proof of Theorem 3.2 28
C.3 Proof of Theorem 3.3 29

D Deferred Results **36**
D.1 Spectral stability of the integral operator T_1 36
D.2 The spiked covariance expansion framework 40
D.3 Applying the expansion to the kernel k_1 43
D.4 The remainder term R is dominated 44
D.5 Eigenfunctions of the isotropic ReLU kernel 49
D.6 Approximating the first-order term of the ReLU expansion 49

E Experimental details **56**
E.1 Experiments 56
E.2 Different models 57
E.3 Metrics and Significance 57
E.4 Alignment experiment 58
E.5 Generalization performance 58
E.6 Implementation, Software Stack and Hardware 59

Appendix 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

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

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

$$\mathbb{E}[f^*(\boldsymbol{x})\boldsymbol{x}]\mathbb{E}[f^*(\boldsymbol{x})\boldsymbol{x}^\top] = \mathbb{E}[g'(\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle)]^2 \boldsymbol{w}^*(\boldsymbol{w}^*)^\top = \mu_1^2 \boldsymbol{w}^*(\boldsymbol{w}^*)^\top .$$

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

$$\begin{aligned}
\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] &= \frac{1}{d}\left(\boldsymbol{I}_d + \frac{\eta^2\lambda_{\alpha,\beta}}{d^2}\Big\{2s\boldsymbol{w}^*(\boldsymbol{w}^*)^\top + \mathbb{E}[f^*(\boldsymbol{x})^2]\boldsymbol{I}_d\Big\} + \frac{\eta^2\lambda_{\alpha,\beta}}{d^2}\sigma_\varepsilon^2\boldsymbol{I}_d + \frac{\eta^2\lambda_{\alpha,\beta}}{d}\Big(\alpha - \frac{1}{d}\Big)\mu_1^2\boldsymbol{w}^*(\boldsymbol{w}^*)^\top\right) \\
&= \frac{1}{d}\left(\boldsymbol{I}_d + \frac{\eta^2\lambda_{\alpha,\beta}}{d^2}\mathbb{E}[f^*(\boldsymbol{x})^2]\boldsymbol{I}_d + \frac{\eta^2\lambda_{\alpha,\beta}}{d^2}\sigma_\varepsilon^2\boldsymbol{I}_d + \frac{\eta^2\lambda_{\alpha,\beta}}{d}\Big(\alpha - \frac{1}{d} + \frac{2s}{\mu_1^2 d}\Big)\mu_1^2\boldsymbol{w}^*(\boldsymbol{w}^*)^\top\right) \\
&= \frac{1}{d}\left(\boldsymbol{I}_d + \frac{\eta^2\lambda_{\alpha,\beta}}{d^2}\mathbb{E}[f^*(\boldsymbol{x})^2]\boldsymbol{I}_d + \frac{\eta^2\lambda_{\alpha,\beta}}{d^2}\sigma_\varepsilon^2\boldsymbol{I}_d + \frac{\eta^2}{d}\frac{\mu_1^4}{\alpha\beta^2}\Big(\alpha - \frac{1}{d} + \frac{2s}{\mu_1^2 d}\Big)\boldsymbol{w}^*(\boldsymbol{w}^*)^\top\right) .
\end{aligned}$$

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

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

Also note that g is bounded, therefore

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

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

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

Appendix B. Technical Lemmas and their proofs

B.1. Characteristic-function comparison for activation products

Lemma B.1 *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an L-Lipschitz function, and define $F(u, v) = \sigma(u)\sigma(v)$. Let μ and ν be probability measures on $\mathbb{R}^2$ with characteristic functions ϕ_μ and ϕ_ν such that $|\phi_\mu(\xi) - \phi_\nu(\xi)| \leq \delta |\xi|^2 e^{-\xi^\top \Sigma \xi}$ for a positive semidefinite matrix Σ such that $\operatorname{Tr}(\Sigma) \leq C$. Furthermore, assume that μ and ν satisfy $P(|x| > R) \leq C_0 e^{-\alpha R}$ for some $\alpha > 0$. Then*

$$\left| \int F(x) d\mu(x) - \int F(x) d\nu(x) \right| = \mathcal{O}\left(\delta \log^3(1/\delta)\right)$$

Proof First, we note that because σ is L-Lipschitz, its growth is at most linear: $|\sigma(x)| \leq |\sigma(0)| + L|x|$. Consequently, the product $F(u, v)$ has at most quadratic growth

$$|F(u, v)| \leq (|\sigma(0)| + L|u|)(|\sigma(0)| + L|v|) \leq C_1(1 + |x|^2)$$

Let $\eta : \mathbb{R}^2 \to \mathbb{R}$ be a smooth (C^∞) function such that $0 \leq \eta(x) \leq 1$ for all x, $\eta(x) = 1$ for all $|x| \leq 1$ and $\eta(x) = 0$ for all $|x| \geq 2$.

Because $\eta(x)$ is constant (either 1 or 0) outside the region where $1 < |x| < 2$, its gradient $\nabla \eta(x)$ is exactly zero everywhere except inside that closed, compact set. As a continuous function inside a compact set, and given this property of the gradient, there exists some

absolute, finite constant C such that

$$|\nabla\eta(x)| \leq C \quad \text{for all } x \in \mathbb{R}^2$$

Now, we define the specific cutoff function $\chi_R(x)$ by

$$\chi_R(x) = \eta\left(\frac{x}{R}\right).$$

That makes χ_R smooth and $\chi_R(x) = 1$ if $|x| \leq R$, $\chi_R(x) = 0$ if $|x| > 2R$, otherwise $1 \leq \chi_R(x) \leq 0$ when $R < |x| < 2R$.

By the Chain Rule, we have

$$\nabla\chi_R(x) = \nabla\left[\eta\left(\frac{x}{R}\right)\right] = \frac{1}{R}(\nabla\eta)\left(\frac{x}{R}\right)$$

and therefore

$$|\nabla\chi_R(x)| = \left|\frac{1}{R}(\nabla\eta)\left(\frac{x}{R}\right)\right| = \frac{1}{R}\left|(\nabla\eta)\left(\frac{x}{R}\right)\right| \leq \frac{C}{R}.$$

Using that $\chi_R(x)$ supported on B_{2R} with $|\nabla\chi_R| \leq \frac{C}{R}$, we define $F_R(x) = F(x)\chi_R(x)$. We split the integral into the truncated core and the tail

$$\int F(x)d\mu(x) = \int F(x)\chi_R(x)d\mu(x) + \int_{|x|>R} F(x)(1-\chi_R(x))d\mu(x)$$

Using the quadratic growth bound $|F(x)| \leq C_1(1+|x|^2)$ and the sub-exponential tail of μ, the tail error evaluates to

$$\int_{|x|>R} F(x)(1-\chi_R(x))d\mu(x) \leq \int_{|x|>R} C_1(1+|x|^2)d\mu(x) \leq C_2R^2e^{-\alpha R}$$

The total error is bounded by the core difference plus the tail errors

$$\left|\int F(x)d\mu(x) - \int F(x)d\nu(x)\right| \leq \left|\int F(x)\chi_R(x)d\mu(x) - \int F(x)\chi_R(x)d\nu(x)\right| + 2C_2R^2e^{-\alpha R}.$$

On the core term, if we let $F_R(x) := F(x)\chi_R(x)$, we use Parseval's identity to obtain

$$\int F(x)\chi_R(x)d\mu(x) - \int F(x)\chi_R(x)d\nu(x) = \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2} \widehat{F_R}(\xi)(\phi_\mu(-\xi) - \phi_\nu(-\xi))d\xi.$$

Next, we look at the distributional gradient ∇F_R which can be written as

$$\nabla F_R = \chi_R\nabla F + F\nabla\chi_R,$$

and we will bound the L^∞ norm of both terms on the support ball B_{2R}.

The gradient is given by $\nabla F = (\sigma'(u)\sigma(v), \sigma(u)\sigma'(v))$. Since σ is L-Lipschitz, $|\sigma'| \leq L$ almost everywhere. Also, on the ball of radius $2R$, the activation function is bounded by

$|\sigma(u)| \leq C_\sigma R$ for some absolute constant $C_\sigma > 0$. Therefore, $|\nabla F| = \mathcal{O}(R)$. Furthermore, $|F|$ grows quadratically, and since $|\nabla \chi_R| \leq \frac{C}{R}$, their product is $\mathcal{O}(R)$. Thus, the maximum value of the gradient is bounded by $\|\nabla F_R\|_{L^\infty} \leq C_3 R$.

If the symbol $\mathcal{F}$ denotes the Fourier transform operator, by the properties of the Fourier transform, we have that

$$\mathcal{F}\{\nabla F_R\} = i\xi \widehat{F_R}(\xi)\,.$$

Next, the maximum absolute value of any Fourier transform is always bounded by

$$|\mathcal{F}\{\nabla F_R\}| \leq \|\nabla F_R\|_{L^1}$$

and the L^1 norm is bounded by the L^∞ norm times the area of the support ball, which is $\pi(2R)^2$, thus

$$\|\nabla F_R\|_{L^1} \leq \|\nabla F_R\|_{L^\infty} \pi(2R)^2 \leq (C_3 R)\pi(2R)^2 = C_4 R^3\,.$$

Combining these facts together we get

$$\mathcal{F}\{\nabla F_R\} = |\xi||\widehat{F_R}(\xi)| \leq \|\nabla F_R\|_{L^1} \leq C_4 R^3$$

and finally arrive at

$$|\widehat{F_R}(\xi)| \leq \frac{C_4 R^3}{|\xi|}\,.$$

Substituting this new bound and the anisotropic characteristic function estimate into the Parseval integral gives

$$\left| \int F(x)\chi_R(x) d\mu(x) - \int F(x)\chi_R(x) d\nu(x) \right| \leq \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \left(\frac{C_4 R^3}{|\xi|} \right) \left(\delta |\xi|^2 e^{-\frac{1}{2}\xi^\top \Sigma \xi} \right) d\xi.$$

Canceling one power of $|\xi|$ yields

$$\left| \int F(x)\chi_R(x) d\mu(x) - \int F(x)\chi_R(x) d\nu(x) \right| \leq C_5 \delta R^3 \int_{\mathbb{R}^2} |\xi| e^{-\frac{1}{2}\xi^\top \Sigma \xi} d\xi.$$

Now diagonalize the covariance matrix:

$$\Sigma = Q^\top \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} Q,$$

with orthogonal Q, and perform the rotation $v = Q\xi$. Since orthogonal transformations preserve Lebesgue measure and Euclidean norm,

$$\int_{\mathbb{R}^2} |\xi| e^{-\frac{1}{2}\xi^\top \Sigma \xi} d\xi = \int_{\mathbb{R}^2} |v| e^{-\frac{1}{2}(\lambda_1 v_1^2 + \lambda_2 v_2^2)} dv.$$

Because Σ is positive semidefinite and

$$\mathrm{Tr}(\Sigma) \leq C,$$

the Gaussian factor provides exponential decay in every nondegenerate direction. Even when one eigenvalue degenerates, the integral remains effectively one-dimensional in the degenerate direction and therefore finite. Consequently,

$$\int_{\mathbb{R}^2} |v| e^{-\frac{1}{2}(\lambda_1 v_1^2 + \lambda_2 v_2^2)} dv \leq C',$$

for some absolute constant $C' > 0$

Therefore,

$$\left| \int F(x)\chi_R(x) d\mu(x) - \int F(x)\chi_R(x) d\nu(x) \right| \leq C_6 \delta R^3.$$

Thus, the total bound as a function of the truncation radius R is

$$\left| \int F(x) d\mu(x) - \int F(x) d\nu(x) \right| \leq C_6 \delta R^3 + 2C_2 R^2 e^{-\alpha R}.$$

We balance the terms by setting the tail decay equal to the δ parameter

$$e^{-\alpha R} = \delta \implies R = \frac{1}{\alpha} \log(1/\delta).$$

Plugging this R back into the total bound we obtain the result

$$\left| \int F d\mu - \int F d\nu \right| \leq C_6 \delta \left(\frac{1}{\alpha} \log(1/\delta) \right)^3 + 2C_2 \delta \left(\frac{1}{\alpha} \log(1/\delta) \right)^2 = \mathcal{O}\left(\delta \log^3(1/\delta) \right).$$

■

B.2. Closed form of the infinite sums

Lemma B.2 *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.3. Concentration for Gaussian integrals

Lemma B.3 *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 $\mathcal{A}^c$:

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

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. ■

Appendix C. Proofs of the main results

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 $o(f/d)$.

C.1. Proof of Theorem 3.1

Before we start, we note that our result builds upon macroscopic results of order $\mathcal{O}\left(\frac{|B|}{d}\right)$, by repeatedly showing the trailing terms of our approximations decay with rate $o_d\left(\frac{|B|}{d}\right)$. 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 $\mathcal{O}\left(\frac{\eta^4 \ln^3 d}{d^5}\right) = \mathcal{O}\left(\frac{|B|^2 \ln^3 d}{d^3}\right) \subset o_d\left(\frac{|B|}{d}\right)$. Note that this result also guarantees an asymptotic approximation in the most aggressive learning rate regime where $\eta = \Theta(d)$.

Proof

We split the proofs into several parts. First, we show that the characteristic function of the true distribution is sufficiently close to the one from the Gaussian distribution governed by $\boldsymbol{\Gamma}$. Then, we use that to bound $\|k_1^* - k_1\|_{L^2(\rho_{\mathcal{X}})\times L^2(\rho_{\mathcal{X}})}$, which immediately implies the operators are also close in norm. Since these objects live in the $L^2(\rho_{\mathcal{X}})$ space, the inputs are inherently unbounded, thus we further 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.

Convergence of the true distribution to the spiked Gaussian: Unconditionally, $\boldsymbol{z}$ is not Gaussian in general (it is a mixture of sub-exponential random variables), but conditional on $\boldsymbol{\varsigma}$ it is Gaussian. To be specific, if we let $c := \frac{\mu_1 \eta}{n\sqrt{m}}$, with $a \sim \mathcal{N}(0, \frac{1}{m})$ and $\boldsymbol{w} \sim \mathcal{N}(0, \frac{1}{d}\boldsymbol{I}_d)$, condition on $\boldsymbol{\varsigma} = \sum y_i \boldsymbol{x}_i$, we have

$$\boldsymbol{z} \mid \boldsymbol{\varsigma} \sim \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\varsigma})), \quad \boldsymbol{\Sigma}(\boldsymbol{\varsigma}) = \frac{1}{d}\boldsymbol{I}_d + \frac{c^2}{m}\boldsymbol{\varsigma}\boldsymbol{\varsigma}^\top .$$

So conditioning on $\boldsymbol{\varsigma}$, $\boldsymbol{z}$ is Gaussian with an anisotropic rank-one spike along the direction $\boldsymbol{\varsigma}$. Therefore, if we fix a test vector $\boldsymbol{t} \in \mathbb{R}^d$ and consider the projection $\langle \boldsymbol{z}, \boldsymbol{t} \rangle$, conditionally on $\boldsymbol{\varsigma}$ we have

$$\langle \boldsymbol{z}, \boldsymbol{t} \rangle \mid \boldsymbol{\varsigma} \sim \mathcal{N}\left(0, \frac{\|\boldsymbol{t}\|^2}{d} + \frac{c^2}{m}\langle \boldsymbol{\varsigma}, \boldsymbol{t} \rangle^2\right) .$$

And consequently, the unconditional characteristic function is given by

$$\phi_{\langle \boldsymbol{z},\boldsymbol{t}\rangle}(u) = \exp\left(-\frac{u^2}{2}\frac{\|\boldsymbol{t}\|^2}{d}\right)\mathbb{E}_{\boldsymbol{\varsigma}}\left[\exp\left(-\frac{u^2}{2}\frac{c^2\langle \boldsymbol{\varsigma},\boldsymbol{t}\rangle^2}{m}\right)\right].$$

Now, to investigate the concentration of this variable, we look at the difference

$$\left|\mathbb{E}_{\boldsymbol{\varsigma}}\left[\exp\left(-\frac{u^2}{2}\frac{c^2\langle \boldsymbol{\varsigma},\boldsymbol{t}\rangle^2}{m}\right)\right] - \exp\left(-\frac{u^2}{2}\frac{c^2\mathbb{E}_{\boldsymbol{\varsigma}}[\langle \boldsymbol{\varsigma},\boldsymbol{t}\rangle^2]}{m}\right)\right|.$$

If we define $X = \frac{c^2}{m}\langle \boldsymbol{\varsigma},\boldsymbol{t}\rangle^2$, $\mu = \frac{c^2}{m}\mathbb{E}_{\boldsymbol{\varsigma}}[\langle \boldsymbol{\varsigma},\boldsymbol{t}\rangle^2]$ and $\alpha = \frac{u^2}{2}$, we perform a second-order Taylor expansion of the function $f(x) = e^{-\alpha x}$ around μ. By Taylor's theorem, there exists some ξ between X and μ such that

$$e^{-\alpha X} = e^{-\alpha\mu} - \alpha e^{-\alpha\mu}(X-\mu) + \frac{\alpha^2 e^{-\alpha\xi}}{2}(X-\mu)^2$$

Taking the expectation of both sides yields:

$$\mathbb{E}[e^{-\alpha X}] = e^{-\alpha\mu} - \alpha e^{-\alpha\mu}\mathbb{E}[X-\mu] + \frac{\alpha^2}{2}\mathbb{E}[e^{-\alpha\xi}(X-\mu)^2]$$

Because $\mu = \mathbb{E}[X]$, the first-order term cancels out exactly ($\mathbb{E}[X-\mu] = 0$), leaving:

$$\mathbb{E}[e^{-\alpha X}] - e^{-\alpha\mu} = \frac{\alpha^2}{2}\mathbb{E}[e^{-\alpha\xi}(X-\mu)^2]$$

Since $\alpha = u^2/2 > 0$ and $X \geq 0$ (as it is a scaled squared projection), it follows that $\xi \geq 0$ and therefore $e^{-\alpha\xi} \leq 1$. Taking the absolute value, we can bound the difference directly by the variance of X

$$|\mathbb{E}[e^{-\alpha X}] - e^{-\alpha\mu}| \leq \frac{\alpha^2}{2}\mathbb{E}[(X-\mu)^2] = \frac{\alpha^2}{2}\mathrm{Var}(X)$$

Substituting our definitions back in, we have $\frac{\alpha^2}{2}\mathrm{Var}(X) = \frac{u^4}{8}\mathrm{Var}(X)$. Given that $\frac{c^2}{m} = \mathcal{O}(\frac{\eta^2}{d^4})$, the variance of the sub-exponential variable yields the following bound

$$|\mathbb{E}[e^{-\alpha X}] - e^{-\alpha\mu}| \leq \mathcal{O}\left(\frac{\eta^4}{d^5}\right) \tag{C.1}$$

Bounding the kernel difference: We let $p_{\boldsymbol{\Gamma}}$ be the density function of the Gaussian distribution $\mathcal{N}(0, \frac{1}{d}\boldsymbol{\Gamma})$ and $p_1^* = \mathbb{E}_{\boldsymbol{\varsigma}}[p_{\boldsymbol{I}_d}(\boldsymbol{z} \mid \boldsymbol{\varsigma})]$ be the true density function from the non-Gaussian distribution followed by z.

Denote $G(\boldsymbol{w}) := \sigma(\langle \boldsymbol{w},\boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{w},\boldsymbol{x}'\rangle)$, the true kernel after the deterministic update is given by

$$k_1^*(\boldsymbol{x},\boldsymbol{x}') = \int G(\boldsymbol{w})p_1^*(\boldsymbol{w})\mathrm{d}\boldsymbol{v}\,,$$

while the Gaussian kernel k_1 is

$$k_1(\boldsymbol{x}, \boldsymbol{x}') = \int G(\boldsymbol{w}) p_{\boldsymbol{\Gamma}}(\boldsymbol{w}) \mathrm{d}\boldsymbol{v} \,.$$

We define a 2-dimensional vector $\boldsymbol{v}$ representing the two projections

$$\boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} \langle \boldsymbol{w}, \boldsymbol{x} \rangle \\ \langle \boldsymbol{w}, \boldsymbol{x}' \rangle \end{bmatrix} \in \mathbb{R}^2 \,,$$

then we can write G in terms of this 2D variable: $G(\boldsymbol{v}) = \sigma(v_1)\sigma(v_2)$. The error between both kernels is the difference in expectations over this 2D plane

$$|k_1^*(\boldsymbol{x}, \boldsymbol{x}') - k_1(\boldsymbol{x}, \boldsymbol{x}')| = \left| \int_{\mathbb{R}^2} G(\boldsymbol{v}) p_1^*(\boldsymbol{v}) \mathrm{d}\boldsymbol{v} - \int_{\mathbb{R}^2} G(\boldsymbol{v}) p_{\boldsymbol{\Gamma}}(\boldsymbol{v}) \mathrm{d}\boldsymbol{v} \right|$$

Taking the Fourier transforms, we map this into the 2D frequency domain. In this context, the frequency variable is $\boldsymbol{u} = (u_1, u_2)$ and the error is given by

$$|k_1^*(\boldsymbol{x}, \boldsymbol{x}') - k_1(\boldsymbol{x}, \boldsymbol{x}')| = \frac{1}{(2\pi)^2} \left| \int_{\mathbb{R}^2} \widehat{G}(\boldsymbol{u}) [\Phi_1^*(\boldsymbol{u}) - \Phi_{\boldsymbol{\Gamma}}(\boldsymbol{u})] \mathrm{d}\boldsymbol{u} \right| \,,$$

where Φ are the respective characteristic functions of the distributions. We can see that the inner product between $\boldsymbol{u}$ and $\boldsymbol{v}$ gives

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = u_1 \langle \boldsymbol{w}, \boldsymbol{x} \rangle + u_2 \langle \boldsymbol{w}, \boldsymbol{x}' \rangle = \langle \boldsymbol{w}, u_1 \boldsymbol{x} + u_2 \boldsymbol{x}' \rangle \,.$$

Therefore, if we define our test vector as $\boldsymbol{t}_u = u_1 \boldsymbol{x} + u_2 \boldsymbol{x}' \in \mathbb{R}^d$, $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \langle \boldsymbol{w}, \boldsymbol{t}_u \rangle$, the characteristic function for $\boldsymbol{u}$ is given by

$$\Phi(\boldsymbol{u}) = \mathbb{E}[\exp(i \langle \boldsymbol{u}, \boldsymbol{v} \rangle)] = \mathbb{E}[\exp(i \langle \boldsymbol{w}, \boldsymbol{t}_u \rangle)] \,.$$

This implies the characteristic function acting on $\boldsymbol{u}$ is precisely the characteristic function of a 1D projection of $\boldsymbol{w}$, as considered in the bound from Eq. (C.1). Thus, we can plug in the bound and factor the d-dependent estimate out of the integral

$$|\Phi_1^*(\boldsymbol{u}) - \Phi_{\boldsymbol{\Gamma}}(\boldsymbol{u})| \leq \mathcal{O}\left(\frac{\eta^4 \ln^3 d}{d^5} \right) \|\boldsymbol{u}\|^2 \exp\left(-\frac{1}{2} \boldsymbol{u}^T \boldsymbol{\Sigma} \boldsymbol{u} \right) \,,$$

where we write the covariance matrix $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \frac{1}{d} \begin{bmatrix} \|\boldsymbol{x}\|^2 & \langle \boldsymbol{x}, \boldsymbol{x}' \rangle \\ \langle \boldsymbol{x}, \boldsymbol{x}' \rangle & \|\boldsymbol{x}'\|^2 \end{bmatrix} \,.$$

We will 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 to the Hilbert-Schmidt norm and the operator norm of the difference of integral operators. For this we split this proof with a truncation argument considering a bounded domain and then use the Lipschitz property of σ to ensure the tails

decays exponentially. Ultimately, we want analyze the integral

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

so we start controlling the difference inside a bounded set.

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

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

Then, since $\boldsymbol{\Sigma}$ is positive semidefinite and

$$\mathrm{Tr}\,(\boldsymbol{\Sigma}) = \frac{\|\boldsymbol{x}\|^2 + \|\boldsymbol{x}'\|^2}{d} \leq 2R^2\,,$$

and because both weight distributions are sub-exponential, using Lemma B.1, we get that

$$\sup_{\boldsymbol{x},\boldsymbol{x}' \in B_{R,c}\times B_{R,c}} |k_1^*(\boldsymbol{x},\boldsymbol{x}') - k_1(\boldsymbol{x},\boldsymbol{x}')| = \mathcal{O}\left(\frac{\eta^4 \ln^3 d}{d^5}\right)\,.$$

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 $|\sigma(t)| \leq |\sigma(0)| + L_\sigma|t|$. 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

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

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

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

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

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

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

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

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

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

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

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

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

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

and applying our bound gives

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

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

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

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

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

Noting that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Analogously, we have

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

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

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

for the absolute constant $K = \min\left\{\frac{(R-1)^2}{4}, \frac{c'}{2}\right\} > 0$, 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

$$\iint_{B_{R,c}\times B_{R,c}} |k_1^* - k_1|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') = \mathcal{O}\left(\frac{\eta^4 \ln^3 d}{d^5}\right)^2$$

and

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

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

$$\|k_1^* - k_1\|_{L^2(\rho_{\mathcal{X}})\times L^2(\rho_{\mathcal{X}})} = \mathcal{O}\left(\frac{\eta^4 \ln^3 d}{d^5}\right) ,$$

and since

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

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

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

with constants A and B from the original definition satisfying $A + B > |B|$. 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 $\boldsymbol{z} \sim \mathcal{N}\left(0, \frac{1}{d}\boldsymbol{\Gamma}\right)$ we can write

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

and the kernels are related by

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

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

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

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

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

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

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

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

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

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

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

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

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

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

■

C.2. Proof of Theorem 3.2

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.3. Proof of Theorem 3.3

Proof

We have from Lemma D.5 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 3.2 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

$$T_1 f(\boldsymbol{x}) = A T_0 f(\boldsymbol{x}) + \frac{B}{2d} T_S f(\boldsymbol{x}) + T_R f(\boldsymbol{x})$$

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 $\operatorname{span}\{f_1, f_2\}$. To bridge the gap between both operators, we recall the eigenbasis of the isotropic operator given by Lemma D.4, 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

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

$$f_1(\boldsymbol{x}) = rY_0(\boldsymbol{\omega}), \quad f_2(\boldsymbol{x}) = r\widehat{Y}_2(\boldsymbol{\omega}, \boldsymbol{w}^*) + \frac{r}{d}Y_0(\boldsymbol{\omega}).$$

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

$$\varphi = c_0[rY_0(\boldsymbol{\omega})] + c_2[r\widehat{Y}_2(\boldsymbol{\omega}, \boldsymbol{w}^*)]$$

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

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

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, more intuitively, using the Taylor expansion of $\sqrt{x+\epsilon}$ we can write

$$\widetilde{\lambda}_{+} = A\lambda_{\max}(T_0) + \frac{B}{2\pi d} + o_d\left(\frac{|B|}{d}\right), \quad \widetilde{\lambda}_{-} = \mathcal{O}\left(A\lambda_2(T_0) + \frac{B^2}{d^3}\right).$$

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

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

and therefore the normalized approximate eigenfunctions are given by

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

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:

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

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 $\widetilde{\lambda}_+ > \widetilde{\lambda}_-$.

The approximate eigenvalue $\widetilde{\lambda}_+$ dominates every other eigenvalue of T_1: From the previous derivations, if we look at the approximate eigenvalue as functions of the dimension $\lambda_+(d)$, we established that

$$\lim_{d\to\infty}\widetilde{\lambda}_+(d) = A\lambda_{\max}(T_0) > 0$$

Thus, for any chosen $\epsilon > 0$, there exists an integer d_0 such that for all $d > d_0$

$$\widetilde{\lambda}_+(d) > A\lambda_{\max}(T_0) - \epsilon$$

Choose $\epsilon = \frac{A\lambda_{\max}(T_0)}{2}$ such that there exists a d_0 such that for all $d > d_0$

$$\widetilde{\lambda}_+(d) > \frac{A\lambda_{\max}(T_0)}{2}.$$

By Theorem 3.2, we know $\lambda_1(T_1) = A\lambda_1(T_0) + \frac{B}{4d} = \mathcal{O}(d^{-1})$. Thus, there exists a real constant $M > 0$ and an integer d_1 such that for all $d > d_1$

$$\lambda_1(T_1) \leq \frac{M}{d}.$$

Because the ReLU kernel has monotonically decreasing eigenvalues, the linear eigenvalue serves as a strict upper bound for all higher-degree subspaces

$$\lambda_1(T_0) > \lambda_k(T_0) \quad \text{for all } k \geq 2.$$

Furthermore, by Theorem D.1 there exists a radius $R > 0$ and $\varepsilon > 0$ such that for constant $C_R > 0$ independent of the dimension we have

$$\lambda_k(T_1) \leq C_R \lambda_k(T_0) + \varepsilon < 2C_R \lambda_1(T_0)$$

for all $k \geq 1$ which implies

$$\lambda_k(T_1) \leq \frac{2C_R M}{d}\,.$$

Thus, if we can prove $\widetilde{\lambda}_+(d) > \lambda_1(T_1)$, we automatically prove it dominates all $\lambda_k(T_1)$ for $k \geq 1$.

We want to guarantee that $\widetilde{\lambda}_+(d) > \lambda_1(T_1)$ which is true as long as

$$\frac{A\lambda_{\max}(T_0)}{2} > \frac{2C_R M}{d} \implies d > \frac{4C_R M}{A\lambda_{\max}(T_0)}.$$

Define

$$d^* = \max\left(d_0, d_1, \left\lfloor \frac{4C_R M}{A\lambda_{\max}(T_0)} \right\rfloor + 1\right)$$

then for any dimension $d > d^*$, the following holds

$$\widetilde{\lambda}_+(d) > \frac{A\lambda_{\max}(T_0)}{2} > \frac{2C_R M}{d} \geq \lambda_1(T_1) \geq \lambda_{k\geq 2}(T_1).$$

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 Theorem D.1, we have the lower bound

$$c\lambda_{\max}(T_0) \leq \lambda_{\max}(T_1)$$

where $c > 0$ is an absolute constant. We know $\lambda_{\max}(T_0) = \Theta(1)$ thus $c\lambda_{\max}(T_0)$ is bounded away from zero. which gives the lower bound

$$\lambda_{\max}(T_1) \geq c\lambda_{\max}(T_0) = \Omega(1).$$

Now if $\delta := \lambda_{\max}(T_1) - \lambda_1(T_1)$ is the true spectral gap, we know from Theorem 3.2 that $\lambda_1(T_1) = \Theta(1/d)$. Substituting our lower bound for $\lambda_{\max}(T_1)$ into the gap we have

$$\delta \geq c\lambda_{\max}(T_0) - \lambda_1(T_1) = \Omega(1),$$

for a high enough dimension d. The difference between the gaps is given by

$$|\lambda_{\max}(T_1) - \widetilde{\lambda}_+| = |\delta - \widetilde{\delta}|$$

and since $|\lambda_{\max}(T_1) - \widetilde{\lambda}_+| = o_d(|B|/d)$ the triangle inequality implies

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

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. ■

Appendix D. Deferred Results

D.1. Spectral stability of the integral operator T_1

Theorem D.1 *Consider the integral operators T_1 and T_0 defined in Eq. (3.2). Under Theorem 2.1, for any $\varepsilon > 0$ there exists a truncation radius $R > 0$, a constant $C_R > 0$ that depends on R and an absolute constant $c > 0$ such that the j-th eigenvalues of the operators satisfy*

$$c\lambda_j(T_0) \le \lambda_j(T_1) \le C_R\lambda_j(T_0) + \varepsilon\,,$$

for all $k \ge 0$. Consequently, the spectra of both operators exhibit the same asymptotic decay rate.

Proof Since we work with the Gaussian measures, the domain is naturally unbounded, thus we split the proof in cases.

Difference under bounded and unbounded domain Supposing we consider an unbounded integration space $\mathcal{X}$, introduce the bounded set

$$B_R = \{\boldsymbol{y} \in \mathbb{R}^d : \|\boldsymbol{y}\| < R\},$$

and we want to show that outside this set we able to control the fluctuations of the eigenvalues well enough.

Since the Gaussian measures are regular, for a given $\varepsilon > 0$, we can choose $R = R(\varepsilon) > 0$ such that

$$\rho_{\boldsymbol{\Gamma}}(B_R^c) < \varepsilon \quad \rho_{\boldsymbol{I}_d}(B_R^c) < \varepsilon$$

and using this we write

$$T_1 f(\boldsymbol{x}) = T_1^R f(\boldsymbol{x}) + E_1 f(\boldsymbol{x}) = \int_{B_R} k_1(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') \mathbf{1}_{\boldsymbol{B_R}}(\boldsymbol{x}) \mathrm{d}\rho_{\boldsymbol{I}_d}(\boldsymbol{x}') + \int_{B_R^c} k_1(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') \mathrm{d}\rho_{\boldsymbol{I}_d}(\boldsymbol{x}')$$

and similarly $T_0 = T_0^R + E_0$, noting that since T_1, T_0 are compact and self adjoint, T_0^R, T_1^R, E_0, E_1 must also be.

For the bounded domain case, we assume the following result holds: for every $k \geq 0$,

$$c\lambda_k(T_0^R) \leq \lambda_k(T_1^R) \leq C\lambda_k(T_0^R).$$

In fact, since the lower bound of the function r does not depend on the norm, its value is the same for the bounded and unbounded case, thus the lower bound on the eigenvalues comes for free

$$c\lambda_k(T_0) \leq \lambda_k(T_1).$$

Furthermore, since σ is L_σ-Lipschitz, we have $|\sigma(z) - \sigma(0)| \leq L|z|$ and

$$k_1(\boldsymbol{x}, \boldsymbol{x}) \leq 2L^2 \mathbb{E}_{\boldsymbol{w}}[(\boldsymbol{w}^T \boldsymbol{x})^2] + 2\sigma(0)^2.$$

Since for k_1, $\boldsymbol{w} \sim \mathcal{N}(0, \frac{1}{d}\boldsymbol{\Gamma})$, we have

$$k_1(\boldsymbol{x}, \boldsymbol{x}) \leq C\left(1 + \frac{\|\boldsymbol{x}\|^2}{d}\right).$$

for some absolute constant $C > 0$ depending on $\lambda_{\max}(\boldsymbol{\Gamma}), L$ and $\sigma(0)$.

$$\|E_1\|_{op} \leq \mathrm{Tr}\,(E_1) \leq \int_{B_R^c} k_1(\boldsymbol{x}, \boldsymbol{x}) \mathrm{d}\rho_{\boldsymbol{I}_d}(\boldsymbol{x}) \leq \sqrt{\mathbb{E}_{\boldsymbol{x}}\left[C\left(1 + \frac{\|\boldsymbol{x}\|^2}{d}\right)^2\right]} \sqrt{\rho_{\boldsymbol{I}_d}(B_R^c)}\,.$$

First, we have $\sqrt{\rho_{\boldsymbol{I}_d}(B_R^c)} < \sqrt{\varepsilon}$. Also, we can calculate

$$\mathbb{E}_{\boldsymbol{x}}\left[C\left(1 + \frac{\|\boldsymbol{x}\|^2}{d}\right)^2\right] = C\mathbb{E}_{\boldsymbol{x}}\left[1 + \frac{2\|\boldsymbol{x}\|^2}{d} + \frac{\|\boldsymbol{x}\|^4}{d^2}\right] \leq 6C$$

hence

$$\|E_1\|_{op} \leq \sqrt{6C\varepsilon} := C_1' \sqrt{\varepsilon}\,.$$

Similarly, we have that $\|E_1\|_{op} \leq C_0' \sqrt{\varepsilon}$ for some absolute constant C_0'.

Using Weyl's monotonicity theorem we get, for all $k \geq 0$,

$$|\lambda_k(T_1) - \lambda_k(T_1^R)| \leq \|E_1\|_{op} < C_1' \sqrt{\varepsilon}$$

and

$$|\lambda_k(T_0) - \lambda_k(T_0^R)| \leq \|E_0\|_{op} < C_0' \sqrt{\varepsilon}$$

which shows the real eigenvalues and the truncated ones are close whenever you fix a radius R.

Result for bounded domain Based on the above discussion we set out to prove that the eigenvalue decay is maintained in the bounded domain.

We assume that the integration space $\mathcal{X}$ is bounded, i.e. $\|\boldsymbol{x}\| < \infty$ for all $\boldsymbol{x} \in \mathcal{X}$.

Consider the operator $\mathcal{T}_\# : L^2(\rho_{\boldsymbol{I}_d}) \to L^2(\rho_{\boldsymbol{\Gamma}})$ acting on f by the following rule

$$\mathcal{T}_\# f = f \circ \boldsymbol{\Gamma}^{-1/2}$$

which is designed to push f forward to the transformed space where the new kernel is acting. Since all measures in play are Gaussian, and because $\boldsymbol{\Gamma}$ is full rank, we can perform a change of variables $\boldsymbol{y} = \boldsymbol{\Gamma}^{1/2}\boldsymbol{x}$ to get

$$\|\mathcal{T}_\# f\|^2_{L^2(\rho_{\boldsymbol{\Gamma}})} = \int |f \circ \boldsymbol{\Gamma}^{-1/2}(\boldsymbol{y})|^2 \mathrm{d}\rho_{\boldsymbol{\Gamma}}(\boldsymbol{y}) = \int |f(\boldsymbol{x})|^2 \mathrm{d}\rho_{\boldsymbol{I}_d}(\boldsymbol{x}) = \|f\|^2_{L^2(\rho_{\boldsymbol{I}_d})},$$

thus $\mathcal{T}_\#$ is an isometry between both spaces and in particular an isomorphism. Similarly, we introduce its conjugate operator $(\mathcal{T}_\#)^* = \mathcal{T}_\#^{-1} := \mathcal{T}^\# : L^2(\rho_{\boldsymbol{\Gamma}}) \to L^2(\rho_{\boldsymbol{I}_d})$ which corresponds to the pullback operator

$$\mathcal{T}^\# g = g \circ \boldsymbol{\Gamma}^{1/2},$$

and is also an isometry between the spaces.

Courant-Fischer's theorem characterizes the k-th eigenvalue of any operator by the identity

$$\lambda_k(T_1) = \sup_{V:\dim(V)=k} \inf_{f \in V} \frac{\langle f, T_1 f\rangle}{\|f\|^2_{L^2(\rho_{\boldsymbol{I}_d})}}$$

and we can expand quadratic form into

$$\begin{aligned}
\langle f, T_1 f\rangle_{L^2(\rho_{\boldsymbol{I}_d})} &= \iint f(\boldsymbol{x}) k_1(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') \mathrm{d}\rho_{\boldsymbol{I}_d}(\boldsymbol{x}) \mathrm{d}\rho_{\boldsymbol{I}_d}(\boldsymbol{x}') \\
&= \iint f(\boldsymbol{\Gamma}^{-1/2}\boldsymbol{y}) k_0(\boldsymbol{y}, \boldsymbol{y}') f(\boldsymbol{\Gamma}^{-1/2}\boldsymbol{y}') \mathrm{d}\rho_{\boldsymbol{\Gamma}}(\boldsymbol{y}) \mathrm{d}\rho_{\boldsymbol{\Gamma}}(\boldsymbol{y}') \\
&= \iint f(\boldsymbol{\Gamma}^{-1/2}\boldsymbol{y}) k_0(\boldsymbol{y}, \boldsymbol{y}') f(\boldsymbol{\Gamma}^{-1/2}\boldsymbol{y}') \left[\frac{\mathrm{d}\rho_{\boldsymbol{\Gamma}}}{\mathrm{d}\rho_{\boldsymbol{I}_d}}(\boldsymbol{y})\right] \mathrm{d}\rho_{\boldsymbol{I}_d}(\boldsymbol{y}) \left[\frac{\mathrm{d}\rho_{\boldsymbol{\Gamma}}}{\mathrm{d}\rho_{\boldsymbol{I}_d}}(\boldsymbol{y}')\right] \mathrm{d}\rho_{\boldsymbol{I}_d}(\boldsymbol{y}') \\
&:= \iint r(\boldsymbol{y}) f(\boldsymbol{\Gamma}^{-1/2}\boldsymbol{y}) k_0(\boldsymbol{y}, \boldsymbol{y}') r(\boldsymbol{y}') f(\boldsymbol{\Gamma}^{-1/2}\boldsymbol{y}') \mathrm{d}\rho_{\boldsymbol{I}_d}(\boldsymbol{y}) \mathrm{d}\rho_{\boldsymbol{I}_d}(\boldsymbol{y}') \\
&= \langle r\mathcal{T}_\# f, T_0(r\mathcal{T}_\# f)\rangle_{L^2(\rho_{\boldsymbol{I}_d})}
\end{aligned}$$

where we use the equality $k_1(\boldsymbol{x}, \boldsymbol{x}') = k_0(\boldsymbol{\Gamma}^{1/2}\boldsymbol{x}, \boldsymbol{\Gamma}^{1/2}\boldsymbol{x}')$ and r is the Radon-Nikodym derivative induced by the two Gaussian measures:

$$r(\boldsymbol{y}) := \frac{\mathrm{d}\rho_{\boldsymbol{\Gamma}}}{\mathrm{d}\rho_{\boldsymbol{I}_d}}(\boldsymbol{y}) = \frac{1}{\sqrt{\det \boldsymbol{\Gamma}}} \exp\left(-\frac{1}{2}\boldsymbol{y}^\top(\boldsymbol{\Gamma}^{-1} - \boldsymbol{I}_d)\boldsymbol{y}\right).$$

Given all these definitions, we introduce the operator $\widetilde{T}_1 := \mathcal{T}_{\#} T_1 \mathcal{T}^{\#}$, leading to the identity

$$\widetilde{T}_1 h(\boldsymbol{y}) = \int k_0(\boldsymbol{y}, \boldsymbol{y}') h(\boldsymbol{y}') \mathrm{d}\rho_{\boldsymbol{\Gamma}}(\boldsymbol{y}'),$$

furthermore since T_1 and $\widetilde{T}_1$ are related by isometries, they are unitarily equivalent and share the same eigenvalues.

To continue we define the operator $M : L^2(\rho_{\boldsymbol{\Gamma}}) \to L^2(\rho_{\boldsymbol{I}_d})$ that maps $g \to rg$ and note that it is bounded and invertible since r is bounded below and above under the assumption that the integration space is bounded, i.e. it is an isomorphism.

If we let $h = M(g) = rg$ we can see

$$\begin{aligned}\langle g, \widetilde{T}_1 g \rangle_{L^2(\rho_{\boldsymbol{\Gamma}})} &= \iint k_0(\boldsymbol{y}, \boldsymbol{y}') g(\boldsymbol{y}) g(\boldsymbol{y}') \mathrm{d}\rho_{\boldsymbol{\Gamma}}(\boldsymbol{y}) \mathrm{d}\rho_{\boldsymbol{\Gamma}}(\boldsymbol{y}') \\ &= \iint k_0(\boldsymbol{y}, \boldsymbol{y}') [r(\boldsymbol{y}) g(\boldsymbol{y})][r(\boldsymbol{y}') g(\boldsymbol{y}')] \mathrm{d}\rho_{\boldsymbol{I}_d}(\boldsymbol{y}) \mathrm{d}\rho_{\boldsymbol{I}_d}(\boldsymbol{y}')\end{aligned}$$

where we use that $r(\boldsymbol{y}) d\rho_{\boldsymbol{I}_d}(\boldsymbol{y}) = d\rho_{\boldsymbol{\Gamma}}(\boldsymbol{y})$, and thus

$$\langle g, \widetilde{T}_1 g \rangle_{L^2(\rho_{\boldsymbol{\Gamma}})} = \langle M(g), T_0 M(g) \rangle_{L^2(\rho_{\boldsymbol{I}_d})} = \langle h, T_0 h \rangle_{L^2(\rho_{\boldsymbol{I}_d})}. \tag{D.1}$$

Analyzing the norms, and writing $g = \frac{1}{r} h$, we get

$$\begin{aligned}\|f\|^2_{L^2(\rho_{\boldsymbol{I}_d})} = \int |f(\boldsymbol{x})|^2 \mathrm{d}\rho_{\boldsymbol{I}_d}(\boldsymbol{x}) &= \int |g(\boldsymbol{y})|^2 \mathrm{d}\rho_{\boldsymbol{\Gamma}}(\boldsymbol{y}) \\ &= \int \left| \frac{1}{r(\boldsymbol{y})} h(\boldsymbol{y}) \right|^2 \mathrm{d}\rho_{\boldsymbol{\Gamma}}(\boldsymbol{y}) \\ &= \int \left| \frac{1}{r(\boldsymbol{x})} h(\boldsymbol{y}) \right|^2 r(\boldsymbol{y}) \mathrm{d}\rho_{\boldsymbol{I}_d}(\boldsymbol{y}) \\ &= \int \frac{1}{r(\boldsymbol{y})} |h(\boldsymbol{y})|^2 \mathrm{d}\rho_{\boldsymbol{I}_d}(\boldsymbol{y}),\end{aligned}$$

and we have that

$$\|f\|^2_{L^2(\rho_{\boldsymbol{I}_d})} = \left\| \frac{1}{\sqrt{r}} h \right\|^2_{L^2(\rho_{\boldsymbol{I}_d})}. \tag{D.2}$$

Thus we can rewrite the variational form as the following identity

$$\frac{\langle f, T_1 f \rangle_{L^2(\rho_{\boldsymbol{I}_d})}}{\|f\|_{L^2(\rho_{\boldsymbol{I}_d})}} = \frac{\langle g, \widetilde{T}_1 g \rangle_{L^2(\rho_{\boldsymbol{\Gamma}})}}{\|g\|^2_{L^2(\rho_{\boldsymbol{\Gamma}})}} = \frac{\langle h, T_0 h \rangle_{L^2(\rho_{\boldsymbol{I}_d})}}{\left\| \frac{1}{\sqrt{r}} h \right\|^2_{L^2(\rho_{\boldsymbol{I}_d})}}. \tag{D.3}$$

Since $\boldsymbol{\Gamma} - \boldsymbol{I}_d \succ 0$, we have that $\boldsymbol{y}^\top (\boldsymbol{\Gamma}^{-1} - \boldsymbol{I}_d) \boldsymbol{y} \leq 0$ for all $\boldsymbol{y} \in \mathbb{R}^d$ and

$$r(\boldsymbol{y}) = \frac{1}{\sqrt{\det \boldsymbol{\Gamma}}} \exp\left(-\frac{1}{2} \boldsymbol{y}^\top (\boldsymbol{\Gamma}^{-1} - \boldsymbol{I}_d) \boldsymbol{y} \right) \geq \frac{1}{\sqrt{\det \boldsymbol{\Gamma}}}, \ \forall \boldsymbol{y} \in \mathbb{R}^d.$$

And because we assumed $\mathcal{X}$ to be bounded, we can find constants $c, C > 0$ such that $c \leq r(\boldsymbol{y}) \leq C$ for all $\boldsymbol{y} \in \mathcal{X}$, thus the relationship between the norms of f and h in Eq. (D.2) gives

$$\frac{1}{C}\|h\|^2_{L^2(\rho_{\boldsymbol{I}_d})} \leq \|f\|^2_{L^2(\rho_{\boldsymbol{I}_d})} = \|g\|^2_{L^2(\rho_{\boldsymbol{\Gamma}})} \leq \frac{1}{c}\|h\|^2_{L^2(\rho_{\boldsymbol{I}d})}.$$

Combining these bounds with the identity for the quadratic forms (D.1), we know that for a function g from a fixed subspace V' with dimension k, if $h = M(g)$, we have

$$c\frac{\langle h, T_0 h\rangle}{\|h\|^2_{L^2(\rho_{\boldsymbol{I}_d})}} \leq \frac{\langle g, \widetilde{T}_1 g\rangle}{\|g\|^2_{L^2(\rho_{\boldsymbol{\Gamma}})}} \leq C\frac{\langle h, T_0 h\rangle}{\|h\|^2_{L^2(\rho_{\boldsymbol{I}_d})}}.$$

Now, given a subspace $V' \subset L^2(\rho_{\boldsymbol{\Gamma}})$ of dimension k, we define

$$W = M(V') := \{rg : g \in V'\} \subset L^2(\rho_{\boldsymbol{I}_d}).$$

Because M is bijective and isomorphic, W has dimension k and the mapping $V' \mapsto W$ is a bijection, allowing us to associate every subspace $V' \subset L^2(\rho_{\boldsymbol{\Gamma}})$ with a respective unique subspace $W \subset L^2(\rho_{\boldsymbol{I}_d})$.

Thus, taking the infimum over all functions inside V' is the same as taking the infimum over the associated subspace W and

$$c\left(\inf_{h\in W}\frac{\langle h, T_0 h\rangle}{\|h\|^2_{L^2(\rho_{\boldsymbol{I}_d})}}\right) \leq \inf_{g\in V'}\frac{\langle g, \widetilde{T}_1 g\rangle}{\|g\|^2_{L^2(\rho_{\boldsymbol{\Gamma}})}} \leq C\left(\inf_{h\in W}\frac{\langle h, T_0 h\rangle}{\|h\|^2_{L^2(\rho_{\boldsymbol{I}_d})}}\right),$$

and finally taking the supremum over all possible sets V' of dimension k is the same as taking the supremum over the associated sets W and we get

$$c\left(\sup_{\dim W=k}\inf_{h\in W}\frac{\langle h, T_0 h\rangle}{\|h\|^2_{L^2(\rho_{\boldsymbol{I}_d})}}\right) \leq \sup_{\dim V'=k}\inf_{g\in V'}\frac{\langle g, \widetilde{T}_1 g\rangle}{\|g\|^2_{L^2(\rho_{\boldsymbol{\Gamma}})}} \leq C\left(\sup_{\dim W=k}\inf_{h\in W}\frac{\langle h, T_0 h\rangle}{\|h\|^2_{L^2(\rho_{\boldsymbol{I}_d})}}\right),$$

which are exactly the expressions for the eigenvalues of $\lambda(T_0)$ and $\lambda_k(\widetilde{T}_1) = \lambda_k(T_1)$, giving the desired result

$$c\lambda_k(T_0) \leq \lambda_k(\widetilde{T}_1) = \lambda_k(T_1) \leq C\lambda_k(T_0).$$

■

D.2. The spiked covariance expansion framework

Theorem D.2 *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 $D^{(j)}_{\boldsymbol{u}} G$ 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*

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

Then, we have that

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

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

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

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

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

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

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. ■

D.3. Applying the expansion to the kernel k_1

To bring Theorem D.2 into our context, for a fixed pair $(\boldsymbol{x}, \boldsymbol{x}')$, we define the function $G(\boldsymbol{w}) := G_{\boldsymbol{x},\boldsymbol{x}'}(\boldsymbol{w}) = \sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{w}, \boldsymbol{x}'\rangle)$. Since $\boldsymbol{\Gamma}$ is positive definite, setting $\gamma_1 := \frac{A}{d}$ and $\gamma_2 := \frac{B}{d}$, we apply the theorem to obtain the expansion:

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

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

$$\begin{aligned} S(\boldsymbol{x}, \boldsymbol{x}') := \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{A}{d}\boldsymbol{I}_d)}[D_{\boldsymbol{w}^*}^{(2)} G(\boldsymbol{w})] &= \langle \boldsymbol{w}^*, \boldsymbol{x}'\rangle^2 \mathbb{E}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\sigma^{(2)}(\langle \boldsymbol{w}, \boldsymbol{x}'\rangle)] \\ &\quad + 2\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle\langle \boldsymbol{w}^*, \boldsymbol{x}'\rangle \mathbb{E}[\sigma'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\sigma'(\langle \boldsymbol{w}, \boldsymbol{x}'\rangle)] \\ &\quad + \langle \boldsymbol{w}^*, \boldsymbol{x}\rangle^2 \mathbb{E}[\sigma^{(2)}(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{w}, \boldsymbol{x}'\rangle)]\,. \end{aligned} \tag{D.4}$$

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

$$k_1(\boldsymbol{x}, \boldsymbol{x}') = k_0^{(A)}(\boldsymbol{x}, \boldsymbol{x}') + \frac{B}{2d} S(\boldsymbol{x}, \boldsymbol{x}') + R(\boldsymbol{x}, \boldsymbol{x}')\,, \tag{D.5}$$

where $S(\boldsymbol{x}, \boldsymbol{x}')$ is defined by Eq. (D.4), and $R(\boldsymbol{x}, \boldsymbol{x}')$ are the residual terms formally defined as the tail of the expansion for $j \geq 2$:

$$R(\boldsymbol{x}, \boldsymbol{x}') = \sum_{j=2}^{\infty} \frac{1}{n!} \left(\frac{B}{2d}\right)^j \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{A}{d}\boldsymbol{I}_d)}[D_{\boldsymbol{w}^*}^{(2j)} G(\boldsymbol{w})]\,. \tag{D.6}$$

Theorem 2.1 ensures k_1, $k_0^{(A)}$, and S grow at most polynomially, thus R defines a bounded integral operator on the Gaussian space, and is dominated by the first terms of the expansion by the following lemma.

D.4. The remainder term R is dominated

Lemma D.3 *Under Theorem 2.1 for the function R defined by Eq. (D.6) with bounded integral operator $T_R f(\boldsymbol{x}) = \int R(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}')$, we have that $\|T_R\|_{op} = \mathcal{O}\left(B^2/d^2\right)$.*

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 $\mathcal{N}(0, \frac{1}{d}\boldsymbol{\Gamma}(t))$, 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

$$F(1) = F(0) + F'(0) + \frac{F''(\xi)}{2!}$$

with $\xi \in [0,1]$.

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

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

therefore

$$F''(\xi) = \left(\frac{B}{2d}\right)^2 \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{1}{d}\boldsymbol{\Gamma}(\xi))}[D^{(4)}_{\boldsymbol{w}^*} G(\boldsymbol{w})],$$

and 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.

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

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

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

Then we can write

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

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

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

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

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

Thus, the absolute value of R can be bounded with

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

where we define

$$\begin{aligned}
P^*(\boldsymbol{U},\boldsymbol{Q}_\xi) &= M_L^2\|\boldsymbol{U}\|^4\cdot\mathbb{E}_{\boldsymbol{h}}\left[(2+\|\boldsymbol{h}\|^2)(\|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}^4\|\boldsymbol{h}\|^4 + 6\|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}^3\|\boldsymbol{h}\|^2 + 3\|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}^2)\right] \\
&= M_L^2\|\boldsymbol{U}\|^4\cdot\mathbb{E}_{\boldsymbol{h}}\Big[\|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}^4\|\boldsymbol{h}\|^6 + (2\|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}^4 + 6\|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}^3)\|\boldsymbol{h}\|^4 \\
&\quad + (3\|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}^2 + 12\|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}^3)\|\boldsymbol{h}\|^2 + 6\|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}^2\Big].
\end{aligned}$$

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

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

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

$$\begin{aligned}
P^*(\boldsymbol{U},\boldsymbol{Q}_\xi) \le M_{L_\sigma}^2\|\boldsymbol{U}\|^4\cdot\Big[&(120\|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}^4\|\boldsymbol{Q}_\xi\|_{\mathrm{op}}^3 + 12(2\|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}^4\| + 6\|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}^3)\|\boldsymbol{Q}_\xi\|_{\mathrm{op}}^2 \\
&+ 2(2\|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}^2 + 6\|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}^3)\|\boldsymbol{Q}_\xi\|_{\mathrm{op}} + 4\|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}^2\Big].
\end{aligned} \tag{D.7}$$

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

Lastly, notice that if we square the bound in D.7, we obtain a combination of powers of $\|\boldsymbol{U}\|, \|\boldsymbol{Q}_\xi^{-1}\|_{\mathrm{op}}$ and $\|\boldsymbol{Q}_\xi\|_{\mathrm{op}}$. 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

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

Thus, we have the following result for the HS norm

$$\begin{aligned}\|T_R\|_{\mathrm{HS}}^2 &= \iint |R(\boldsymbol{x}, \boldsymbol{x}')|^2 \mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\rho_{\mathcal{X}}(\boldsymbol{x}') \\ &\le \frac{B^4}{64d^4}\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[P^*(\boldsymbol{U}, \boldsymbol{Q}_\xi)^2]\end{aligned}$$

$$= \mathcal{O}\left(\frac{B^4}{d^4}\right)$$

which implies

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

and the proof is completed. ■

D.5. Eigenfunctions of the isotropic ReLU kernel

Lemma D.4 *The normalized eigenfunctions 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 $\mathbb{S}^{d-1}$.*

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 $\frac{r}{\sqrt{d}} Y_{k,m}(\boldsymbol{\omega})$. ■

D.6. Approximating the first-order term of the ReLU expansion

Wconsider the expansion of the kernel for the ReLU activation, and determine the first two terms from Eq. (D.5). When $\sigma(t) = \max(0,t)$, the scaling effect is captured by the identity $k_0^{(A)}(\boldsymbol{x}, \boldsymbol{x}') = A k_0(\boldsymbol{x}, \boldsymbol{x}')$. 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 $\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\frac{A}{d}\boldsymbol{I}_d)}[\sigma'(\langle \boldsymbol{w},\boldsymbol{x}\rangle)\sigma'(\langle \boldsymbol{w},\boldsymbol{x}'\rangle)] = \frac{\pi-\theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi}$.[1] Therefore, we have the following

$$S(\boldsymbol{x},\boldsymbol{x}') = \frac{\pi-\theta_{\boldsymbol{x},\boldsymbol{x}'}}{\pi}[\langle \boldsymbol{x},\boldsymbol{w}^*\rangle\langle \boldsymbol{x}',\boldsymbol{w}^*\rangle] + \frac{\sin\theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi}\left(\frac{\|\boldsymbol{x}'\|}{\|\boldsymbol{x}\|}\langle \boldsymbol{x},\boldsymbol{w}^*\rangle^2 + \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{x}'\|}\langle \boldsymbol{x}',\boldsymbol{w}^*\rangle^2\right), \quad \text{(D.8)}$$

and the ReLU k_1 kernel is given by

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

and we have the following result

Lemma D.5 *Consider the function given by*

$$S(\boldsymbol{x},\boldsymbol{x}') = \frac{\pi-\theta_{\boldsymbol{x},\boldsymbol{x}'}}{\pi}[\langle \boldsymbol{x},\boldsymbol{w}^*\rangle\langle \boldsymbol{x}',\boldsymbol{w}^*\rangle] + \frac{\sin\theta_{\boldsymbol{x},\boldsymbol{x}'}}{2\pi}\left(\frac{\|\boldsymbol{x}'\|}{\|\boldsymbol{x}\|}\langle \boldsymbol{x},\boldsymbol{w}^*\rangle^2 + \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{x}'\|}\langle \boldsymbol{x}',\boldsymbol{w}^*\rangle^2\right), \quad \text{(D.9)}$$

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 Theorem 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 $\|E\|_{L^2(\rho_{\mathcal{X}})} = o_d(1)$.

Proof For this proof we write

$$\begin{aligned} T_1 f(\boldsymbol{x}) &= AT_0 f(\boldsymbol{x}) + \frac{B}{2d}T_S f(\boldsymbol{x}) + T_R f(\boldsymbol{x}) \\ &= AT_0 f(\boldsymbol{x}) + \frac{B}{2d}\left(T_{(1*)}f(\boldsymbol{x}) + T_{(2*)}f(\boldsymbol{x})\right) + T_R f(\boldsymbol{x}) \end{aligned}$$

and, as we have seen by Lemma D.3, $\|T_R f\|_{L^2(\rho_{\mathcal{X}})} = o_d\left(\frac{|B|}{d}\right)$. Thus, we will study the terms $T_{(1*)}f$ and $T_{(2*)}f$.

Analysis of the first term: We start by analyzing

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

1. 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 $\delta(ct) = \frac{\delta(t)}{|c|}$.

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.3, 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 D.4, 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

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

$$= \left(\frac{1}{d} \int_{\mathbb{R}} |r^2|^2 \phi(r) dr \right)$$

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

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

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.3, 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

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

$$= \frac{1}{2\pi}\frac{\langle \boldsymbol{x}, \boldsymbol{w}^* \rangle^2}{\|\boldsymbol{x}\|}\langle \|\cdot\|, f\rangle + \frac{1}{2\pi}\|\boldsymbol{x}\| \left\langle \frac{\langle \cdot, \boldsymbol{w}^* \rangle^2}{\|\cdot\|}, f \right\rangle + E_1^{(2)}(\boldsymbol{x}) + E_2^{(2)}(\boldsymbol{x}) + E_3^{(2)}(\boldsymbol{x})$$
$$:= T_S^{(2*)}(\boldsymbol{x}) + E_1^{(2)}(\boldsymbol{x}) + E_2^{(2)}(\boldsymbol{x}) + E_3^{(2)}(\boldsymbol{x}).$$

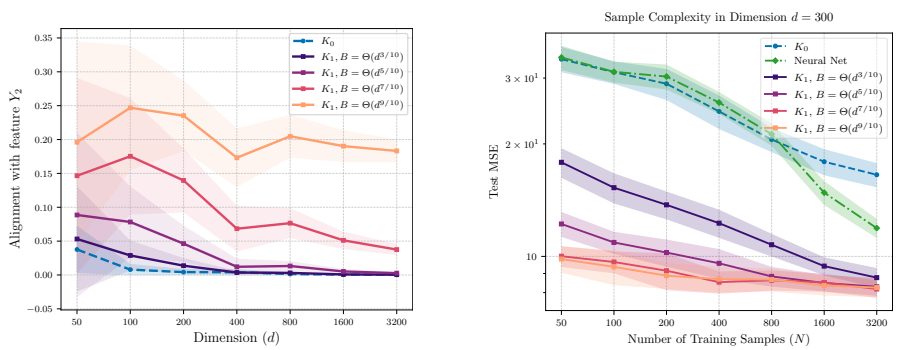
Finally, since we chose $\epsilon = d^{-1/3}$, we have

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

■

Appendix E. Experimental details

E.1. Experiments



(a) Feature alignment (b) Test MSE

Figure 1: (a) Alignment with the directional feature Y_2. (b) Generalization performance at $d = 300$.

Over 10 trials (variance shown as shaded areas around the curves), we sample $\boldsymbol{Z} \in \mathbb{R}^{N\times d}$ with $\boldsymbol{z}_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$, and compare the ReLU kernel k_0, the ReLU kernel k_1 and a two-layer ReLU MLP (width 400).

In Fig. 1(a), we compute the alignment $\langle \boldsymbol{v}_i^{\text{top}}, Y_2(\boldsymbol{Z})\rangle$, where $\boldsymbol{v}_i^{\text{top}}$ is the lead eigenvector of the kernel matrices for $i \in \{0, 1\}$. As expected, the alignment for k_1 grows with B, while k_0 always remains near-zero. In Fig. 1(b) we track the Mean Squared Error ($d = 300$) on a test set of 600 samples when learning $g(t) = 2t^2 + 3t + 4\sin(2t)$, comparing Kernel Ridge Regression and the network across varying training sample sizes (N). We observe the network converging toward the spiked kernel's performance; noting that while the kernels have privileged access to $\boldsymbol{w}^*$, the network must recover it from the data.

Going forward, we provide comprehensive details regarding the experimental setup. For all experiments, we always consider the matrix $\boldsymbol{\Gamma} = A\boldsymbol{I}_d + B\boldsymbol{w}^*(\boldsymbol{w}^*)^\top$, where we set

the artificial scaling $A = 1.2$ and choose multiple values for B to observe of its influence. Denoting by N the number of training samples, we sample the input data matrix $\boldsymbol{Z} \in \mathbb{R}^{N \times d}$, where each row $\boldsymbol{z}_i$ is drawn independently from a standard multivariate Gaussian distribution $\boldsymbol{z}_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$.

E.2. Different models

Throughout the experiments we consider three different models detailed as follows:

1. **Base ReLU Kernel (k_0):** The standard ReLU activation kernel with closed form given by the arc-cosine kernel

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

   with $\gamma \coloneqq \frac{\langle \boldsymbol{x}, \boldsymbol{x}' \rangle}{\|\boldsymbol{x}\|\|\boldsymbol{x}'\|}$.

2. **ReLU Kernel (k_1):** The closed form for kernel k_1

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

   with $\gamma_{\boldsymbol{\Gamma}} \coloneqq \frac{\boldsymbol{x}^\top \boldsymbol{\Gamma} \boldsymbol{x}'}{\sqrt{(\boldsymbol{x}^\top \boldsymbol{\Gamma} \boldsymbol{x})(\boldsymbol{x}'^\top \boldsymbol{\Gamma} \boldsymbol{x}')}}$.

3. **ReLU MLP:** A two-layer Multi-Layer Perceptron (MLP) with hidden layer width of 400 and linear output layer.

To obtain the closed form of the kernels, the calculations leverage the coordinate transformation trick (Liao and Couillet, 2018, Appendix A), and we omit the details here.

Kernels: For $i \in \{0, 1\}$, we calculate the kernel matrix $\boldsymbol{K}_i(\boldsymbol{Z}) \in \mathbb{R}^{N \times N}$ whose entries correspond to

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

ReLU NN: The ReLU neural network was trained using the Adam optimizer (`torch.optim.Adam`) with the following parameters:

- **First layer dimension:** 400
- **Output layer dimension:** 1
- **Epochs:** 1
- **Batch size:** 64
- **Loss Function:** Mean Squared Error (`torch.nn.MSELoss`)

E.3. Metrics and Significance

All experiments were executed over 10 independent trials. Shaded regions in figures indicate the variance across these trials. For every experiment routine implemented, we set random seeds to ensure reproducibility across all sources of randomness (e.g. data sampling, NN initialization, K-Fold validation, etc.).

E.4. Alignment experiment

For the results in Fig. 1(*a*) we set the seed 54643. For the alignment analysis, we vary the dimension $d \in \{50, 100, 200, 400, 800, 1600, 3200\}$, and we also vary the magnitude of $B \in \{5d^{3/10}, 5d^{5/10}, 5d^{7/10}, 5d^{9/10}\}$ obtaining four different kernels. The alignment reported is calculated as $\langle \boldsymbol{v}_i^{\text{top}}, Y_2(\boldsymbol{Z})\rangle$, where $\boldsymbol{v}_i^{\text{top}}$ is the lead eigenvector of the kernel matrix $\boldsymbol{K}_i(\boldsymbol{Z})$ for $i \in \{0, 1\}$ and Y_2 is the normalized version of the function

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

where $\boldsymbol{\omega} = \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}$.

E.5. Generalization performance

For the results in Fig. 1(*b*) we set the seed 558812. For the learning performance analysis, we work on the fixed dimension $d = 300$ and vary the number of training samples $N \in \{50, 100, 200, 400, 800, 1600, 3200\}$ and again we build three different kernels $A = 1.2$ with $B \in \{5d^{3/10}, 5d^{5/10}, 5d^{7/10}, 5d^{9/10}\}$. All models are trying to learn the target function $g(t) = 2t^2 + 3t + 4\sin(2t)$.

We use Kernel Ridge Regression (KRR) with k_0 and k_1 to obtain the solutions $\widehat{\boldsymbol{a}}_i = (\boldsymbol{K}_i + \lambda \boldsymbol{I}_N)^{-1}\boldsymbol{y}$, with regularization parameter λ. The regularization is chosen through 5-fold cross-validation, for each sample size N, from a logarithmic grid $\{10^{-3}, 10^{-2}, \ldots, 10^3\}$ with random state 38182.

We sample a test set $\widetilde{\boldsymbol{Z}}$, from the same distribution as $\boldsymbol{Z}$, consisting of $M = 600$ independent samples. Given the test samples, for $i \in \{0, 1\}$, we 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 predictor

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

for each $i \in \{0, 1\}$. Lastly, given the `torch.nn.Linear` and `torch.nn.ReLU` implementations from `PyTorch`, we implement the predictor $\widehat{f}_{\text{NN}} = \text{Linear}_2(\text{ReLU}(\text{Linear}_1(\boldsymbol{z})))$. The network was trained with learning rate obtained through 5-fold cross validation from a logarithmic grid $\{10^{-3}, 10^{-2}, 10^{-1}, 10^0\}$, for every sample size N, with random state 123114.

To obtain the metric from the figure, we calculate the Test Mean Squared Error by computing

$$\frac{1}{M}\sum_{i=1}^{M}\Big(y_i - \widehat{f}_p(\widetilde{\boldsymbol{z}}_i)\Big)^2$$

for all $\widehat{f}_p \in \{\widehat{f}_0, \widehat{f}_1, \widehat{f}_{\text{NN}})$.

E.6. Implementation, Software Stack and Hardware

The experimental framework was implemented in `Python` (v3.10.12). We use `PyTorch` (v.2.11.0) (Paszke et al., 2019), such that layers and activation are out-of-the-box calls to the methods `torch.nn.Linear` and `torch.nn.ReLU`, specifying only the input, first layer, and output sizes. We utilized the `NumPy` (v.1.26.4) (Harris et al., 2020) and `SciPy` (v.1.15.3) (Virtanen et al., 2020) libraries for numerical linear algebra operations, specifically for the eigendecomposition (`scipy.linalg.eigh`) of the kernel matrices, and for the KRR implementation (`numpy.linalg.solve`). Lastly, for KFold validation we used `sklearn`'s (v.1.4.1.post1) `model_selection.KFold` object (Pedregosa et al., 2011).

All experiments were conducted on a laptop with a 11th Gen Intel(R) Core(TM) i7-11800H @ 2.30GHz processor, 8GB of RAM 3200 MHz and a single NVIDIA RTX 3060 Laptop GPU, and should be reproducible, even with a CPU.