THE RKHS COST OF NEURAL ACTIVATIONS FOR SPARSE VARIATIONAL GAUSSIAN PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Mean functions of sparse variational Gaussian processes are equivalent to feed-forward neural networks, with neurons determined by the covariance between the interdomain inducing variables and the GP function values. We investigate whether inducing variables can be chosen so that these cross-covariances are neurons with typical activation functions, such as ReLU. This requires the desired neurons to lie in the prior's reproducing kernel Hilbert space (RKHS). We study their membership and variational cost. Under the stated growth and parameter assumptions, in dimension $d \geq 2$, a neuron with non-zero weight and bias in the parameter support belongs to its own single-layer NNGP RKHS if and only if the activation is polynomial. Under Gaussian input and parameter measures and the stated regularity assumptions, an activation whose first non-zero derivative jump has order s yields neurons outside any fixed trace-class candidate RKHS for almost every parameter once $d \geq 2s + 1$. Smooth non-polynomial neurons can belong to suitable RKHSs, but their scale-invariant RKHS cost grows superpolynomially in probability under the stated fan-in and bias laws. This cost obstruction also holds for random neural-network means. In Gaussian regression, such means are either inadmissible or, under a polynomial combined prior-variance and data budget, have a lower ELBO than their centred counterparts with high probability. These results identify when exact neural-network mean representations are feasible in sparse GPs and suggest that activation-shaped inducing variables may be most promising in low-dimensional settings.

1 INTRODUCTION

It has long been known that infinite-width neural networks yield Gaussian process (GP) priors (Neal, 1996; Williams, 1997; Lee et al., 2018; Matthews et al., 2018). On the other hand, the mean function of a sparse variational Gaussian process (SVGP) has the form of a finite-width neural network, since it is a linear combination of cross-covariances between interdomain inducing variables and GP function values. In particular, these cross-covariance functions play the role of neurons. This correspondence then extends layerwise to deep Gaussian processes (DGPs) (Damianou & Lawrence, 2013; Salimbeni & Deisenroth, 2017). Since the posterior mean at each layer has the form of a neural network, composing these means recovers the computation of a feed-forward neural network.

This correspondence raises a natural question: can the inducing variables be chosen such that the resulting neurons take the familiar form $x \mapsto \sigma(w^\top x + b)$, for typical activation functions σ, such as ReLU, tanh, or Softplus? To answer this question, we must understand which neurons can be realised as cross-covariances of valid, finite-variance inducing variables.

The key observation is that a neuron is an admissible cross-covariance exactly when it belongs to the prior's reproducing kernel Hilbert space (RKHS). However, membership alone does not ensure practical usefulness. An SVGP mean h incurs an ELBO penalty of at least $\frac{1}{2}\|h\|^2_{\mathcal{H}_K}$, which can outweigh the available likelihood improvement even if the norm is finite. We establish four results on membership and cost.

- **NNGP membership requires polynomial activations.** For continuous activations of at most polynomial growth, non-degenerate Gaussian weights, and Gaussian or uniform bi-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ases, a neuron with non-zero weight belongs to its own single-layer NNGP RKHS on $\mathbb{R}^d$, $d \geq 2$, if and only if the activation is polynomial (Theorem 1).

- **Restricted spectra impose a tight RKHS cost barrier.** We give a tight spectral lower bound on scale-invariant RKHS cost over every measurable subset of neuron parameters (Theorem 2). Divergence of the restricted square-root trace on every positive-measure subset forces almost every neuron outside any fixed trace-class candidate RKHS.
- **Finite derivative jumps obstruct membership.** Under the Gaussian measures and regularity assumptions of Theorem 4, almost every neuron with an activation having a finite jump discontinuity in some derivative lies outside any fixed trace-class candidate RKHS when $d \geq 2s + 1$, where s is the first non-zero derivative-jump order.
- **Smooth activations retain a high-dimensional cost.** Under isotropic fan-in scaling and the stated bias and growth assumptions, neurons with smooth non-polynomial activations have superpolynomial scale-invariant RKHS cost in probability for every sequence of candidate priors (Theorem 5). In Gaussian regression, this yields an ELBO cost for random networks of any deterministic finite width (Section 5).

2 Background and Related Work

We will start by reviewing deep Gaussian processes (DGPs), sparse variational Gaussian processes (SVGPs), and how we can relate mean propagation in SVGPs to computation in feed-forward neural networks. We then connect this construction to RKHS membership and variational cost, placing our results alongside existing approaches.

2.1 Sparse GPs and Neural-Network Mean Propagation

A Gaussian process is a collection of random variables $\{f(x) : x \in \mathcal{X}\}$ whose every finite subcollection is jointly Gaussian. It is specified by a mean function $\mu(x) = \mathbb{E}[f(x)]$ and covariance kernel $K(x, x') = \mathrm{Cov}(f(x), f(x'))$, written $f \sim \mathcal{GP}(\mu, K)$ (Rasmussen & Williams, 2006). We work with centred priors, for which $\mu = 0$.

A deep Gaussian process composes independent GP layers, $F = f_L \circ \cdots \circ f_1$ (Damianou & Lawrence, 2013). We take the output coordinates of each DGP layer to be independent scalar GPs. We use sparse variational approximations for the individual layers (Titsias, 2009; Salimbeni & Deisenroth, 2017), whose form we now review.

For a centred GP layer, let $u = (u_1, \ldots, u_M)^\top$ be the inducing vector and let $q(u) = \mathcal{N}(m, S)$. Let

$$K_{uu} = \mathrm{Cov}(u, u), \qquad k_u(x) = \mathrm{Cov}(u, f(x)) \in \mathbb{R}^M, \qquad B = K_{uu}^{-1} m.$$

Entrywise, $[K_{uu}]_{ij} = \mathrm{Cov}(u_i, u_j)$ for $1 \leq i, j \leq M$, and $[k_u(x)]_i = \mathrm{Cov}(u_i, f(x))$. The induced approximate posterior is $q(f) = \mathcal{GP}(h, C_q)$, where

$$h(x) = B^\top k_u(x), \tag{1}$$

$$C_q(x, x') = K(x, x') + k_u(x)^\top K_{uu}^{-1}(S - K_{uu})K_{uu}^{-1}k_u(x'). \tag{2}$$

Here m and S are the variational mean and covariance. On training inputs $X = (x_1, \ldots, x_N)$, write $h_X = (h(x_1), \ldots, h(x_N))^\top$ and define

$$\begin{aligned}
K_{ff} &= K(X, X) \in \mathbb{R}^{N \times N}, && [K_{ff}]_{ij} = K(x_i, x_j), \\
K_{uf} &= [k_u(x_1), \ldots, k_u(x_N)] \in \mathbb{R}^{M \times N}, && [K_{uf}]_{ij} = \mathrm{Cov}(u_i, f(x_j)), \\
K_{fu} &= K_{uf}^\top \in \mathbb{R}^{N \times M}, && [K_{fu}]_{ij} = \mathrm{Cov}(f(x_i), u_j).
\end{aligned}$$

The parameters of an SVGP, including kernel hyperparameters, inducing-variable parameters, and variational parameters, are generally learned by maximising the evidence lower bound (ELBO) (Titsias, 2009; Hensman et al., 2013):

$$\mathcal{L}(q) = \sum_{n=1}^{N} \mathbb{E}_{q(f(x_n))}[\log p(y_n \mid f(x_n))] - \mathrm{KL}(q(u) \,\|\, p(u)) \tag{3}$$

$$= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{\|y - h_X\|^2 + \mathrm{tr}(S_f)}{2\sigma^2} - \mathrm{KL}(q(u) \,\|\, p(u)). \tag{4}$$

Here, equation 4 is for the Gaussian likelihood $p(y \mid f_X) = \mathcal{N}(f_X, \sigma^2 I_N)$, where $f_X = (f(x_1), \ldots, f(x_N))^\top$ and

$$S_f = K_{ff} - Q_{ff} + K_{fu} K_{uu}^{-1} S K_{uu}^{-1} K_{uf}. \tag{5}$$

Here $Q_{ff} = K_{fu} K_{uu}^{-1} K_{uf}$ is the Nyström approximation term and $S_f = \mathrm{Cov}_q(f_X, f_X)$. The KL term is

$$\mathrm{KL}(q(u) \,\|\, p(u)) = \frac{1}{2} \left[\mathrm{tr}(K_{uu}^{-1} S) + m^\top K_{uu}^{-1} m - M + \log \frac{\det K_{uu}}{\det S} \right]. \tag{6}$$

Returning to the form of our variational posterior over function values $f_\ell(\cdot)$, the most important thing to note for our purposes is that the propagation of the mean of each SVGP layer is given by

$$\mathbb{E}_q[f_L(\cdot)] \circ \cdots \circ \mathbb{E}_q[f_1(\cdot)] = B_L^\top k_{u_L} \circ \cdots \circ B_1^\top k_{u_1} \tag{7}$$

which is equivalent to the computation of the forward pass of a fully connected feed-forward neural network (Sun et al., 2020; Dutordoir et al., 2021)

$$V_L^\top \sigma(W_L^\top(\cdot) + b_L) \circ \cdots \circ V_1^\top \sigma(W_1^\top(\cdot) + b_1) \tag{8}$$

via the identifications

$$B_\ell = V_\ell, \qquad k_{u_\ell}(\cdot) = \sigma(W_\ell^\top(\cdot) + b_\ell) \tag{9}$$

Here u_ℓ denotes the inducing vector at layer ℓ, with k_{u_ℓ} and B_ℓ defined as above for that layer. The matrices W_ℓ are the usual pre-activation weights in layer ℓ, and V_ℓ are post-activation weights for layer ℓ. The main obstruction to taking advantage of this correspondence between SVGP mean propagation and neural network computation is our ability to choose our inducing variables u_ℓ such that our cross-covariances $k_{u_\ell}(\cdot) = \sigma(W_\ell^\top(\cdot) + b_\ell)$ for some typical activation function σ.

The original SVGP construction uses inducing points, $u_i = f(z_i)$, for inducing locations $z_i \in \mathcal{X}$. Their cross-covariances are therefore kernel sections, $\mathrm{Cov}(u_i, f(x)) = K(z_i, x)$, which in general do not coincide with standard neurons such as $x \mapsto \sigma(w^\top x + b)$.

2.2 Interdomain Inducing Variables and RKHS Functions

Since the inducing points of the original SVGP construction do not provide our desired neurons, we need a sparse approximation procedure that allows for more general cross-covariances. Interdomain inducing variables (Lázaro-Gredilla & Figueiras-Vidal, 2009) are defined as linear functionals of the GP, $u = \int_{\mathcal{X}} q(x) f(x)\, \mu(dx)$, for some choice of function $q(x)$ and measure μ.

Rather than restricting our attention to a particular integral representation, we use a more general and abstract formulation of admissible inducing variables (Burt, 2018; Kanagawa et al., 2025). We take an admissible inducing variable to be a finite-variance linear functional of the GP. For a centred GP f with kernel K, define its Gaussian Hilbert space by

$$\mathcal{G}_f := \overline{\mathrm{span}}\{f(x) : x \in \mathcal{X}\} \subset L^2(\Omega)$$

Since the elements here are finite-variance linear functionals of the GP values, obtained as $L^2(\Omega)$-limits of finite linear combinations, this is exactly the space of admissible inducing variables. By the Loève isometry, $\mathcal{G}_f$ is canonically isometrically isomorphic to the RKHS $\mathcal{H}_K$ (van der Vaart & van Zanten, 2008). Consequently, every $u \in \mathcal{G}_f$ corresponds uniquely to a function $\psi_u \in \mathcal{H}_K$, with

$$\mathrm{Cov}(u, f(x)) = \psi_u(x), \qquad \mathrm{Var}(u) = \|\psi_u\|_{\mathcal{H}_K}^2.$$

Conversely, every $\psi \in \mathcal{H}_K$ is the cross-covariance of a unique admissible inducing variable, with uniqueness understood in $L^2(\Omega)$. We give a short proof of the Loève isometry and connect it to the standard interdomain integral formulation in Appendix A.1.

2.3 RKHS Membership of Neurons

We are interested in SVGPs on $\mathcal{X} = \mathbb{R}^d$ whose inducing-variable cross-covariances are exactly neural-network neurons $[k_u(x)]_i = \sigma(w_i^\top x + b_i)$, for a typical activation function σ.

Previous work has pursued this goal through particular interdomain constructions. Sun et al. (2020) interpret neurons as inducing features for dot-product kernels, while Dutordoir et al. (2021) construct spherical activation features whose cross-covariances yield neural-network-like DGP mean propagation. Their construction also exposes compatibility conditions between the activation and the kernel, requiring, for example, truncation for ReLU.

Our analysis provides a general RKHS perspective on these compatibility constraints. Rather than beginning from a particular inducing construction, we ask instead a representation-independent question. By Section 2.2, a function can arise as the cross-covariance of a valid finite-variance inducing variable if and only if it lies in the prior RKHS, $\mathcal{H}_K$. Our question therefore becomes: for which kernels K does $\mathcal{H}_K$ contain ridge functions of the form $x \mapsto \sigma(w^\top x + b)$, ideally for a broad family of parameters (w, b)?

Existing results already indicate that neural-kernel RKHS membership can be restrictive. Bach (2017) shows that a single ReLU feature does not belong to the RKHS generated by uniformly distributed ReLU features on a sphere. For the two-layer ReLU NTK on the sphere, Bietti & Mairal (2019) derive its spherical-harmonic decomposition and hence an explicit RKHS membership criterion. In particular, they show that sufficiently smooth even functions belong to the RKHS. Holzmüller & Schölpple (2026, Theorem 9) extend this type of characterisation to the NNGP and NTK kernels of a broad class of piecewise-smooth activations, including ELU and Leaky ReLU. Up to norm equivalence, the resulting spherical RKHSs are Sobolev spaces whose regularity is determined by the first derivative order at which the activation has a discontinuity.

These results characterise the RKHSs associated with particular neural kernels on the sphere. Our results instead focus directly on membership of the generating neurons. For the single-layer NNGP kernel, our Theorem 1 characterises whether an individual ridge function $x \mapsto \sigma(w^\top x + b)$ belongs to its RKHS on the full domain $\mathbb{R}^d$, $d \geq 2$, showing under our stated assumptions that this occurs exactly when σ is polynomial. We then move beyond neural kernels altogether. Our Theorem 4 applies to any fixed trace-class candidate covariance, with no assumed rotational invariance, and shows that under Gaussian input and parameter measures almost every neuron lies outside its RKHS when $d \geq 2s + 1$, where s is the first non-zero derivative-jump order.

However, the membership picture is not entirely negative. Our finite-jump obstruction does not apply to smooth activations, and indeed suitable RKHSs can contain such neurons exactly. For example, the Hermite spaces studied by Dick et al. (2017, Section 2) yield trace-class covariances under Gaussian inputs whose RKHSs contain ridge functions with smooth activations such as tanh, sigmoid, and Softplus. Thus, failure of membership in an activation's own NNGP RKHS does not imply that the neuron cannot belong to any suitable RKHS. Our remaining results turn from the binary question of membership to the quantitative question of its cost. That is, even when such neurons belong to a suitable RKHS, we ask whether their RKHS norm can remain moderate as dimension increases.

2.4 Variational Cost of Neurons

Membership of a suitable RKHS is only the first requirement, however. Even when a desired neuron belongs to the prior RKHS, representing it as an inducing cross-covariance may require a prohibitively large RKHS norm. This issue already appears in the construction of Dutordoir et al. (2021, Section 4.3): an exact ReLU feature has infinite norm in the relevant arc-cosine RKHS, while truncated approximations have finite but increasingly large norms as the truncation is reduced. Thus, making a neuron admissible does not by itself make it variationally inexpensive.

The RKHS norm enters the SVGP objective directly. For the induced posterior mean $h(x) = k_u(x)^\top K_{uu}^{-1} m$, the mean-dependent term in the KL divergence is

$$m^\top K_{uu}^{-1} m = \|h\|_{\mathcal{H}_K}^2.$$

Hence a neural-network mean may be exactly representable yet incur a large KL penalty. In particular, replacing an infinite RKHS norm by a very large but finite one need not make the corresponding mean competitive under the ELBO. This motivates studying not only whether neurons belong to a suitable RKHS, but how their RKHS cost behaves in high dimension.

Related high-dimensional limitations for kernel methods were established by Ghorbani et al. (2021). For regression with inputs uniformly distributed on a high-dimensional sphere, they show that

random-feature and neural-tangent methods with width polynomial in d, and rotationally invariant kernel methods with sample size polynomial in d, can effectively learn only bounded-degree polynomial components of a general target function, with the accessible degree increasing only as the corresponding polynomial budget in d increases.

Our results establish a different high-dimensional obstruction. Rather than fixing a random-feature or rotationally invariant kernel method and asking which components of a target function it can learn, we ask how cheaply an individual neuron can be represented by any fixed trace-class candidate prior. Theorem 5 shows that, for a broad class of smooth non-polynomial activations, the scale-invariant RKHS cost of a random neuron grows superpolynomially with dimension for every fixed sequence of trace-class candidate priors. We then show that this obstruction persists for random finite-width neural-network means and translate it into a Gaussian-regression ELBO penalty in Sections 4 and 5.

3 RKHS Membership of Neurons

3.1 Covariance Operators and Trace-Class Priors

From this point onward, we fix a probability measure ν on the input space $\mathbb{R}^d$, which serves as a reference population distribution over inputs. Such a measure allows us, for a given GP kernel K, to define the associated covariance operator $T : L^2(\nu) \to L^2(\nu)$

$$(Tf)(x) = \int K(x, x')f(x')\nu(dx') \tag{10}$$

Moreover, we'll restrict our attention to kernels that give a *trace-class* covariance operator under this measure. These are operators satisfying

$$\mathrm{tr}(T) = \int K(x, x)\nu(dx) = \mathbb{E}\|f\|^2_{L^2(\nu)} < \infty \tag{11}$$

As shown, this basic requirement is equivalent to demanding that the GP has finite expected variance over the measure ν. For the GP priors typically used in statistical learning, with respect to a probability measure ν on inputs, trace-class covariance operators are the natural setting. Since kernels are PSD, we also obtain that trace-class covariance operators are positive, $T \geq 0$, and consequently they are also self-adjoint, compact, and bounded (Simon, 2005).

We will further take K to be continuous and for ν to have full support over $\mathcal{X} = \mathbb{R}^d$. These conditions allow us to use the canonical isometric identification of $\mathcal{H}_K$ with

$$\mathcal{H}_T = \mathrm{Ran}(T^{1/2}) \subset L^2(\nu) \tag{12}$$

This identification is standard in RKHS theory (Steinwart & Christmann, 2008), with further details given in Appendix A.2.

Additionally, let Θ denote the neuron parameter space, writing σ_θ for the neuron associated with $\theta \in \Theta$. We equip Θ with a probability measure π, which weights the family of neurons being considered. Such a measure is crucial for allowing us to formulate average-cost and almost-everywhere statements over neuron parameters.

3.2 The Single-Layer NNGP Kernel

We begin by analysing the RKHS membership problem for a very concrete kernel. Perhaps the most natural candidate for a kernel tailored to a certain neuron family σ_θ is the single-layer neural-network GP kernel ("NNGP kernel"). This kernel is obtained as the infinite-width limit of a single-layer neural network with activation σ and parameter distribution π over parameters $(w, b) \in \Theta$ (Neal, 1996). Writing $\theta = (w, b)$ and $\sigma_\theta(x) = \sigma(w^\top x + b)$, the kernel is

$$K_\sigma(x, x') = \mathbb{E}_{\theta\sim\pi}[\sigma_\theta(x)\sigma_\theta(x')] \tag{13}$$

For appropriate choices of the parameter distribution π, ReLU activations yield the first-order arc-cosine kernel (Cho & Saul, 2009), while cosine activations yield the squared-exponential kernel. (Rahimi & Recht, 2007).

Although this NNGP construction is very natural for a given activation σ, it comes with a catch. We show that neurons with non-polynomial activations do not lie in the RKHS $\mathcal{H}_{K_\sigma}$ for the NNGP kernel K_σ, for any non-zero weight. More precisely, we prove the following result:

Theorem 1 (Single-neuron membership in the NNGP RKHS)**.** *Let $d \geq 2$, let $\Theta = \mathbb{R}^d \times \mathbb{R}$, and write $\theta = (w, b) \in \Theta$. Let $\pi = \pi_w \otimes \pi_b$, where π_w is a non-degenerate Gaussian measure on $\mathbb{R}^d$ and π_b is either a non-degenerate Gaussian measure on $\mathbb{R}$ or the uniform measure on $[a, c]$, with $a < c$. Suppose that $\sigma \colon \mathbb{R} \to \mathbb{R}$ is continuous and has at most polynomial growth, and define*

$$K_\sigma(x, x') := \int_\Theta \sigma(w^\top x + b)\sigma(w^\top x' + b)\, \mathrm{d}\pi(w, b).$$

Then, for every $w_0 \neq 0$ and every $b_0 \in \operatorname{supp}(\pi_b)$,

$$\big[x \mapsto \sigma(w_0^\top x + b_0)\big] \in \mathcal{H}_{K_\sigma} \quad \iff \quad \sigma \text{ is a polynomial}.$$

The proof of this can be found in Appendix B.

3.3 A Spectral Criterion for Candidate Priors

As we have just established, the issue with the NNGP kernel is that the desired neurons have infinite RKHS norm and so cannot be used as inducing cross-covariances. We now wish to consider the behaviour of the RKHS norm of neurons under general kernels. The first issue to tackle is that finite RKHS norm alone is not a meaningful measure of usability. This is because the RKHS cost of a function can always be made artificially smaller by increasing the prior variance. Specifically, if we scale our covariance operator as $T \mapsto cT$, then the trace and RKHS cost scale as

$$\operatorname{tr}(cT) = c \operatorname{tr}(T), \qquad \|f\|^2_{\mathcal{H}_{cT}} = \frac{1}{c}\|f\|^2_{\mathcal{H}_T} \tag{14}$$

Consequently, we can make every finite RKHS cost arbitrarily small just by taking $c \to \infty$. This suggests the natural scale-free quantity, the *relative* RKHS cost, as

$$C_T(\theta) := \operatorname{tr}(T)\|\sigma_\theta\|^2_{\mathcal{H}_T} \tag{15}$$

Recall that we are in the setting of trace-class kernels, where $\operatorname{tr}(T) < \infty$. We therefore ask whether a trace-class candidate kernel can have finite relative RKHS cost on a positive-measure set of neuron parameters. Our first result is the following criterion, which applies to every measurable set of parameters. The inequality is an infinite-dimensional extension of the finite-dimensional variational identity of Argyriou et al. (2006, Theorem 4.1), which yields the non-trivial trace-class existence criterion $\operatorname{tr}(T^{1/2}_{\sigma,E}) < \infty$.

Theorem 2 (Square-root trace criterion)**.** *Let $\theta \mapsto \sigma_\theta$ be a measurable family of neurons on $L^2(\nu)$, with $\int_\Theta \|\sigma_\theta\|^2_{L^2(\nu)}\, \pi(d\theta) < \infty$. Let $E \subseteq \Theta$ be any measurable set, and define the unnormalised restricted neuron covariance*

$$T_{\sigma,E} := \int_E \sigma_\theta \otimes \sigma_\theta\, \pi(d\theta).$$

For every positive trace-class candidate covariance operator T, define $\|\sigma_\theta\|_{\mathcal{H}_T} = +\infty$ outside $\mathcal{H}_T$, and write

$$A_E(T) := \int_E \|\sigma_\theta\|^2_{\mathcal{H}_T}\, \pi(d\theta) = \operatorname{tr}(T_{\sigma,E}T^\dagger).$$

Here $T^\dagger$ is the Moore–Penrose pseudoinverse, and the trace is understood as an extended nonnegative cost, equal to $+\infty$ if the neurons fail to lie in $\mathcal{H}_T$ almost everywhere on E. For $T \neq 0$,

$$\operatorname{tr}(T) \int_E \|\sigma_\theta\|^2_{\mathcal{H}_T}\, \pi(d\theta) \geq \big[\operatorname{tr}(T^{1/2}_{\sigma,E})\big]^2. \tag{SR-E}$$

The inequality is understood in the extended sense, allowing $+\infty$.

Consequently,

$$\exists\, T \geq 0 : \quad \operatorname{tr}(T) < \infty \quad \textit{and} \quad A_E(T) < \infty \qquad \iff \qquad \operatorname{tr}(T^{1/2}_{\sigma,E}) < \infty.$$

When $T_{\sigma,E} \neq 0$ and the right-hand side is finite, equality in (SR-E) *is attained by $T = cT^{1/2}_{\sigma,E}$ for any $c > 0$, with T zero on $\ker T_{\sigma,E}$. If $T_{\sigma,E} = 0$, then $A_E(T) = 0$ for every non-zero positive trace-class T.*

Corollary 3 (Full-distribution square-root trace criterion)**.** *Under the assumptions of Theorem 2, take $E = \Theta$ and write $T_\sigma = T_{\sigma,\Theta}$ and $A(T) = A_\Theta(T) = \mathbb{E}_{\theta\sim\pi}[\|\sigma_\theta\|^2_{\mathcal{H}_T}]$. Then every non-zero positive trace-class T satisfies*

$$\mathrm{tr}(T)A(T) \geq \left[\mathrm{tr}(T_\sigma^{1/2})\right]^2. \tag{SR}$$

A positive trace-class covariance with $A(T) < \infty$ exists if and only if $\mathrm{tr}(T_\sigma^{1/2}) < \infty$. When $T_\sigma \neq 0$ and this trace is finite, equality is attained by $T = cT_\sigma^{1/2}$ for any $c > 0$.

The proofs of Theorem 2 and Corollary 3 are given in Appendix C. The restricted inequality equation SR-E is essential for turning spectral divergence into an almost-everywhere obstruction. Suppose that $\mathrm{tr}(T_{\sigma,E}^{1/2}) = \infty$ for every measurable E with $\pi(E) > 0$. For any fixed non-zero positive trace-class T, each bounded-cost set

$$E_m := \{\theta : \mathrm{tr}(T)\|\sigma_\theta\|^2_{\mathcal{H}_T} \leq m\}$$

must then have measure zero by equation SR-E. This implies almost every neuron lies outside $\mathcal{H}_T$, since $\{\theta : \sigma_\theta \in \mathcal{H}_T\} = \bigcup_{m=1}^\infty E_m$.

3.4 Neural Activations with a Finite Derivative Jump

The previous section shows that the relevant spectral quantities are therefore the restricted square-root traces $\mathrm{tr}(T_{\sigma,E}^{1/2})$ over positive-measure sets E. The full-distribution trace of Corollary 3 characterises finite expected cost, while divergence on every such E rules out even a positive-measure subfamily with finite RKHS cost. For activations with a finite jump at some derivative order, this stronger obstruction holds above a dimension threshold, made precise in the following theorem.

Theorem 4 (Almost-everywhere finite-jump obstruction)**.** *Let ν be a non-degenerate Gaussian measure on $\mathbb{R}^d$, and let π be a non-degenerate Gaussian measure on $\Theta = \mathbb{R}^d \times \mathbb{R}$. For $\theta = (w, b)$, set*

$$\sigma_\theta(x) = \sigma(w^\top x + b).$$

Suppose that σ has at most polynomial growth and is smooth on each side of some $t_0 \in \mathbb{R}$. Let $s \geq 0$ be an integer. Assume in addition that the two one-sided restrictions admit C^N extensions to t_0 for some integer $N \geq s$ with $N > d/2$, and thus all one-sided derivatives through order N have finite limits at the breakpoint. Assume that the first non-zero derivative jump occurs at order s, meaning

$$\sigma^{(k)}(t_0+) = \sigma^{(k)}(t_0-) \quad (k < s), \qquad \sigma^{(s)}(t_0+) \neq \sigma^{(s)}(t_0-).$$

If $d \geq 2s + 1$, then, for every fixed positive trace-class covariance operator T on $L^2(\nu)$,

$$\pi\{\theta : \sigma_\theta \in \mathcal{H}_T\} = 0.$$

Equivalently,

$$\|\sigma_\theta\|_{\mathcal{H}_T} = \infty \qquad \text{for } \pi\text{-almost every } \theta.$$

This theorem definitively rules out the existence of RKHSs containing a positive-measure set of neurons with certain common activation functions, such as ReLU and Leaky ReLU (for $d \geq 3$) and ELU (for $d \geq 5$). It does not rule out certain exceptional neurons lying in the RKHS, nor does it rule out the construction of a prior around a particular measure-zero set of neurons.

However, this obstruction for neurons whose activation functions have a finite jump at some order clearly does not apply to neurons with smooth activations. In fact, membership is possible in suitable RKHSs. For example, the results of Dick et al. (2017) imply that tanh, sigmoid, and Softplus ridge functions belong to certain RKHSs with trace-class covariance under a Gaussian measure ν on inputs, in every dimension. Although the RKHS cost of these functions can therefore be finite, there is the remaining question of how this cost scales with dimension and whether it remains moderate enough for practical variational inference.

4 High-Dimensional RKHS Cost for Smooth Non-Polynomial Activations

We now consider neurons with smooth non-polynomial activations and ask how their scale-invariant RKHS cost grows with the input dimension. For the theorem statement, we specialise to standard

Gaussian fan-in scaling, with input measure $\nu_d = \mathcal{N}(0, I_d)$ and weight covariance $(\sigma_w^2/d)I_d$ for a fixed $\sigma_w > 0$.

The proof applies more generally to spherically symmetric weight laws whose radii concentrate at a positive constant and satisfy suitable uniform moment bounds. We give the precise assumptions and prove the result in this broader setting in Appendix E, together with the required Hermite and Gaussian-chaos decompositions. Gaussian fan-in weights satisfy these assumptions because $\|w\| \to \sigma_w$ in probability and their radial moments of every fixed order are uniformly bounded.

Theorem 5 (High-dimensional non-polynomial cost obstruction)**.** *Suppose that σ is smooth, non-polynomial, and has at most polynomial growth. Fix $\sigma_w > 0$ and a bias law that is either non-degenerate Gaussian or uniform on a bounded interval of positive length. For each $d \geq 1$, let $\nu_d = \mathcal{N}(0, I_d)$ and let π_d be the law of $\theta = (w, b)$, where*

$$w \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{d} I_d\right)$$

is independent of b, whose law is fixed across dimensions. Write $\sigma_\theta(x) = \sigma(w^\top x + b)$.

Let $(T_d)_{d\geq 1}$ be any fixed sequence of non-zero positive trace-class covariance operators on $L^2(\nu_d)$, and write $C_d(\theta; T_d) := C_{T_d}(\theta)$ for the relative RKHS cost defined previously. Then, for every $p > 0$ and every $M < \infty$,

$$\pi_d\big\{\theta : C_d(\theta; T_d) \leq M d^p\big\} \longrightarrow 0. \tag{AE}$$

Equivalently, for every $p > 0$,

$$\frac{C_d(\theta; T_d)}{d^p} \longrightarrow \infty \qquad \textit{in } \pi_d\textit{-probability}.$$

5 Variational Cost of Neural-Network Means

We now connect high-dimensional RKHS cost to variational inference. For a fixed prior, the Gaussian regression ELBO imposes a deterministic budget on the posterior mean. Under a polynomial combined prior-variance and data budget, a typical random network exceeds this budget with high probability, independently of its finite width. Proofs are given in Appendix F.

5.1 The Observed-Data Mean Budget

Let a centred GP prior have covariance operator T_d, RKHS $\mathcal{H}_{T_d}$ and integrated variance $v_d := \mathrm{tr}(T_d) > 0$. Fix observed data (X, y) and Gaussian noise variance $\sigma_d^2 > 0$. For any finite collection of admissible inducing variables, compare the finite-KL posteriors

$$q(u) = \mathcal{N}(m, S), \qquad q^\circ(u) = \mathcal{N}(0, S).$$

Write $h(x) := \mathbb{E}_{q(f)}[f(x)] = k_u(x)^\top K_{uu}^{-1} m$ for the induced mean and $h_X = (h(x_1), \ldots, h(x_N))^\top$ for its evaluations at $X = (x_1, \ldots, x_N)$. The Loève isometry in Appendix A.1 gives $J_u := m^\top K_{uu}^{-1} m = \|h\|^2_{\mathcal{H}_{T_d}}$.

Proposition 6 (Deterministic ELBO mean budget)**.** *For every observed dataset,*

$$\mathcal{L}(q^\circ) - \mathcal{L}(q) = \frac{1}{2} J_u + \frac{\|y - h_X\|^2 - \|y\|^2}{2\sigma_d^2} \geq \frac{1}{2}\|h\|^2_{\mathcal{H}_{T_d}} - \frac{\|y\|^2}{2\sigma_d^2}. \tag{16}$$

Consequently,

$$\mathcal{L}(q) \geq \mathcal{L}(q^\circ) \quad \Longrightarrow \quad v_d \|h\|^2_{\mathcal{H}_{T_d}} \leq v_d \frac{\|y\|^2}{\sigma_d^2}. \tag{17}$$

This holds for all finite-KL posteriors and data-dependent inducing variables. Remove redundant inducing variables before taking inverses, as justified in Appendix F.1.

5.2 High-Probability Exclusion of Random Network Means

Assume Theorem 5's activation and parameter hypotheses. Fix non-zero positive trace-class operators (T_d) independently of the target network. For deterministic finite widths $L_d \geq 1$, take the hidden parameters and Gaussian output weights below to be mutually independent:

$$g_d(x) = \frac{1}{\sqrt{L_d}} \sum_{\ell=1}^{L_d} a_\ell \sigma_{\theta_\ell}(x), \qquad \theta_\ell \overset{\text{iid}}{\sim} \pi_d, \qquad a \sim \mathcal{N}(0, I_{L_d}), \tag{18}$$

Proposition 7 (High-probability network cost)**.** *Under these assumptions, for every $p > 0$,*

$$\frac{v_d \|g_d\|^2_{\mathcal{H}_{T_d}}}{d^p} \longrightarrow \infty \qquad \textit{in probability}. \tag{19}$$

The proof controls the lower tail of a Gaussian quadratic form, ruling out cancellation of expensive neurons with high probability.

Theorem 8 (Observed-data exclusion under a combined budget)**.** *Under the assumptions of Proposition 7, let $0 \leq B_d \leq Cd^r$ for some $C > 0$ and $r \geq 0$. With probability tending to one over g_d, every finite-KL inducing posterior with mean g_d and every finite dataset satisfying*

$$v_d \frac{\|y\|^2}{\sigma_d^2} \leq B_d \tag{20}$$

obey

$$v_d \big[\mathcal{L}(q^\circ) - \mathcal{L}(q)\big] \geq \frac{1}{2}\Big(v_d \|g_d\|^2_{\mathcal{H}_{T_d}} - B_d\Big) > 0. \tag{21}$$

For every $p > 0$, the lower bound divided by d^p tends to infinity in probability. If $g_d \notin \mathcal{H}_{T_d}$, no finite-KL inducing posterior with this mean exists.

6 Conclusion

In this paper, we have studied when SVGPs can reproduce the mean computation of feed-forward neural networks through activation-shaped inducing cross-covariances. Our results identify two distinct obstacles to this construction. First, the desired neurons must belong to the RKHS of the GP prior in order to arise from valid finite-variance inducing variables. Second, even when membership is satisfied, their RKHS norms determine how costly the resulting mean is under the ELBO.

These constraints are restrictive in high dimensions. Under its stated hypotheses, Theorem 1 shows that a neuron with non-zero weight belongs to its own NNGP RKHS only for polynomial activations. Theorem 4 excludes almost every neuron from any fixed trace-class candidate RKHS when the first non-zero derivative jump has order s and $d \geq 2s + 1$. Smooth non-polynomial activations avoid this jump obstruction, and suitable trace-class priors can contain their neurons. Under Theorem 5's assumptions, however, typical neurons have superpolynomial scale-invariant RKHS cost in probability. This extends to random finite-width network means. Under Theorem 8's combined budget, such means are inadmissible or lose to their centred counterparts. The ELBO gap multiplied by v_d grows superpolynomially; the unscaled gap does too if v_d is bounded above by a polynomial in d.

The resulting picture is therefore not that activation-shaped inducing variables are intrinsically unsuitable for sparse GP inference. Rather, reproducing a prescribed neural-network mean is likely to become fundamentally expensive. An optimiser free to choose its posterior mean may instead favour combinations of neurons that are cheap under the prior RKHS. Our results delineate where the exact neural-network correspondence is feasible and suggest that it is most promising in low-dimensional settings or when exact reproduction of a pre-specified network is not required.

AI use statement

In this work, we used generative AI tools to assist in the writing and streamlining of the proofs in the appendices. The authors confirm that they have reviewed and edited these proofs as necessary and take full responsibility for the work's correctness. Additionally, we used generative AI tools to assist in minor aspects of the writing and editing of the main text, such as spelling and grammar checks and improvements to clarity of exposition.

REFERENCES

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems*, volume 19, 2006. URL https://home.ttic.edu/~argyriou/papers/nips06_online.pdf.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. doi: 10.1090/S0002-9947-1950-0051437-7.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. URL https://jmlr.org/papers/v18/14-546.html.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Statistics for Engineering and Information Science. Kluwer Academic Publishers, Boston, 2004.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1905.12173.

David R. Burt. Spectral methods in gaussian process approximations. Mphil thesis, University of Cambridge, 2018.

Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, volume 22, 2009. URL https://papers.nips.cc/paper/3628-kernel-methods-for-deep-learning.pdf.

Feng Dai and Yuan Xu. *Approximation Theory and Harmonic Analysis on Spheres and Balls*. Springer, 2013. doi: 10.1007/978-1-4614-6660-4.

Andreas Damianou and Neil D. Lawrence. Deep Gaussian processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pp. 207–215, 2013. URL https://proceedings.mlr.press/v31/damianou13a.html.

Julio Delgado and Michael Ruzhansky. Schatten classes on compact manifolds: Kernel conditions. *Journal of Functional Analysis*, 267(3):772–798, 2014. doi: 10.1016/j.jfa.2014.04.016. URL https://arxiv.org/abs/1403.6158.

Josef Dick, Christian Irrgeher, Gunther Leobacher, and Friedrich Pillichshammer. On the optimal order of integration in hermite spaces with finite smoothness, 2017. URL https://arxiv.org/abs/1608.06061.

Vincent Dutordoir, James Hensman, Mark van der Wilk, Carl Henrik Ek, Zoubin Ghahramani, and Nicolas Durrande. Deep neural networks as point estimates for deep Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2105.04504.

Semyon Dyatlov. Lecture notes for 18.155: Distributions, elliptic regularity, and applications to PDEs, 2022. URL https://math.mit.edu/~dyatlov/18.155/155-notes.pdf. Massachusetts Institute of Technology, December 10, 2022.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021. URL https://arxiv.org/abs/1904.12191.

Friedrich Haslinger. *Complex Analysis: A Functional Analytic Approach*. De Gruyter, Berlin, Boston, 2018. doi: 10.1515/9783110417241.

James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 282–290, 2013. URL https://auai.org/uai2013/prints/papers/244.pdf.

James Hensman, Nicolas Durrande, and Arno Solin. Variational Fourier features for Gaussian processes. *Journal of Machine Learning Research*, 18(151):1–52, 2018. URL https://jmlr.org/papers/v18/16-579.html.

David Holzmüller and Max Schölpple. Beyond ReLU: How activations affect neural kernels and random wide networks. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3403–3411, 2026. URL https://proceedings.mlr.press/v300/holzmuller26a.html.

Tuomas Hytönen, Jan van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach Spaces: Volume I: Martingales and Littlewood–Paley Theory*, volume 63 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics*. Springer, Cham, 2016. ISBN 978-3-319-48519-5. doi: 10.1007/978-3-319-48520-1.

Svante Janson. *Gaussian Hilbert Spaces*, volume 129 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 1997. doi: 10.1017/CBO9780511526169.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K. Sriperumbudur. Gaussian processes and reproducing kernels: Connections and equivalences, 2025.

Miguel Lázaro-Gredilla and Aníbal Figueiras-Vidal. Inter-domain Gaussian processes for sparse inference using inducing features. In *Advances in Neural Information Processing Systems*, volume 22, 2009. URL https://papers.nips.cc/paper_files/paper/2009/hash/5ea1649a31336092c05438df996a3e59-Abstract.html.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1711.00165.

Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1804.11271.

Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer, New York, 1996. doi: 10.1007/978-1-4612-0745-0.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. URL https://gaussianprocess.org/gpml/.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1705.08933.

Barry Simon. *Trace Ideals and Their Applications*, volume 120 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2 edition, 2005.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008. doi: 10.1007/978-0-387-77242-4.

Shengyang Sun, Jiaxin Shi, and Roger Grosse. Neural networks as inter-domain inducing points. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2020. URL https://openreview.net/forum?id=NgqYp7sAW6t.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 567–574, 2009. URL https://proceedings.mlr.press/v5/titsias09a.html.

A. W. van der Vaart and J. H. van Zanten. Reproducing kernel hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, volume 3 of *IMS Collections*, pp. 200–222. Institute of Mathematical Statistics, 2008. doi: 10.1214/074921708000000156. URL https://arxiv.org/abs/0805.3252.

Christopher K. I. Williams. Computing with infinite networks. In *Advances in Neural Information Processing Systems 9*, pp. 295–301. MIT Press, 1997.

A RKHS AND INDUCING-VARIABLE PRELIMINARIES

A.1 ADMISSIBLE INDUCING VARIABLES AND THE LOÈVE ISOMETRY

Let f be a centred GP with covariance kernel K, and let

$$\mathcal{G}_f := \overline{\operatorname{span}\{f(x) : x \in \mathcal{X}\}}^{L^2(\Omega)}$$

be its Gaussian Hilbert space. The map

$$\mathcal{L} : \mathcal{G}_f \to \mathcal{H}_K, \qquad (\mathcal{L}u)(x) = \mathbb{E}[uf(x)],$$

known as the Loève isometry, is an isometric isomorphism (Berlinet & Thomas-Agnan, 2004; van der Vaart & van Zanten, 2008). Every admissible inducing variable $u \in \mathcal{G}_f$ therefore corresponds uniquely to $\psi_u = \mathcal{L}u \in \mathcal{H}_K$, with

$$\operatorname{Cov}(u, f(x)) = \psi_u(x), \qquad \operatorname{Var}(u) = \|\psi_u\|^2_{\mathcal{H}_K}.$$

Conversely, every RKHS function has a unique preimage in $\mathcal{G}_f$, where random variables are identified up to almost-sure equality. We now give a short proof of this isometry and show how the typical interdomain integral inducing variable fits into this framework.

Proof of the isometry. First define the map on finite linear combinations by

$$\mathcal{L}_0\left(\sum_i a_i f(x_i)\right) := \sum_i a_i K(x_i, \cdot).$$

The RKHS inner product gives

$$\left\|\sum_i a_i K(x_i, \cdot)\right\|^2_{\mathcal{H}_K} = \sum_{i,j} a_i a_j K(x_i, x_j) = \mathbb{E}\left[\left(\sum_i a_i f(x_i)\right)^2\right].$$

We first check that this map is well-defined. Taking any two representations of the same random variable $u \in L^2(\Omega)$

$$u = \sum_i a_i f(x_i) = \sum_j b_j f(y_j)$$

we must check that both representations give the same RKHS function under $\mathcal{L}_0$. Writing the difference of the two representations as a single sum, $v = \sum_i a_i f(x_i) - \sum_j b_j f(y_j) = \sum_k c_k f(z_k) = 0$, we have $\|v\|^2_{L^2(\Omega)} = \mathbb{E}[v^2] = 0$. Applying the above norm identity to this linear combination gives $\|\sum_k c_k K(z_k, \cdot)\|^2_{\mathcal{H}_K} = \mathbb{E}[v^2] = 0$, so $\sum_k c_k K(z_k, \cdot) = 0$ in the RKHS. Thus both representations give the same RKHS function under $\mathcal{L}_0$, and our map is well-defined. It is also linear and the same identity shows that it preserves norms.

Now take $u \in \mathcal{G}_f$ and choose finite sums u_n converging to u in $L^2(\Omega)$. Since

$$\|\mathcal{L}_0 u_n - \mathcal{L}_0 u_m\|_{\mathcal{H}_K} = \|u_n - u_m\|_{L^2(\Omega)},$$

the images form a Cauchy sequence and converge to some $\psi \in \mathcal{H}_K$ by completeness. So, we can define $\mathcal{L} : \mathcal{G}_f \to \mathcal{H}_K$ by $\mathcal{L}u := \lim_{n\to\infty} \mathcal{L}_0 u_n$. Again this definition does not depend on the sequence, as if $v_n \to u$ as well, then $\|\mathcal{L}_0 u_n - \mathcal{L}_0 v_n\|_{\mathcal{H}_K} = \|u_n - v_n\|_{L^2(\Omega)} \to 0$, so both sequences $\mathcal{L}_0 u_n$ and $\mathcal{L}_0 v_n$ have the same limit. Taking limits also shows that $\mathcal{L}$ is linear and $\|\mathcal{L}u\|_{\mathcal{H}_K} = \|u\|_{L^2(\Omega)}$. This is the extension by continuity, and it is unique because the finite sums are dense in $\mathcal{G}_f$.

From the definition of $\mathcal{L}_0$, we have $(\mathcal{L}_0 u_n)(x) = \mathbb{E}[u_n f(x)]$ for each n, and we now show that the limit of the left-hand side is the same as the limit of the right-hand side. Fix $x \in \mathcal{X}$. The reproducing

property makes evaluation at x continuous in RKHS norm, giving $|(\mathcal{L}u)(x) - (\mathcal{L}_0 u_n)(x)| \leq \|\mathcal{L}u - \mathcal{L}_0 u_n\|_{\mathcal{H}_K} \sqrt{K(x,x)} \to 0$. On the other hand, Cauchy–Schwarz gives

$$|\mathbb{E}[(u_n - u)f(x)]| \leq \|u_n - u\|_{L^2(\Omega)} \sqrt{K(x,x)} \longrightarrow 0.$$

Consequently,

$$(\mathcal{L}u)(x) = \lim_n (\mathcal{L}_0 u_n)(x) = \lim_n \mathbb{E}[u_n f(x)] = \mathbb{E}[u f(x)]$$

as desired. Every $u \in \mathcal{G}_f$ is centred, since it is an L^2 limit of centred variables. This proves the covariance and variance identities, $\mathrm{Cov}(u, f(x)) = \mathbb{E}[uf(x)] = (\mathcal{L}u)(x)$ and $\mathrm{Var}(u) = \|\mathcal{L}u\|^2_{\mathcal{H}_K}$.

Finally, the range of $\mathcal{L}$ is closed because it is the isometric image of the complete space $\mathcal{G}_f$. It contains every kernel section, as $\mathcal{L}f(x) = K(x, \cdot)$, and hence their dense linear span. The range is therefore all of $\mathcal{H}_K$. Thus the map is surjective. The isometry also makes $\mathcal{L}$ injective, proving the claimed one-to-one correspondence. □

For the usual formulation of interdomain variables (Lázaro-Gredilla & Figueiras-Vidal, 2009; Hensman et al., 2018),

$$u = \int_{\mathcal{X}} q(x) f(x)\, \mu(dx),$$

interpret the integral as a mean-square limit of finite weighted sums of GP values. Each sum belongs to $\mathcal{G}_f$, so its limit does too by closedness. Applying $\mathcal{L}$ to these sums gives

$$\psi_u = \int_{\mathcal{X}} q(x) K(x, \cdot)\, \mu(dx),$$

with this integral understood as the corresponding limit in RKHS norm.

A.2 Covariance Operators and RKHS Range Spaces

For a positive trace-class operator T, its RKHS is

$$\mathcal{H}_T = \mathrm{Ran}(T^{1/2}), \qquad \|f\|^2_{\mathcal{H}_T} = \|T^{\dagger/2} f\|^2_{L^2(\nu)},$$

where $T^\dagger$ denotes the Moore–Penrose pseudoinverse. The norm is defined to be $+\infty$ outside $\mathrm{Ran}(T^{1/2})$. Explicitly, if (t_j, e_j) are the positive eigenpairs of T, then

$$\|f\|^2_{\mathcal{H}_T} = \begin{cases} \displaystyle\sum_j \frac{|\langle f, e_j\rangle_{L^2(\nu)}|^2}{t_j}, & f \perp \ker T, \\ +\infty, & \text{otherwise}, \end{cases}$$

where a divergent series also has value $+\infty$. This extended RKHS cost definition makes the cost a measurable function for all $f \in L^2(\nu)$.

The inequalities below are formulated in terms of covariance operators, whereas our application concerns the RKHS norm associated with a kernel. We therefore take a moment to explain why the operator-defined space $\mathcal{H}_T$ recovers that RKHS and its norm. This identification lets us use operator inequalities to study the RKHS costs appearing in Theorem 2.

The elements of $L^2(\nu)$ are equivalence classes, so functions that agree ν-almost everywhere represent the same element. By contrast, the usual RKHS $\mathcal{H}_K$ of a kernel K consists of functions with specified values at every input. For this identification, suppose that K is continuous, $\int K(x,x)\, \nu(dx) < \infty$, and ν has full support, meaning that every non-empty open set has positive measure. Define

$$J : \mathcal{H}_K \longrightarrow L^2(\nu), \qquad Jf = [f],$$

where $[f]$ denotes the almost-everywhere equivalence class of f. The reproducing property gives

$$\|Jf\|^2_{L^2(\nu)} \leq \|f\|^2_{\mathcal{H}_K} \int K(x,x)\, \nu(dx),$$

so J is well-defined and bounded. It is also injective. Indeed, continuity of K implies continuity of every $f \in \mathcal{H}_K$. If such an f is non-zero at a point, continuity makes it non-zero on an open

neighbourhood, which has positive measure by full support. Thus an RKHS function that vanishes almost everywhere must vanish everywhere, and passing to $L^2(\nu)$ loses no information about it.

Let T now be the integral operator associated with K,

$$(Tg)(x) = \int K(x, y)g(y)\,\nu(dy).$$

The reproducing property yields $T = JJ^*$. The range identity for a bounded operator and its adjoint then gives

$$J(\mathcal{H}_K) = \mathrm{Ran}(J) = \mathrm{Ran}((JJ^*)^{1/2}) = \mathrm{Ran}(T^{1/2}) = \mathcal{H}_T.$$

The same identity preserves the minimum preimage norm:

$$\|g\|_{\mathcal{H}_T} = \inf\{\|f\|_{\mathcal{H}_K} : Jf = g\}.$$

One way to see both statements is through the singular-value decomposition: J and $(JJ^*)^{1/2}$ have the same non-zero singular values and output vectors, so their ranges and minimum preimage norms coincide. Since J is injective, each g in its range has a unique preimage, and hence

$$\|Jf\|_{\mathcal{H}_T} = \|f\|_{\mathcal{H}_K}.$$

This is equality of the two RKHS norms, not equality with the ordinary $L^2(\nu)$ norm. Thus $\mathcal{H}_T$ consists precisely of the equivalence classes having an RKHS representative, with that representative's RKHS norm. The continuity and full-support assumptions above justify this pointwise interpretation. The operator inequalities below use only the range-space definition and apply to general positive trace-class operators.

B Single-Neuron Membership in the NNGP RKHS

Proof overview

Theorem 1 is proved directly, without auxiliary lemmas. The first step translates RKHS membership into the existence of a parameter-space representer $\psi \in L^2(\pi)$ satisfying $\sigma(w_0^\top x + b_0) = \int \psi(w, b)\sigma(w^\top x + b)\,d\pi(w, b)$ for every x. If σ is polynomial, the parameter-space features span a finite-dimensional space. Evaluation at (w_0, b_0) is then a bounded linear functional, so the Riesz representation theorem supplies such a ψ.

For non-polynomial σ, the proof rules out this representation by Fourier testing. The target neuron's Fourier transform is supported on $L = \mathrm{span}(w_0)$ and has non-zero support away from the origin. Step 1 constructs tests Φ_ε supported in shrinking angular neighbourhoods of L, away from the origin, whose response to the target is a fixed non-zero constant. Step 2 proves equation 23: a neuron's response is zero unless its weight direction lies in that neighbourhood, and otherwise is bounded by a polynomial in $\|w\| + |b|$, uniformly in ε. This uniform bound is the essential estimate, since the tests' transverse derivatives themselves grow as the neighbourhood shrinks.

Step 3 integrates that bound against ψ. Cauchy–Schwarz and the parameter moments give an integrable majorant, while Gaussian weights assign zero probability to L when $d \geq 2$. The mixture's response therefore tends to zero, contradicting the target's fixed non-zero response. This proves non-membership for each target covered by the theorem, rather than only for a randomly chosen target.

Proof of Theorem 1. For $x \in \mathbb{R}^d$, let

$$\phi_x(w, b) := \sigma(w^\top x + b), \qquad V := \mathrm{span}\{\phi_x : x \in \mathbb{R}^d\} \subset L^2(\pi).$$

Polynomial growth and the parameter moments ensure that $\phi_x \in L^2(\pi)$. Here ϕ_x is a function of the parameters (w, b) for a fixed input x, whereas an element of $\mathcal{H}_{K_\sigma}$ is a function of the input. The link between these two spaces is the kernel identity

$$K_\sigma(x, x') = \langle \phi_x, \phi_{x'} \rangle_{L^2(\pi)}.$$

Indeed, the correspondence $K_\sigma(\cdot, x) \mapsto \phi_x$ preserves inner products and hence norms of finite linear combinations:

$$\left\| \sum_{i=1}^{N} a_i K_\sigma(\cdot, x_i) \right\|^2_{\mathcal{H}_{K_\sigma}} = \sum_{i,j=1}^{N} a_i a_j K_\sigma(x_i, x_j) = \left\| \sum_{i=1}^{N} a_i \phi_{x_i} \right\|^2_{L^2(\pi)}.$$

It therefore extends by completion to an isometric isomorphism from $\mathcal{H}_{K_\sigma}$ onto $\overline{V}$. This is the feature-space construction of the RKHS (Aronszajn, 1950; Steinwart & Christmann, 2008). If h corresponds to $\psi \in \overline{V}$, the reproducing property gives

$$h(x) = \langle K_\sigma(\cdot, x), h \rangle_{\mathcal{H}_{K_\sigma}} = \langle \phi_x, \psi \rangle_{L^2(\pi)} = \int_\Theta \psi(w, b) \sigma(w^\top x + b)\, d\pi(w, b).$$

Conversely, every $\psi \in \overline{V}$ defines an RKHS function by this formula, with $\|h\|_{\mathcal{H}_{K_\sigma}} = \|\psi\|_{L^2(\pi)}$. We may also allow $\psi \in L^2(\pi)$, as its component orthogonal to $\overline{V}$ pairs to zero with every ϕ_x, so replacing ψ by $P_{\overline{V}}\psi$ leaves the represented function unchanged. The integral is finite for each x by Cauchy–Schwarz.

Thus, for the target $h(x) = \sigma(w_0^\top x + b_0)$, RKHS membership is exactly the existence of a single $\psi \in L^2(\pi)$ giving the displayed integral for every input x. We use this representation criterion to prove the polynomial and non-polynomial cases below.

Polynomial activations. If σ is a polynomial of degree at most $m \geq 0$, then V has dimension at most $\binom{d+1+m}{m}$, the number of parameter monomials of total degree at most m. Since $V \subset L^2(\pi)$ consists of equivalence classes of functions that are equal almost everywhere, we must first establish that evaluation on V is well-defined. Every class in V has a continuous polynomial representative. If two such representatives are equal π-almost everywhere, they are equal on the support of π, and thus they agree at $\theta_0 = (w_0, b_0) \in \operatorname{supp} \pi$. Every linear functional on a finite-dimensional normed space is bounded. The Riesz representation theorem therefore supplies $\psi \in V$ with

$$\langle \phi_x, \psi \rangle_{L^2(\pi)} = \phi_x(\theta_0) = \sigma(w_0^\top x + b_0),$$

which proves membership of $[x \mapsto \sigma(w_0^\top x + b_0)]$ in $\mathcal{H}_{K_\sigma}$.

Non-polynomial activations. Suppose instead that σ is not a polynomial and that a representer $\psi \in L^2(\pi)$ exists. Define

$$F(x) := \int \psi(w, b) \sigma(w^\top x + b)\, d\pi(w, b), \qquad G(x) := \sigma(w_0^\top x + b_0), \qquad L := \operatorname{span}(w_0).$$

We will find Fourier tests for which G has a fixed non-zero response but F has a response tending to zero. The reason is essentially geometric: a single ridge function has frequencies on a single line, whereas continuously distributed weights assign zero mass to that line.

Here Fourier transforms are understood as tempered distributions. This allows us to apply the Fourier transform to a much larger array of functions, including polynomially growing functions, such as ReLU, by testing them against smooth functions that decay faster than every polynomial. Indeed, choosing an integer m with $|\sigma(t)| \leq C(1 + |t|)^m$, Cauchy–Schwarz gives $|F(x)| + |G(x)| \leq C_\psi(1 + \|x\|)^m$. We use the convention

$$\widehat{f}(\xi) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-ix\cdot\xi} f(x)\, dx.$$

Tempered distribution pairings with test functions are written $\langle G, \Phi \rangle$, for example, and are bilinear. Fourier test functions may be complex-valued. We now move on to the body of the proof.

1. Choose a test around the target line. We first make precise why a ridge has frequencies only along its weight direction. Write $H_{w,b}(x) := \sigma(w^\top x + b)$. For $w \neq 0$, put $r := \|w\|$ and $v := w/r$, and decompose $x = sv + y$ with $y \in v^\perp$. Then

$$w^\top(sv + y) + b = rs + b, \qquad H_{w,b}(sv + y) = \sigma(rs + b).$$

Thus changing y leaves the ridge unchanged. In Fourier coordinates, this will mean that the frequency perpendicular to v must be zero. Since the ridge need not be integrable, we establish this statement by pairing with a test, rather than by taking its Fourier integral pointwise.

For a smooth compactly supported test function Ψ, our bilinear convention gives

$$\langle \widehat{H}_{w,b}, \Psi \rangle = \langle H_{w,b}, \widehat{\Psi} \rangle = \int_{\mathbb{R}^d} H_{w,b}(x) \widehat{\Psi}(x)\, dx.$$

Here $\widehat{\Psi}$ decays faster than any polynomial, so this integral is absolutely convergent. Explicitly, integration by parts gives, for every integer $N \geq 0$,

$$(1 + \|x\|^2)^N |\widehat{\Psi}(x)| = \left|\widehat{(1 - \Delta)^N \Psi}(x)\right| \leq (2\pi)^{-d/2} \|(1 - \Delta)^N \Psi\|_{L^1},$$

where $\Delta = \sum_{j=1}^d \partial_{\xi_j}^2$ differentiates the test function. For fixed (w, b), polynomial growth gives $|H_{w,b}(x)| \leq C_{w,b}(1 + \|x\|)^m$. Hence, outside the unit ball,

$$|H_{w,b}(x) \widehat{\Psi}(x)| \leq C \|x\|^{m-2N}.$$

In polar coordinates, the volume element contributes a factor r^{d-1}, so

$$\int_{\|x\| \geq 1} |H_{w,b}(x) \widehat{\Psi}(x)|\, dx \leq C |\mathbb{S}^{d-1}| \int_1^\infty r^{m-2N+d-1}\, dr,$$

where $|\mathbb{S}^{d-1}|$ is the surface area of the unit sphere. The last integral is finite precisely when its exponent is less than -1, that is, when $m - 2N + d - 1 < -1$, or equivalently $2N > m + d$. Inside the unit ball both functions are bounded, so the integral there is also finite. We may choose N this large because Ψ is smooth and compactly supported. This establishes absolute convergence of the pairing. Now, the frequency-line support property follows from the next calculation. Defining the restriction of the test to the frequency line by $f_v(t) := \Psi(tv)$, we claim that

$$\langle \widehat{H}_{w,b}, \Psi \rangle = (2\pi)^{(d-1)/2} \int_{\mathbb{R}} \sigma(rs + b) \widehat{f_v}(s)\, ds. \tag{22}$$

To check this, write the spatial and frequency coordinates as

$$x = sv + y, \qquad \xi = tv + \eta, \qquad y, \eta \in v^\perp.$$

Here s, t are the coordinates along v, while y, η are the perpendicular coordinates. Since $\|v\| = 1$ and v is perpendicular to both y and η, the cross terms vanish:

$$x \cdot \xi = (sv + y) \cdot (tv + \eta) = st + y \cdot \eta.$$

The orthogonal frequency change of coordinates has $d\xi = dt\, d\eta$. Substitution into the definition of the Fourier transform therefore gives

$$\widehat{\Psi}(sv + y) = (2\pi)^{-d/2} \int_{v^\perp} \int_{\mathbb{R}} e^{-ist} e^{-iy \cdot \eta} \Psi(tv + \eta)\, dt\, d\eta.$$

We can compute this transform in two stages. First, transform only in t, leaving η fixed, and call the result

$$q_s(\eta) := (2\pi)^{-1/2} \int_{\mathbb{R}} e^{-ist} \Psi(tv + \eta)\, dt.$$

Factoring $(2\pi)^{-d/2} = (2\pi)^{-1/2} (2\pi)^{-(d-1)/2}$ in the double integral then yields

$$\widehat{\Psi}(sv + y) = (2\pi)^{-(d-1)/2} \int_{v^\perp} e^{-iy \cdot \eta} q_s(\eta)\, d\eta.$$

Thus, for each fixed s, the function of y on the left is the $(d-1)$-dimensional Fourier transform of the function $\eta \mapsto q_s(\eta)$. Fourier inversion reconstructs that function:

$$q_s(\eta) = (2\pi)^{-(d-1)/2} \int_{v^\perp} e^{iy \cdot \eta} \widehat{\Psi}(sv + y)\, dy.$$

At $\eta = 0$, the exponential equals one, so

$$\int_{v^\perp} \widehat{\Psi}(sv + y)\, dy = (2\pi)^{(d-1)/2} q_s(0).$$

Finally, evaluating the definition of q_s at zero gives

$$q_s(0) = (2\pi)^{-1/2} \int_{\mathbb{R}} e^{-ist} \Psi(tv)\, dt = \widehat{f_v}(s), \qquad f_v(t) = \Psi(tv).$$

Consequently,

$$\int_{v^\perp} \widehat{\Psi}(sv + y)\, dy = (2\pi)^{(d-1)/2} \widehat{f_v}(s).$$

In words, integrating over the perpendicular spatial coordinates leaves only the one-dimensional Fourier transform of the test restricted to the frequency line. Using $dx = ds\, dy$ and $H_{w,b}(sv+y) = \sigma(rs + b)$, we can now substitute this identity directly into the pairing:

$$\begin{aligned}
\langle \widehat{H}_{w,b}, \Psi \rangle &= \int_{\mathbb{R}^d} H_{w,b}(x) \widehat{\Psi}(x)\, dx \\
&= \int_{\mathbb{R}} \int_{v^\perp} \sigma(rs + b) \widehat{\Psi}(sv + y)\, dy\, ds \\
&= \int_{\mathbb{R}} \sigma(rs + b) \left(\int_{v^\perp} \widehat{\Psi}(sv + y)\, dy \right) ds \\
&= (2\pi)^{(d-1)/2} \int_{\mathbb{R}} \sigma(rs + b) \widehat{f_v}(s)\, ds.
\end{aligned}$$

This is precisely equation 22. All integrals in this calculation are justified by the rapid decay of the transformed tests and the polynomial growth of σ.

In particular, if Ψ is supported away from $\operatorname{span}(w)$, then $f_v(t) = \Psi(tv) = 0$ for every t, and hence

$$\langle \widehat{H}_{w,b}, \Psi \rangle = 0.$$

By definition, a distribution is supported on a closed set when it pairs to zero with every test supported outside that set. We have thus shown

$$\operatorname{supp} \widehat{H}_{w,b} \subset \operatorname{span}(w).$$

The case $w = 0$ can be ignored in the parameter integral, since $\pi_w(\{0\}) = 0$.

Next we need a test that detects a non-zero frequency of G. The relevant distributional fact is

$$\operatorname{supp} \widehat{H} \subset \{0\} \quad \Longleftrightarrow \quad H \text{ is a polynomial}$$

for continuous functions H of at most polynomial growth, see Dyatlov (2022, Theorem 4.19 and Section 11.2). Thus any non-polynomial G must have some Fourier support away from the origin.

In our case, the affine line

$$x(t) := \frac{t - b_0}{\|w_0\|^2} w_0 \quad \text{satisfies} \quad G(x(t)) = \sigma(t).$$

If G were a polynomial in x, its composition with this affine map would be a polynomial in t, contrary to the assumption on σ. Thus G is not a polynomial, and $\widehat{G}$ is not supported only at zero. By the definition of support, there is consequently a test

$$\Phi \in C_c^\infty(\mathbb{R}^d \setminus \{0\}), \qquad c := \langle \widehat{G}, \Phi \rangle \neq 0.$$

In words, Φ detects the part of the target's Fourier transform away from the origin. Fix $0 < \rho < R$ such that

$$\operatorname{supp} \Phi \subset \{\xi : \rho \leq \|\xi\| \leq R\}.$$

We now narrow this test around L while preserving its response to G. Let $v_0 := w_0/\|w_0\|$, and define the angle to the *unoriented* line L by

$$\angle(\xi, L) := \arccos\left(\frac{|\xi \cdot v_0|}{\|\xi\|} \right) \in [0, \pi/2], \qquad \xi \neq 0.$$

Both directions along L therefore have angle zero. Take a smooth function a on $[0,\infty)$, equal to one on $[0,1/2]$, zero on $[1,\infty)$, and valued in $[0,1]$. For $0<\varepsilon<\pi/4$, set

$$\Phi_\varepsilon(\xi) := \Phi(\xi)a\left(\frac{\angle(\xi,L)}{\varepsilon}\right) \quad (\xi\neq 0), \qquad \Phi_\varepsilon(0) := 0.$$

The cut-off a leaves Φ unchanged within angle $\varepsilon/2$ of L and makes it zero outside angle ε. It varies smoothly between these regions. The angle function is smooth wherever the cut-off changes value, and the cut-off is constant near the directions where the angle is not smooth. Since Φ also vanishes near the origin, Φ_ε is smooth everywhere. Its support is contained in the compact set $\operatorname{supp}\Phi$, so it remains a valid compactly supported test function.

The difference $\Phi_\varepsilon-\Phi$ vanishes on an open neighbourhood of L. Explicitly, using the radius ρ above, take

$$U := B_\rho(0)\cup\{\xi\neq 0 : \angle(\xi,L)<\varepsilon/2\}.$$

This set is open and contains all of L. On the angular region the cut-off factor is exactly one, while on $B_\rho(0)$ both tests are identically zero, so $\Phi_\varepsilon=\Phi$ throughout U. The ball is needed because $0\in L$ and the angular region alone contains no ball around the origin, since every such ball includes perpendicular directions. Since $\operatorname{supp}\widehat{G}\subset L$, it follows that

$$\langle\widehat{G},\Phi_\varepsilon-\Phi\rangle = 0, \qquad \langle\widehat{G},\Phi_\varepsilon\rangle = c \quad \text{for every } \varepsilon.$$

Agreement on a neighbourhood is useful here, as a distribution may depend on derivatives of a test function, so equality of the test functions just at points of L would not in general suffice.

2. Bound the response of every other ridge. The angular cut-off factor can have large derivatives across L as ε decreases. The key is that equation 22 only involves the restriction of a test *along* each ridge's frequency line. For $w=rv\neq 0$, set

$$f_{\varepsilon,v}(t) := \Phi_\varepsilon(tv), \qquad a_{\varepsilon,v} := a\left(\frac{\angle(v,L)}{\varepsilon}\right).$$

Because the line is unoriented, for both positive and negative t,

$$\angle(tv,L)=\angle(v,L), \qquad f_{\varepsilon,v}(t)=a_{\varepsilon,v}\Phi(tv).$$

The second identity also holds at $t=0$, since both sides vanish there. We bound the integral of each derivative by the length of its support multiplied by its maximum size. First, since $\|v\|=1$, we have $\|tv\|=|t|$. The support condition on Φ therefore gives

$$\operatorname{supp} f_{\varepsilon,v}\subset[-R,-\rho]\cup[\rho,R].$$

Every derivative also vanishes outside these intervals, whose total length is at most $2R$.

Next, the chain rule gives

$$\frac{d}{dt}\Phi(tv)=\sum_{j=1}^{d}v_j\partial_j\Phi(tv)=(v\cdot\nabla)\Phi(tv).$$

For fixed ε, v, the angular factor $a_{\varepsilon,v}$ is constant with respect to t. Repeating the chain rule yields, for $k\geq 1$,

$$\begin{aligned} f^{(k)}_{\varepsilon,v}(t) &= a_{\varepsilon,v}(v\cdot\nabla)^k\Phi(tv) \\ &= a_{\varepsilon,v}\sum_{j_1,\ldots,j_k=1}^{d} v_{j_1}\cdots v_{j_k}\partial_{j_1}\cdots\partial_{j_k}\Phi(tv). \end{aligned}$$

Define

$$M_{k,\Phi} := \max_{j_1,\ldots,j_k\in\{1,\ldots,d\}}\sup_{\xi\in\mathbb{R}^d}|\partial_{j_1}\cdots\partial_{j_k}\Phi(\xi)|, \qquad M_{0,\Phi}:=\|\Phi\|_\infty.$$

These constants are finite because Φ is smooth and compactly supported. Since $|a_{\varepsilon,v}|\leq 1$, each $|v_j|\leq 1$, and there are d^k terms in the sum, we obtain

$$|f^{(k)}_{\varepsilon,v}(t)|\leq d^k M_{k,\Phi} \qquad (k\geq 0).$$

Combining this pointwise bound with the support length gives

$$\|f_{\varepsilon,v}^{(k)}\|_{L^1(\mathbb{R})} = \int_{-R}^{R} |f_{\varepsilon,v}^{(k)}(t)|\, dt \le 2R\, d^k M_{k,\Phi} =: C_{k,\Phi}.$$

The bound is independent of v and ε. In particular, no factors of ε^{-1} appear, since differentiating in t moves along a fixed frequency line, where the angular factor is constant, rather than between directions where that factor changes.

We now use the standard Fourier-decay estimate: for $f \in C_c^\infty(\mathbb{R})$ and every integer $N \ge 1$,

$$|\widehat{f}(s)| \le C_N(1+|s|)^{-N}\big(\|f\|_{L^1} + \|f^{(N)}\|_{L^1}\big),$$

where C_N depends only on N and the Fourier convention, see Dyatlov (2022, Section 11.1, in particular Proposition 11.7). Applying this with $N = m+2$ and the uniform derivative bounds above gives

$$|\widehat{f}_{\varepsilon,v}(s)| \le C_{m+2}(C_{0,\Phi} + C_{m+2,\Phi})(1+|s|)^{-m-2}.$$

Consequently,

$$\int_{\mathbb{R}} (1+|s|)^m |\widehat{f}_{\varepsilon,v}(s)|\, ds \le C_\Phi \int_{\mathbb{R}} (1+|s|)^{-2}\, ds < \infty,$$

with a bound independent of ε and v. This uniformity is important for controlling the constants as $\varepsilon \to 0$. Here and below C_Φ denotes a constant that may increase between estimates and depend on the fixed m, d, σ, Φ, but not on ε, v, w or b.

Apply equation 22 with $\Psi = \Phi_\varepsilon$ to obtain

$$A_\varepsilon(w,b) := \langle \widehat{H}_{w,b}, \Phi_\varepsilon \rangle = (2\pi)^{(d-1)/2} \int_{\mathbb{R}} \sigma(rs+b)\widehat{f}_{\varepsilon,v}(s)\, ds.$$

The growth bound on σ and the elementary inequality

$$1 + |rs+b| \le 1 + r|s| + |b| \le (1 + r + |b|)(1+|s|)$$

give

$$|\sigma(rs+b)| \le C(1+|rs+b|)^m \le C(1+r+|b|)^m(1+|s|)^m.$$

Taking absolute values inside the ridge-pairing integral and then using this bound, we obtain

$$\begin{aligned} |A_\varepsilon(w,b)| &\le (2\pi)^{(d-1)/2} \int_{\mathbb{R}} |\sigma(rs+b)|\, |\widehat{f}_{\varepsilon,v}(s)|\, ds \\ &\le C(2\pi)^{(d-1)/2}(1+r+|b|)^m \int_{\mathbb{R}} (1+|s|)^m |\widehat{f}_{\varepsilon,v}(s)|\, ds \\ &\le C_\Phi(1+\|w\|+|b|)^m. \end{aligned}$$

In the second line, $(1+r+|b|)^m$ is independent of s and can be moved outside the integral. The last line uses the uniform weighted Fourier bound established above and $r = \|w\|$, absorbing the fixed factors $C(2\pi)^{(d-1)/2}$ into C_Φ. Moreover, if $\angle(w, L) \ge \varepsilon$, then $a_{\varepsilon,v} = 0$, so $f_{\varepsilon,v} = 0$ and $A_\varepsilon(w,b) = 0$ exactly. Thus, defining the set of non-zero weights inside the angular neighbourhood by

$$E_\varepsilon := \{w \ne 0 : \angle(w, L) < \varepsilon\},$$

we obtain, for $w \ne 0$, the more informative bound

$$|A_\varepsilon(w,b)| \le C_\Phi(1+\|w\|+|b|)^m \mathbf{1}_{E_\varepsilon}(w). \tag{23}$$

Since $\pi_w(\{0\}) = 0$, this bound holds for π-almost every (w,b), which suffices for the parameter integral below.

3. Integrate the bound and reach a contradiction. We have previously shown that the target G has the same non-zero response c to every test function Φ_ε, whereas an individual ridge contributes only when its weight lies in E_ε. We now show that the response of the mixture F tends to zero as this angular neighbourhood narrows. To use the individual-ridge bound we found for $|A_\varepsilon(w,b)|$, we first need to express the response of F as the integral of those individual responses.

Fix $\varepsilon > 0$. Substituting the definition of F into its Fourier pairing gives

$$\begin{aligned}\langle \widehat{F}, \Phi_\varepsilon \rangle &= \int_{\mathbb{R}^d} F(x)\widehat{\Phi}_\varepsilon(x)\,dx \\ &= \int_{\mathbb{R}^d} \left(\int_\Theta \psi(w,b)\sigma(w^\top x + b)\,d\pi(w,b) \right) \widehat{\Phi}_\varepsilon(x)\,dx.\end{aligned}$$

In this expression we integrate over the parameters first, and then over x. We would like to reverse that order, integrating over x first to produce exactly $A_\varepsilon(w,b)$. Fubini's theorem permits this reversal provided the absolute value of the full integrand is integrable over both variables. And so, we must check that the following quantity is finite:

$$I_\varepsilon := \int_\Theta \int_{\mathbb{R}^d} |\psi(w,b)\sigma(w^\top x + b)\widehat{\Phi}_\varepsilon(x)|\,dx\,d\pi(w,b).$$

The useful observation here is that polynomial growth separates this integrand into a parameter-dependent factor and an x-dependent factor. Indeed,

$$\begin{aligned}|\sigma(w^\top x + b)| &\le C(1 + |w^\top x + b|)^m \\ &\le C(1 + \|w\|\|x\| + |b|)^m \\ &\le C(1 + \|w\| + |b|)^m(1 + \|x\|)^m.\end{aligned}$$

Consequently, integrating this nonnegative bound gives

$$\begin{aligned}I_\varepsilon \le C &\left(\int_\Theta |\psi(w,b)|(1 + \|w\| + |b|)^m\,d\pi(w,b) \right) \\ &\times \left(\int_{\mathbb{R}^d} (1 + \|x\|)^m |\widehat{\Phi}_\varepsilon(x)|\,dx \right).\end{aligned}$$

This factorisation just uses Tonelli's theorem for nonnegative functions. We have not yet assumed that either factor is finite.

The parameter factor is finite by Cauchy–Schwarz:

$$\begin{aligned}B &:= \int_\Theta |\psi(w,b)|(1 + \|w\| + |b|)^m\,d\pi(w,b) \\ &\le \|\psi\|_{L^2(\pi)} \left(\int_\Theta (1 + \|w\| + |b|)^{2m}\,d\pi(w,b) \right)^{1/2} < \infty.\end{aligned}$$

Here $\psi \in L^2(\pi)$ by the assumed representation, and the remaining moment is finite because Gaussian variables have all polynomial moments, as does a uniform bias on a bounded interval. The spatial factor is finite because Φ_ε is smooth and compactly supported, so its Fourier transform decays faster than any polynomial, by the integration-by-parts argument given earlier. We have therefore proved $I_\varepsilon < \infty$. This is a statement for each fixed ε.

Fubini now gives the desired representation of the mixture's response:

$$\begin{aligned}\langle \widehat{F}, \Phi_\varepsilon \rangle &= \int_\Theta \psi(w,b) \left(\int_{\mathbb{R}^d} \sigma(w^\top x + b)\widehat{\Phi}_\varepsilon(x)\,dx \right) d\pi(w,b) \\ &= \int_\Theta \psi(w,b)A_\varepsilon(w,b)\,d\pi(w,b).\end{aligned}$$

Taking absolute values and applying equation 23, which holds for π-almost every (w,b), yields

$$\begin{aligned}|\langle \widehat{F}, \Phi_\varepsilon \rangle| &\le \int_\Theta |\psi(w,b)|\,|A_\varepsilon(w,b)|\,d\pi(w,b) \\ &\le C_\Phi \int_\Theta |\psi(w,b)|(1 + \|w\| + |b|)^m \mathbf{1}_{E_\varepsilon}(w)\,d\pi(w,b).\end{aligned}$$

This bound has a constant C_Φ independent of ε. Its only dependence on ε is through the indicator of E_ε. We finish by showing that the integral on the right tends to zero.

For each fixed $w \notin L$, the angle $\alpha := \angle(w, L)$ is strictly positive. Once $0 < \varepsilon < \alpha$, that weight is outside E_ε, and so $\mathbf{1}_{E_\varepsilon}(w) = 0$. The only non-zero weights that remain in every E_ε lie on L. These exceptional weights have probability zero because, for $d \geq 2$, a line has zero d-dimensional Lebesgue measure (in orthogonal coordinates it is $\mathbb{R} \times \{0\}^{d-1}$). The non-degenerate Gaussian weight measure has a density p_w, hence

$$\pi_w(L) = \int_L p_w(w)\, dw = 0, \qquad \pi(L \times \mathbb{R}) = \pi_w(L) = 0.$$

It follows that $\mathbf{1}_{E_\varepsilon}(w) \to 0$ for π-almost every (w, b).

To pass this limit through the parameter integral, we use the same integrable parameter factor that appeared in the Fubini calculation:

$$\begin{aligned} 0 &\leq |\psi(w,b)|(1 + \|w\| + |b|)^m \mathbf{1}_{E_\varepsilon}(w) \\ &\leq |\psi(w,b)|(1 + \|w\| + |b|)^m, \qquad \int_\Theta |\psi(w,b)|(1 + \|w\| + |b|)^m \, d\pi = B < \infty. \end{aligned}$$

Thus the integrand tends to zero almost everywhere and is dominated by one integrable function independent of ε. Dominated convergence gives

$$\int_\Theta |\psi(w,b)|(1 + \|w\| + |b|)^m \mathbf{1}_{E_\varepsilon}(w)\, d\pi(w,b) \longrightarrow 0,$$

and the preceding bound implies $\langle \widehat{F}, \Phi_\varepsilon \rangle \to 0$.

Finally, the assumed representer gives $F(x) = G(x)$ for every x, so their Fourier distributions must pair equally with each test function:

$$c = \langle \widehat{G}, \Phi_\varepsilon \rangle = \langle \widehat{F}, \Phi_\varepsilon \rangle \longrightarrow 0.$$

This contradicts the fixed choice $c \neq 0$. No such representer ψ can therefore exist, which proves the non-polynomial direction. □

C Proof of the Square-Root Trace Criterion

Proof overview

Theorem 2 has two parts: a lower bound valid for every candidate covariance, and a construction attaining it when the square-root trace is finite. Appendix A.2 provides the range-space description of RKHS membership used throughout.

Lemma 9 proves the lower bound. If the integrated cost over E is finite, write each neuron as $\sigma_\theta = T^{1/2} g_\theta$. The neuron synthesis operator then factorises as $A_E = T^{1/2} B_E$, where $\|B_E\|_{\mathcal{S}_2}^2 = \int_E \|\sigma_\theta\|_{\mathcal{H}_T}^2 \, d\pi$. Since $A_E A_E^* = T_{\sigma,E}$, its trace norm is $\operatorname{tr}(T_{\sigma,E}^{1/2})$. The Schatten inequality $\|T^{1/2} B_E\|_{\mathcal{S}_1} \leq \|T^{1/2}\|_{\mathcal{S}_2} \|B_E\|_{\mathcal{S}_2}$ gives the claimed bound after squaring. If the integrated cost is infinite, the bound is immediate.

For the converse, the theorem's proof chooses $T = c T_{\sigma,E}^{1/2}$, with $c > 0$. It first checks that almost every neuron in E has zero component in $\ker T_{\sigma,E}$. Summing over the positive eigenvalues then gives both finite integrated cost and

$$\operatorname{tr}(T) A_E(T) = \left[\operatorname{tr}(T_{\sigma,E}^{1/2})\right]^2.$$

The zero-covariance case is handled separately. Thus finiteness of the square-root trace is both necessary and sufficient, and the lower bound is sharp. Corollary 3 simply takes $E = \Theta$.

C.1 Restricted Square-Root Covariance Inequality

We work in $L^2(\nu)$ with the neurons σ_θ and parameter distribution π from the theorem statement. The neuron family is measurable and satisfies

$$\int_\Theta \|\sigma_\theta\|_{L^2(\nu)}^2 \, \pi(d\theta) < \infty.$$

For a measurable $E \subseteq \Theta$, define the unnormalised restricted covariance

$$T_{\sigma,E} := \int_E \sigma_\theta \otimes \sigma_\theta \, \pi(d\theta), \qquad T_{\sigma,E} h = \int_E \langle \sigma_\theta, h \rangle_{L^2(\nu)} \sigma_\theta \, \pi(d\theta).$$

Thus $T_{\sigma,\Theta} = T_\sigma$. When $\pi(E) > 0$, the conditional neuron covariance is $T_{\sigma,E}/\pi(E)$. We use the range-space conventions of Appendix A.2. We use the notation $\mathrm{tr}(T_{\sigma,E} T^\dagger)$ to express the neuron RKHS cost integrated over E. Taking $E = \Theta$ gives the full-distribution cost $A(T)$. This notation requires care, as it is an extended nonnegative cost, which may be infinite, rather than an instruction to multiply two operators and take an ordinary trace. To make its meaning explicit, let P_0 be the orthogonal projection onto $\ker T$, the directions in which the candidate covariance T assigns zero variance. We define

$$\mathrm{tr}(T_{\sigma,E} T^\dagger) := \begin{cases} \displaystyle\sum_j \frac{\langle T_{\sigma,E} e_j, e_j \rangle_{L^2(\nu)}}{t_j}, & \mathrm{tr}(P_0 T_{\sigma,E} P_0) = 0, \\ +\infty, & \mathrm{tr}(P_0 T_{\sigma,E} P_0) > 0. \end{cases}$$

The sum is over the positive eigenvalues of T, and a divergent sum has value $+\infty$. Each numerator measures the neuron energy in direction e_j, while t_j is the variance assigned to that direction by the candidate covariance. Thus each ratio measures how costly that neuron component is under the candidate RKHS norm. No shared eigenbasis for $T_{\sigma,E}$ and T is assumed.

The restricted neuron covariance gives

$$\langle T_{\sigma,E} e_j, e_j \rangle_{L^2(\nu)} = \int_E |\langle \sigma_\theta, e_j \rangle_{L^2(\nu)}|^2 \, \pi(d\theta), \qquad \mathrm{tr}(P_0 T_{\sigma,E} P_0) = \int_E \|P_0 \sigma_\theta\|^2_{L^2(\nu)} \, \pi(d\theta).$$

Consequently, there are two ways the average RKHS cost can be infinite. First, the neurons may have a non-zero component in a zero-variance direction of T on a set of positive measure. Those neurons cannot belong to $\mathcal{H}_T$, so their cost is infinite. Second, even if no such component exists almost everywhere, the sum of the ratios above may diverge because T assigns too little variance across its positive-eigenvalue directions. Interchanging the nonnegative sum and integral therefore gives, with both possibilities included,

$$\mathrm{tr}(T_{\sigma,E} T^\dagger) = \int_E \|\sigma_\theta\|^2_{\mathcal{H}_T} \, \pi(d\theta).$$

This is why the zero-variance case must be specified separately: the Moore–Penrose inverse itself is zero on $\ker T$, so using it mechanically would discard precisely those components that make RKHS membership impossible.

The next inequality is an application of the standard Schatten Hölder inequality (Simon, 2005). We give the argument for restricted neuron covariances, including the possibility of infinite RKHS cost.

Lemma 9 (Restricted square-root covariance inequality)**.** *Let $T \neq 0$ be a positive trace-class operator on $L^2(\nu)$. For every measurable $E \subseteq \Theta$,*

$$\mathrm{tr}(T) \int_E \|\sigma_\theta\|^2_{\mathcal{H}_T} \, \pi(d\theta) \geq \big(\mathrm{tr}(T^{1/2}_{\sigma,E})\big)^2. \tag{24}$$

Equivalently, when $\pi(E) > 0$,

$$\mathrm{tr}(T) \mathbb{E}\big[\|\sigma_\theta\|^2_{\mathcal{H}_T} \mid E\big] \geq \big(\mathrm{tr}((T_{\sigma,E}/\pi(E))^{1/2})\big)^2. \tag{25}$$

Both inequalities are understood in the extended sense, allowing the value $+\infty$.

Proof of Lemma 9. If $I_E := \int_E \|\sigma_\theta\|^2_{\mathcal{H}_T} \, d\pi$ is infinite, the inequality is immediate, since from $T \neq 0$ we have that $\mathrm{tr}(T) > 0$. Now, suppose $I_E < \infty$. Then $\sigma_\theta \in \mathrm{Ran}(T^{1/2})$ for almost every $\theta \in E$, so

$$g_\theta := T^{\dagger/2} \sigma_\theta, \qquad \sigma_\theta = T^{1/2} g_\theta, \qquad \int_E \|g_\theta\|^2_{L^2(\nu)} \, \pi(d\theta) = I_E.$$

Define the synthesis operators

$$B_E f := \int_E f(\theta) g_\theta \, \pi(d\theta), \qquad A_E f := \int_E f(\theta) \sigma_\theta \, \pi(d\theta),$$

from $L^2(E,\pi)$ into $L^2(\nu)$. It is easy to show that they satisfy $A_E = T^{1/2}B_E$ and $A_E A_E^* = T_{\sigma,E}$.

Here an operator is Hilbert–Schmidt when its squared singular values sum to a finite number, and trace class when its singular values sum to a finite number. These two norms are also the Schatten class norms, denoted $\|\cdot\|_{\mathcal{S}_2}$ and $\|\cdot\|_{\mathcal{S}_1}$. Application of Parseval's identity followed by Tonelli's theorem to B_E gives

$$\|B_E\|_{\mathcal{S}_2}^2 = \int_E \|g_\theta\|_{L^2(\nu)}^2 \, d\pi = I_E. \tag{26}$$

The non-zero eigenvalues of $A_E A_E^*$ are the squared singular values of A_E, so $\operatorname{tr}(T_{\sigma,E}^{1/2}) = \|A_E\|_{\mathcal{S}_1}$. The Schatten Hölder inequality (Hytönen et al., 2016) states that

$$\|UV\|_{\mathcal{S}_r} \le \|U\|_{\mathcal{S}_p}\|V\|_{\mathcal{S}_q}, \qquad \frac{1}{r} = \frac{1}{p} + \frac{1}{q},$$

for operators in the relevant Schatten classes, with $p, q, r \in [1,\infty]$. Here $T^{1/2}$ and B_E belong to $\mathcal{S}_2$, with squared norms $\operatorname{tr}(T)$ and I_E respectively. Applying the Schatten Hölder inequality with $p = q = 2$ and $r = 1$ to $A_E = T^{1/2}B_E$ therefore gives

$$\begin{aligned}\operatorname{tr}(T_{\sigma,E}^{1/2}) &= \|A_E\|_{\mathcal{S}_1} \\ &\le \|T^{1/2}\|_{\mathcal{S}_2}\|B_E\|_{\mathcal{S}_2} = \sqrt{\operatorname{tr}(T) I_E}.\end{aligned}$$

Squaring both sides gives equation 24, as desired. When $\pi(E) > 0$, we can divide by $\pi(E)$ and use $\operatorname{tr}((T_{\sigma,E}/\pi(E))^{1/2}) = \pi(E)^{-1/2}\operatorname{tr}(T_{\sigma,E}^{1/2})$ to obtain equation 25. □

C.2 Proof of Theorem 2

Proof of Theorem 2. The inequality *(SR-E)* is shown by Lemma 9, and shows the necessity of $\operatorname{tr}(T_{\sigma,E}^{1/2}) < \infty$ for the existence of a non-zero positive trace-class T with $A_E(T) < \infty$. If $T = 0$, its RKHS contains only zero, so finite integrated cost forces $\sigma_\theta = 0$ almost everywhere on E, and hence $T_{\sigma,E} = 0$. It remains to show sufficiency and attainment of equality, which we do by explicitly constructing the candidate $T = cT_{\sigma,E}^{1/2}$.

Sufficiency: constructing a covariance with finite cost. Suppose $\operatorname{tr}(T_{\sigma,E}^{1/2}) < \infty$ and $T_{\sigma,E} \neq 0$. We choose $T = cT_{\sigma,E}^{1/2}$ with $c > 0$. This operator is positive and trace class, with

$$\operatorname{tr}(T) = c\operatorname{tr}(T_{\sigma,E}^{1/2}) < \infty.$$

It remains to show that its integrated neuron RKHS cost $A_E(T)$ is finite. Before computing this cost from the positive eigenvalues, we must check that there is no neuron component in $\ker T$ for almost every $\theta \in E$. This is vital as such a component would make the extended RKHS cost infinite regardless of the positive-eigenvalue coefficients. Let P_0 denote the orthogonal projection onto $\ker T_{\sigma,E}$. The restricted neuron covariance records the integrated squared neuron coefficient in each direction. Concretely, for every $u \in L^2(\nu)$,

$$\langle T_{\sigma,E}u, u\rangle_{L^2(\nu)} = \int_E |\langle \sigma_\theta, u\rangle_{L^2(\nu)}|^2 \, \pi(d\theta).$$

Let (u_k) be an orthonormal basis of $\ker T_{\sigma,E}$. Application of Parseval's identity and Tonelli's theorem gives

$$\begin{aligned}\int_E \|P_0\sigma_\theta\|_{L^2(\nu)}^2 \, \pi(d\theta) &= \sum_k \int_E |\langle \sigma_\theta, u_k\rangle_{L^2(\nu)}|^2 \, \pi(d\theta) \\ &= \sum_k \langle T_{\sigma,E}u_k, u_k\rangle_{L^2(\nu)} = 0,\end{aligned}$$

because $T_{\sigma,E}u_k = 0$ for every basis vector in its kernel. A nonnegative function with integral zero is zero almost everywhere. Thus $P_0\sigma_\theta = 0$ for almost every $\theta \in E$. Since $\ker T = \ker T_{\sigma,E}$, these neurons have no component in a direction where the candidate covariance assigns zero variance.

Their entire extended RKHS cost is therefore given by the weighted sum over positive-eigenvalue directions.

Let (λ_j, e_j) be the positive eigenpairs of $T_{\sigma,E}$. The covariance identity above gives

$$\int_E |\langle \sigma_\theta, e_j \rangle|^2 \, \pi(d\theta) = \lambda_j.$$

Using the eigenvalues $c\sqrt{\lambda_j}$ of T and another application of Tonelli's theorem, we obtain

$$\begin{aligned} A_E(T) &= \int_E \sum_j \frac{|\langle \sigma_\theta, e_j \rangle|^2}{c\sqrt{\lambda_j}} \, \pi(d\theta) \\ &= \sum_j \frac{\lambda_j}{c\sqrt{\lambda_j}} = c^{-1} \operatorname{tr}(T_{\sigma,E}^{1/2}) < \infty. \end{aligned}$$

We have consequently constructed a covariance with both finite trace and finite integrated neuron RKHS cost. This proves sufficiency.

For completeness, if $T_{\sigma,E} = 0$, then

$$\int_E \|\sigma_\theta\|^2_{L^2(\nu)} \, \pi(d\theta) = \operatorname{tr}(T_{\sigma,E}) = 0,$$

so the neurons are zero as $L^2(\nu)$ elements almost everywhere on E. The zero element belongs to every $\mathcal{H}_T$ and has norm zero. Hence $A_E(T) = 0$ for every positive trace-class T, including $T = 0$, which proves the existence claim in this case as well.

Equality in the lower bound. When $T_{\sigma,E} \neq 0$, this choice attains the lower bound exactly:

$$\operatorname{tr}(T) A_E(T) = \left(c \sum_j \sqrt{\lambda_j} \right) \left(c^{-1} \sum_j \sqrt{\lambda_j} \right) = \left[\operatorname{tr}(T_{\sigma,E}^{1/2}) \right]^2.$$

Thus changing $c > 0$ trades off prior variance against integrated RKHS cost without changing their product. □

C.3 Proof of Corollary 3

Proof of Corollary 3. Take $E = \Theta$ in Theorem 2. Then $T_{\sigma,E} = T_\sigma$ and $A_E(T) = A(T)$, giving the lower bound, existence criterion and equality case. □

D Finite-Derivative-Jump Obstruction

Proof overview

Theorem 4 follows once we show that every positive-probability parameter set E has $\operatorname{tr}(T_{\sigma,E}^{1/2}) = \infty$. Indeed, if a candidate RKHS contained a positive-probability set of neurons, some subset E_M would have uniformly bounded relative cost. The square-root inequality equation SR-E would then give

$$M\pi(E_M) \geq \left[\operatorname{tr}(T_{\sigma,E_M}^{1/2}) \right]^2,$$

contradicting that divergence. The technical lemmas establish the divergence as follows.

Lemma 10 gives a test for failure of trace-class membership: uniformly bounded operators cannot have unbounded trace pairings with a trace-class operator. Lemma 11 shows that sufficiently smooth local kernel terms are trace class, so they contribute only bounded terms to these pairings. Lemma 12 constructs the required tests for a local ridge operator. The first derivative jump, of order s, produces a non-zero leading trace proportional to

$$\int_1^R t^{(d-2s-3)/2} \, dt.$$

This diverges exactly when $d \geq 2s + 1$. Higher-order jumps contribute lower-order terms, and the smooth remainder has bounded trace response, so neither cancels the leading divergence.

Lemma 13 transfers this local test to an arbitrary measurable E of positive probability. It localises the neuron synthesis operator near a jump hyperplane, groups scales sharing that hyperplane, and obtains the operator form required by Lemma 12, with a non-vanishing leading weight. All these operations are bounded, so non-trace-class behaviour of the resulting operator rules out trace-class membership of the original synthesis operator A_E. Finally, $\operatorname{tr}(T_{\sigma,E}^{1/2}) = \|A_E\|_{\mathcal{S}_1}$ gives the required divergence. The last subsection applies the bounded-cost-set contradiction above and treats $T = 0$ separately.

D.1 Trace-Class Tools

We prove this local obstruction by constructing bounded test operators whose trace pairing with the neuron synthesis operator diverges. We begin with the following two lemmas.

The first is the following standard pairing bound between trace-class and bounded operators (Simon, 2005), restated as a non-trace-class test for use below.

Lemma 10 (Bounded trace test)**.** *Let H_1, H_2 be Hilbert spaces, let $A : H_1 \to H_2$ be trace class, and let $B : H_1 \to H_2$ be bounded. Then*

$$|\operatorname{tr}(B^*A)| \leq \|B\| \, \|A\|_{\mathcal{S}_1}.$$

Consequently, a bounded operator $A : H_1 \to H_2$ is not trace class if there is a family of bounded operators $B_R : H_1 \to H_2$ such that each B_R^*A is trace class and

$$\sup_R \|B_R\| < \infty, \qquad |\operatorname{tr}(B_R^*A)| \longrightarrow \infty.$$

Proof. Since A is trace class and B^* is bounded, the Schatten Hölder inequality (with $p = \infty, q = r = 1$) shows that B^*A is trace class and

$$\|B^*A\|_{\mathcal{S}_1} \leq \|B^*\| \, \|A\|_{\mathcal{S}_1} = \|B\| \, \|A\|_{\mathcal{S}_1}.$$

For every trace-class operator C, one has $|\operatorname{tr}(C)| \leq \|C\|_{\mathcal{S}_1}$. Hence

$$|\operatorname{tr}(B^*A)| \leq \|B^*A\|_{\mathcal{S}_1} \leq \|B\| \, \|A\|_{\mathcal{S}_1}.$$

If A were trace class and the B_R were uniformly bounded, this inequality would make the traces $\operatorname{tr}(B_R^*A)$ uniformly bounded as well, proving the final statement by contradiction. □

The second lemma will allow us to discard sufficiently smooth pieces of the localised neuron kernel, which is crucial for isolating the effects of the jump discontinuity.

The following is a version of the standard Sobolev regularity criterion for trace-class integral operators. For example, see Delgado & Ruzhansky (2014, Corollary 4.1). We include a proof allowing an arbitrary parameter measure and requiring regularity only in the output variable x.

Lemma 11 (Smooth kernels define trace-class operators)**.** *Let (Θ, π) be a measure space and let $Q \subset \mathbb{R}^d$ be a bounded cube. Let $K : Q \times \Theta \to \mathbb{C}$ be measurable, with its x-support contained in a compact subset of the interior of Q. Suppose that for some integer $N > d/2$, the weak derivatives in x satisfy*

$$\sum_{|\alpha| \leq N} \int_\Theta \int_Q |\partial_x^\alpha K(x, \theta)|^2 \, dx \, d\pi(\theta) < \infty.$$

Then the integral operator

$$(Af)(x) := \int_\Theta K(x, \theta) f(\theta) \, d\pi(\theta), \qquad A : L^2(\pi) \to L^2(Q),$$

is trace class.

Proof. The term $\alpha = 0$ makes K square-integrable, so A is a well-defined Hilbert–Schmidt operator, by a standard argument. Since K vanishes in a neighbourhood of the boundary of Q, we can safely periodically extend it outside Q without introducing any new discontinuities. This periodic extension therefore retains its weak derivatives $\partial_x^\alpha K$ for all $|\alpha| \le N$. Expanding the x-dependence in the orthonormal Fourier basis $(e_\ell)_{\ell\in\mathbb{Z}^d}$ of $L^2(Q)$ gives

$$K(x,\theta) = \sum_{\ell\in\mathbb{Z}^d} e_\ell(x) k_\ell(\theta), \qquad k_\ell(\theta) := \int_Q \overline{e_\ell(x)} K(x,\theta)\, dx.$$

The expansion converges in $L^2(Q\times\Theta)$. Using the fact that $\partial_x^\alpha e_\ell$ multiplies the ℓ-th Fourier coefficient by a factor of order $|\ell|^{|\alpha|}$, Parseval's identity for the derivatives, followed by integration over Θ, gives

$$\sum_{\ell\in\mathbb{Z}^d} (1+|\ell|^2)^N \|k_\ell\|^2_{L^2(\pi)} < \infty,$$

with harmless constants depending on N, d, and the side length of Q.

For a finite set $F \subset \mathbb{Z}^d$, define the Fourier partial sum

$$K_F(x,\theta) := \sum_{\ell\in F} e_\ell(x) k_\ell(\theta).$$

Since the Fourier expansion converges to K in $L^2(Q\times\Theta)$,

$$\|K_F - K\|_{L^2(Q\times\Theta)} \longrightarrow 0 \qquad \text{as } F \uparrow \mathbb{Z}^d.$$

Now, let $A_F : L^2(\pi) \to L^2(Q)$ be the integral operator with kernel K_F:

$$(A_F f)(x) := \int_\Theta K_F(x,\theta) f(\theta)\, d\pi(\theta).$$

The Hilbert–Schmidt norm agrees with the L^2-norm of the integral kernel, so

$$\|A_F - A\|_{\mathcal{S}_2} = \|K_F - K\|_{L^2(Q\times\Theta)} \longrightarrow 0.$$

And so, our A_F converges to A in the Hilbert–Schmidt norm. Expanding the finite sum gives

$$(A_F f)(x) = \sum_{\ell\in F} e_\ell(x) \int_\Theta k_\ell(\theta) f(\theta)\, d\pi(\theta).$$

Thus $A_F = \sum_{\ell\in F} T_\ell$, where

$$T_\ell f := e_\ell \int_\Theta k_\ell(\theta) f(\theta)\, d\pi(\theta).$$

Each T_ℓ has rank at most one, since every output is a scalar multiple of e_ℓ, so A_F has finite rank. For a rank-one operator $f \mapsto \langle f, v\rangle u$, there is only one singular value, giving a trace norm of $\|u\|\,\|v\|$. With the inner product linear in its first argument, the functional in T_ℓ is $f \mapsto \langle f, \overline{k_\ell}\rangle_{L^2(\pi)}$. Since $\|e_\ell\|_{L^2(Q)} = 1$, we therefore have

$$\|T_\ell\|_{\mathcal{S}_1} = \|e_\ell\|_{L^2(Q)}\, \|\overline{k_\ell}\|_{L^2(\pi)} = \|k_\ell\|_{L^2(\pi)}.$$

All that remains for us to show is that these trace norms are summable. Multiplying and dividing by a Sobolev weight and applying Cauchy–Schwarz yields

$$\begin{aligned}\sum_{\ell\in\mathbb{Z}^d} \|k_\ell\|_{L^2(\pi)} &= \sum_{\ell\in\mathbb{Z}^d} (1+|\ell|^2)^{-N/2} (1+|\ell|^2)^{N/2} \|k_\ell\|_{L^2(\pi)} \\ &\le \left(\sum_{\ell\in\mathbb{Z}^d} (1+|\ell|^2)^{-N}\right)^{1/2} \left(\sum_{\ell\in\mathbb{Z}^d} (1+|\ell|^2)^N \|k_\ell\|^2_{L^2(\pi)}\right)^{1/2} < \infty.\end{aligned}$$

The second factor is finite by the derivative estimate derived earlier. The first factor is finite because $2N > d$. Indeed, group the Fourier indices into unit-width shells

$$S_r := \{\ell \in \mathbb{Z}^d : r \le |\ell| < r+1\}, \qquad r = 1, 2, \dots.$$

Each shell contains $O(r^{d-1})$ lattice points. To see this, place a unit cube centred at each lattice point in S_r. These cubes have disjoint interiors and lie in an annulus of bounded width and volume $O(r^{d-1})$. For $\ell \in S_r$, we also have

$$(1 + |\ell|^2)^{-N} = O(r^{-2N}).$$

Hence, for a constant C independent of r,

$$\sum_{\ell \in \mathbb{Z}^d} (1 + |\ell|^2)^{-N} \leq C \left(1 + \sum_{r \geq 1} r^{d-1-2N} \right) < \infty.$$

The final series converges because $d - 1 - 2N < -1$, equivalently $2N > d$, as we demanded in the lemma statement. Together with the preceding Cauchy–Schwarz estimate, this proves

$$\sum_{\ell \in \mathbb{Z}^d} \|T_\ell\|_{\mathcal{S}_1} = \sum_{\ell \in \mathbb{Z}^d} \|k_\ell\|_{L^2(\pi)} < \infty.$$

From this, we can show that along any increasing finite exhaustion of $\mathbb{Z}^d$, the partial sums $A_F = \sum_{\ell \in F} T_\ell$ form a Cauchy sequence in trace norm. Taking $F \subseteq G$,

$$\|A_G - A_F\|_{\mathcal{S}_1} \leq \sum_{\ell \in G \setminus F} \|T_\ell\|_{\mathcal{S}_1} \leq \sum_{\ell \notin F} \|T_\ell\|_{\mathcal{S}_1} \longrightarrow 0.$$

By completeness of the space of trace-class operators $\mathcal{S}_1$, these partial sums converge in norm to some trace-class operator $\widetilde{A}$. On the other hand, $K_F \to K$ in $L^2(Q \times \Theta)$, so $A_F \to A$ in Hilbert–Schmidt norm, as shown above. Since

$$\|A_F - \widetilde{A}\|_{\mathcal{S}_2} \leq \|A_F - \widetilde{A}\|_{\mathcal{S}_1} \longrightarrow 0,$$

uniqueness of the Hilbert–Schmidt limit gives $\widetilde{A} = A$. Hence A is trace class. □

D.2 Local Ridge Trace Test

We now prove one of the principal lemmas of this paper, which shows that a combination of local ridge functions whose first derivative jump occurs at order s cannot be trace class if the ambient dimension is at least $2s + 1$.

Lemma 12 (A trace test for local ridge operators)**.** *Let $d \geq 1$ and $0 \leq s \leq N$ be integers. Write $x = (x_1, x') \in \mathbb{R} \times \mathbb{R}^{d-1}$ and $z = (v, a) \in \mathbb{R}^{d-1} \times \mathbb{R}$, and put*

$$\phi(x, z) = x_1 + v^\top x' - a, \qquad h_k(t) = t_+^k / k!,$$

with $h_0 = \mathbf{1}_{\{t>0\}}$. Let $D \subset \mathbb{R}^d$ be bounded and measurable, with positive measure, and take a real $\chi \in C_c^\infty(\mathbb{R}^d)$. Suppose

$$H_\chi(z) := \int_{\mathbb{R}^{d-1}} \chi(a - v^\top y, y)^2 \, dy \geq h_- > 0 \quad (z \in D).$$

For $d = 1$, this means $H_\chi(a) = \chi(a)^2$. Define $S_k : L^2(D) \to L^2(\mathbb{R}^d)$ by

$$(S_k g)(x) = \chi(x) \int_D h_k(\phi(x, z)) g(z) \, dz.$$

Let $b_k \in L^\infty(D)$ be real, with $b_s \geq b_- > 0$ almost everywhere on D. Let $\Delta_k \in \mathbb{R}$ with $\Delta_s \neq 0$, and let $\mathcal{R} : L^2(D) \to L^2(\mathbb{R}^d)$ be trace class. If $d \geq 2s + 1$, then

$$\mathcal{A} := \sum_{k=s}^{N} \Delta_k S_k M_{b_k} + \mathcal{R} \quad \textit{is not trace class},$$

where M_{b_k} denotes multiplication by b_k.

Proof of Lemma 12. Writing $S_k M_{b_k}$ as an integral operator gives its kernel as

$$K_{S_k M_{b_k}}(x, z) = \chi(x) h_k(\phi(x, z)) b_k(z).$$

We see that this kernel is bounded and supported on $\operatorname{supp} \chi \times D$, which has finite measure, rendering $K_{S_k M_{b_k}}$ square-integrable. Hence $S_k M_{b_k}$ is Hilbert–Schmidt. We work in the complexified L^2 spaces to ease the use of Fourier transforms. This is harmless as singular values and thus trace-class membership are preserved under complexification. We work with the convention of

$$\widehat{f}(\xi) = \int e^{-2\pi i x \cdot \xi} f(x)\, dx,$$

which is unitary on L^2.

1. Separate smooth terms from the jump frequencies. In the sense of tempered distributions, $\partial_t^{k+1} h_k = \delta_0$, so its Fourier transform obeys $(2\pi i\tau)^{k+1} \widehat{h}_k(\tau) = 1$. Choose a smooth even frequency cut-off ψ which is zero for $|\tau| \le 1$ and one for $|\tau| \ge 2$. We can then separate h_k into its high-frequency jump part and a smooth low-frequency part:

$$h_k = h_k^{\mathrm{lo}} + h_k^{\mathrm{hi}}, \qquad \widehat{h}_k^{\mathrm{hi}}(\tau) = \psi(\tau)(2\pi i\tau)^{-(k+1)} =: q_k(\tau), \qquad \widehat{h}_k^{\mathrm{lo}}(\tau) = \widehat{h}_k(\tau) - \widehat{h}_k^{\mathrm{hi}}(\tau). \tag{27}$$

For $|\tau| \ge 2$, we have $\widehat{h}_k^{\mathrm{hi}}(\tau) = (2\pi i\tau)^{-(k+1)} = \widehat{h}_k(\tau)$, and so $\widehat{h}_k^{\mathrm{lo}} = 0$ for $|\tau| \ge 2$, that is, $\widehat{h}_k^{\mathrm{lo}}$ is compactly supported in frequency. The inverse Fourier transform of this compactly supported distribution is smooth. Consequently, the low-frequency part of h_k is smooth, and the high-frequency part contains the jump in the k-th derivative (the function itself when $k = 0$).

We now wish to apply Lemma 11 to show that the low-frequency part of the operator is trace class. Taking $\Theta = D$ and π as Lebesgue measure, the kernel of $S_k^{\mathrm{lo}} M_{b_k}$ is

$$K_{S_k^{\mathrm{lo}} M_{b_k}}(x, z) = \chi(x) h_k^{\mathrm{lo}}(\phi(x, z)) b_k(z).$$

Now, taking Q as a cube whose interior contains the support of χ, we see that $K_{S_k^{\mathrm{lo}} M_{b_k}}$ is supported on $Q \times \Theta$. All that remains is to check that the x derivatives of $K_{S_k^{\mathrm{lo}} M_{b_k}}$ are square-integrable. Writing $\beta = (\beta_1, \beta')$, the derivatives of $K_{S_k^{\mathrm{lo}} M_{b_k}}$ all take the general form

$$\partial_x^\alpha K_{S_k^{\mathrm{lo}} M_{b_k}}(x, z) = \sum_{\beta \le \alpha} \binom{\alpha}{\beta} (\partial_x^{\alpha-\beta} \chi)(x) v^{\beta'} (h_k^{\mathrm{lo}})^{(|\beta|)}(\phi(x, z)) b_k(z),$$

Since χ is smooth and compactly supported, its derivatives are bounded. The values of $\phi(x, z)$ for $x \in \operatorname{supp} \chi$ and $z \in D$ lie in a bounded interval. Smoothness of h_k^{lo} therefore bounds its derivatives on this interval. The factors $v^{\beta'}$ are bounded because D is bounded, and b_k is bounded by assumption. Hence the kernel derivatives are bounded on their finite-measure support, and thus square-integrable.

Thus, we may apply Lemma 11, which shows that every $S_k^{\mathrm{lo}} M_{b_k}$ is trace class. We can include these terms in $\mathcal{R}$ and analyse the high-frequency portion of the integral operator S_k^{hi} below. Having shown that $S_k^{\mathrm{lo}} M_{b_k}$ is trace class, it is also therefore Hilbert–Schmidt. From the opening of the proof, we have that $S_k M_{b_k}$ is Hilbert–Schmidt, and since Hilbert–Schmidt operators form a vector space, we also have that $S_k^{\mathrm{hi}} M_{b_k}$ is Hilbert–Schmidt.

2. Construct uniformly bounded Fourier tests. We now construct uniformly bounded operators that probe increasingly high frequencies of the ridge singularity. To motivate the construction, consider the Fourier transform of a single ridge function $h_k(t)$,

$$h_k(t) = \int_{\mathbb{R}} e^{2\pi i t\tau} \widehat{h}_k(\tau)\, d\tau$$

Here $\widehat{h}_k$ is a tempered distribution rather than an integrable function, so the displayed integral is formal. It is nevertheless useful for motivating the construction. Plugging in the ridge function $\phi(x, z)$ gives

$$h_k(\phi(x, z)) = \int_{\mathbb{R}} e^{2\pi i \tau \phi(x,z)} \widehat{h}_k(\tau)\, d\tau = \int_{\mathbb{R}} e^{2\pi i x \cdot (\tau, \tau v)} e^{-2\pi i \tau a} \widehat{h}_k(\tau)\, d\tau$$

And so, one can see that each mode in the Fourier transform of h_k factors into a spatial frequency $\xi = (\tau, \tau v)$ in the x-variables and a frequency τ in the offset a. The spatial frequency ξ is in the normal direction to the hyperplane $\phi(x, z) = 0$, as the ridge function is constant along the hyperplane. This motivates transforming first in a, then identifying the resulting coordinates (v, τ) with spatial frequency through $(\tau, v) \mapsto (\tau, \tau v)$.

Let $\mathcal{F}_a$ denote Fourier transformation in the parameter a alone. Define $\mathcal{U} : L^2(\mathbb{R}^d_z) \to L^2(\mathbb{R}^d_x)$, mapping parameter-space functions to spatial functions by

$$\mathcal{U} := \mathcal{F}_x^{-1} \circ V \circ \mathcal{F}_a \tag{28}$$

where V is the unitary coordinate change

$$(Vg)(\xi_1, \xi') = |\xi_1|^{-(d-1)/2} g(\xi'/\xi_1, \xi_1). \tag{29}$$

The formula for V is defined for $\xi_1 \neq 0$, and its values on the null hyperplane $\xi_1 = 0$ can be chosen arbitrarily. The change of variables $\xi = (\tau, \tau v)$ has Jacobian $|\tau|^{d-1}$, which cancels the squared weight in equation 29, making V unitary. Thus $\mathcal{U}$ is unitary because it is a composition of unitary operators. To give some additional intuition about what the map $\mathcal{U}$ is doing, we can write the formal expression

$$(\mathcal{U}f)(x) = \int\int\int e^{2\pi i \tau(x_1 + v^\top x' - a)} |\tau|^{(d-1)/2} f(v, a)\, dv\, da\, d\tau.$$

We see that $\mathcal{U}$ is a Fourier integral building a data-space function from a parameter-space function by integrating over all hyperplanes with normal $(1, v)$ and offset a, with a frequency τ in the normal direction. The factor $|\tau|^{(d-1)/2}$ in this integral combines the weight $|\tau|^{-(d-1)/2}$ from V with the Jacobian $|\tau|^{d-1}$ of the change of variables. This integral is not well-defined in the Lebesgue sense, so it is easier to work with the definition equation 28 of $\mathcal{U}$ as a composition of unitary operators.

For $R \geq 2$, let $\mathcal{C}_R$ retain only a-frequencies $1 \leq |\tau| \leq R$. As an operator acting on parameter space, this is rigorously defined as $\mathcal{C}_R = \mathcal{F}_a^{-1} M_{\mathbf{1}_{\{1 \leq |\tau| \leq R\}}} \mathcal{F}_a$. Let $\mathcal{P}_s$ multiply those frequencies by

$$p_s(\tau) := (i \operatorname{sgn} \tau)^{-(s+1)}, \qquad |p_s(\tau)| = 1 \quad (\tau \neq 0). \tag{30}$$

We fix $p_s(0) = 1$, although the value at $\tau = 0$ can be chosen arbitrarily, as $\tau = 0$ has measure zero, and so any chosen value defines the same action on L^2 functions. Note that this phase matches the Fourier multiplier of the leading jump at order s. Indeed, for $|\tau| \geq 2$,

$$\widehat{h}_s^{\mathrm{hi}}(\tau) = q_s(\tau) = (2\pi)^{-(s+1)} |\tau|^{-(s+1)} p_s(\tau). \tag{31}$$

Consequently, the factor $\overline{p_s(\tau)} q_s(\tau)$ in the trace calculation equation 39 below is positive for $|\tau| \geq 2$, preventing cancellation between positive and negative frequencies in the leading term. Again, $\mathcal{P}_s$ is rigorously defined as a Fourier multiplier on $L^2(\mathbb{R}^d)$ via $\mathcal{P}_s = \mathcal{F}_a^{-1} M_{p_s} \mathcal{F}_a$. Let E_D extend an $L^2(D)$ function by zero to the ambient parameter space. Finally, define our test probe operators $B_R : L^2(D) \to L^2(\mathbb{R}^d)$ by

$$B_R := M_\chi \mathcal{U} \mathcal{P}_s \mathcal{C}_R E_D. \tag{32}$$

It is easy to compute the norms of these factors. As already stated, $\mathcal{U}$ is unitary, and so $\|\mathcal{U}\| = 1$. For the cut-off $\mathcal{C}_R$, we have

$$\begin{aligned} \|\mathcal{C}_R g\|^2_{L^2(\mathbb{R}^d)} &= \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} |\mathbf{1}_{\{1 \leq |\tau| \leq R\}} (\mathcal{F}_a g)(v, \tau)|^2 \, d\tau \, dv \\ &\leq \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} |(\mathcal{F}_a g)(v, \tau)|^2 \, d\tau \, dv = \|g\|^2_{L^2(\mathbb{R}^d)}. \end{aligned} \tag{33}$$

And so $\|\mathcal{C}_R\| \leq 1$. Similarly, it is easy to see that the phase multiplier $\mathcal{P}_s$ is unitary, and so $\|\mathcal{P}_s\| = 1$. The extension operator is an isometry, since

$$\|E_D g\|^2_{L^2(\mathbb{R}^d)} = \int_{\mathbb{R}^d} |E_D g(v, a)|^2 \, dv \, da = \int_D |g(v, a)|^2 \, dv \, da = \|g\|^2_{L^2(D)},$$

So indeed $\|E_D\| = 1$. Thus, we have that every factor except M_χ has norm at most one, and, as $\|M_\chi\| = \|\chi\|_\infty$, we have

$$\|B_R\| \leq \|\chi\|_\infty, \tag{34}$$

uniformly in R. Thus B_R probes an expanding band of spatial frequencies normal to the ridge hyperplanes without its operator norm growing with R.

We can now compute the kernel of B_R explicitly by expanding the operators in the definition equation 32 of B_R, and taking the inverse Fourier transform. First, we have

$$\mathcal{U}\mathcal{P}_s\mathcal{C}_R E_D f = \mathcal{F}_x^{-1} V M_{p_s} M_{\mathbf{1}_{\{1\le|\tau|\le R\}}} \mathcal{F}_a E_D f$$

Since $f \in L^2(D)$ and E_D extends by zero, we have that $E_D f \in L^2(\mathbb{R}^d)$. Moreover, since $E_D f$ is supported on a finite-measure set, we also have that $E_D f \in L^1(\mathbb{R}^d)$, allowing us, by Fubini's theorem, to take its Fourier transform in a as

$$(\mathcal{F}_a E_D f)(v,\tau) = \int_{\mathbb{R}} e^{-2\pi i a\tau} (E_D f)(v,a)\, da$$

The partial Fourier transform preserves the bounded support in v, but need not have bounded support in τ. Multiplication by the cut-off $\mathbf{1}_{1\le|\tau|\le R}$ restricts τ to a bounded set, so the resulting function has bounded support in (v,τ). Multiplication by $p_s(\tau)$ does not alter this support. Applying V from equation 29 then maps the support $(\tau, v) \mapsto \xi = (\tau, \tau v)$ and introduces the Jacobian factor $|\tau|^{-(d-1)/2}$. The $\mathcal{C}_R$ cut-off guarantees that $|\tau| \ge 1$ so there is no singularity at zero. So, the application of V maps compact support on (v,τ) to compact support on ξ. From the earlier norm calculations, including equation 33, we have that

$$\|M_{p_s} M_{\mathbf{1}_{\{1\le|\tau|\le R\}}} \mathcal{F}_a E_D f\|_{L^2} \le \|\mathcal{F}_a E_D f\|_{L^2}$$

And so $\tilde{g}(v,\tau) := p_s(\tau)\mathbf{1}_{\{1\le|\tau|\le R\}}(\mathcal{F}_a E_D f)(v,\tau) \in L^2(\mathbb{R}^d)$. Furthermore, since V is unitary, $V\tilde{g} \in L^2(\mathbb{R}^d)$, and since V maps this bounded support to bounded support in ξ, we have that $V\tilde{g} \in L^1(\mathbb{R}^d)$. This enables us to take the explicit inverse Fourier transform with respect to x, as in the definition equation 28 of $\mathcal{U}$. Multiplying by χ and changing variables give the action of B_R on $f \in L^2(D)$ as

$$(B_R f)(x) = \chi(x) \int_D \int_{1\le|\tau|\le R} e^{2\pi i\tau\phi(x,z)} |\tau|^{(d-1)/2} p_s(\tau) f(z) d\tau dz$$

This gives the kernel explicitly as

$$K_{B_R}(x,z) = \chi(x) \int_{1\le|\tau|\le R} e^{2\pi i\tau\phi(x,z)} |\tau|^{(d-1)/2} p_s(\tau)\, d\tau, \qquad z \in D. \tag{35}$$

For each finite R, this kernel is bounded and supported in the finite-measure product $\operatorname{supp}\chi \times D$, so B_R is also Hilbert–Schmidt.

3. Compute the trace response of a jump term. Aiming for an application of Lemma 10, we now evaluate $\operatorname{tr}(B_R^* S_k^{\mathrm{hi}} M_{b_k})$ to see how strongly the probe equation 32 from part 2 detects a jump of order k. The main point is that the probe and the jump term use the same ridge phase $\phi(x,z)$. When we pair their kernels, their frequencies subtract, so frequencies that are close together give the main contribution. We will show that, to leading order, this pairing recovers the jump multiplier $q_k(\tau)$ from equation 27 times the geometric weight $H_\chi(z)$.

Recall that for Hilbert–Schmidt integral operators $B, S : L^2(D) \to L^2(\mathbb{R}^d)$

$$\operatorname{tr}(B^* S) = \int_D \int_{\mathbb{R}^d} \overline{K_B(x,z)} K_S(x,z)\, dx\, dz. \tag{36}$$

In our case,

$$K_{S_k^{\mathrm{hi}} M_{b_k}}(x,z) = \chi(x) b_k(z) h_k^{\mathrm{hi}}(\phi(x,z)). \tag{37}$$

To see how the Fourier modes combine, recall that K_{B_R} in equation 35 is built from modes $e^{2\pi i\tau\phi(x,z)}$, weighted by $|\tau|^{(d-1)/2} p_s(\tau)$. Similarly, expanding $h_k^{\mathrm{hi}}(\phi(x,z))$ gives

$$h_k^{\mathrm{hi}}(\phi(x,z)) = \int_{\mathbb{R}} e^{2\pi i\eta\phi(x,z)} q_k(\eta)\, d\eta$$

The trace identity equation 36 multiplies the second kernel by the complex conjugate of the first, which combines the oscillatory factors as

$$\overline{e^{2\pi i\tau\phi(x,z)}} e^{2\pi i\eta\phi(x,z)} = e^{-2\pi i\tau\phi(x,z)} e^{2\pi i\eta\phi(x,z)} = e^{2\pi i(\eta-\tau)\phi(x,z)}.$$

In particular, matching frequencies $\eta = \tau$ cancel the oscillation entirely. Including the Fourier weights, the product is therefore

$$\chi(x)^2 b_k(z)|\tau|^{(d-1)/2}\overline{p_s(\tau)}q_k(\eta)e^{2\pi i(\eta-\tau)\phi(x,z)}.$$

Thus the integral in x depends only on the frequency difference $h = \eta - \tau$. Define

$$G_z(h) := \int_{\mathbb{R}^d} \chi(x)^2 e^{2\pi i h\phi(x,z)}\, dx. \tag{38}$$

Substituting the kernels equation 35 and equation 37 into the trace identity equation 36 and then setting $\eta = \tau + h$ gives

$$\operatorname{tr}(B_R^* S_k^{\mathrm{hi}} M_{b_k}) = \int_D b_k(z) \int_{1\le|\tau|\le R} |\tau|^{(d-1)/2}\overline{p_s(\tau)} \left[\int_{\mathbb{R}} G_z(h) q_k(\tau+h)\, dh\right] d\tau\, dz. \tag{39}$$

In equation 39 we used the integral Fourier representation of h_k^{hi}, and this needs some additional care. As in part 2 of the proof, the expression

$$h_k^{\mathrm{hi}}(t) = \int_{\mathbb{R}} e^{2\pi i\eta t} q_k(\eta)\, d\eta$$

is not automatically a valid Lebesgue integral, as although $q_k \in L^2(\mathbb{R})$, it need not be in $L^1(\mathbb{R})$ (in particular, $|q_0(\eta)|$ decays only as $|\eta|^{-1}$). We therefore use the standard procedure of truncating the frequency integral, and then passing to the limit in a specified norm. Define the truncated inverse Fourier transform by

$$h_{k,T}^{\mathrm{hi}}(t) := \int_{-T}^{T} e^{2\pi i\eta t} q_k(\eta)\, d\eta. \tag{40}$$

For finite T the integral in equation 40 is absolutely convergent. With R also fixed, the spatial and parameter cut-offs and the bounded frequency ranges make the integrals in the trace calculation absolutely convergent, so Fubini's theorem allows us to interchange them. Moreover, Plancherel's theorem gives

$$\|h_{k,T}^{\mathrm{hi}} - h_k^{\mathrm{hi}}\|^2_{L^2(\mathbb{R})} = \int_{|\eta|>T} |q_k(\eta)|^2\, d\eta \longrightarrow 0. \tag{41}$$

The convergence in equation 41 is preserved when forming the kernels $\chi(x)b_k(z)h_{k,T}^{\mathrm{hi}}(\phi(x,z))$, as the substitution $t = x_1 + v^\top x' - a$ has Jacobian one, x' ranges over a fixed bounded set on the support of χ, and b_k is bounded on the finite-measure set D. Consequently, the squared $L^2(\mathbb{R}^d \times D)$ difference of the kernels is bounded by a constant times the squared $L^2(\mathbb{R})$ difference in equation 41. Let $S_{k,T}^{\mathrm{hi}}$ denote the operator obtained from S_k^{hi} by replacing h_k^{hi} with $h_{k,T}^{\mathrm{hi}}$. Pairing with the fixed Hilbert–Schmidt kernel of B_R then preserves the limit by Cauchy–Schwarz, justifying convergence of the traces

$$\begin{aligned}|\operatorname{tr}(B_R^* S_{k,T}^{\mathrm{hi}} M_{b_k}) - \operatorname{tr}(B_R^* S_k^{\mathrm{hi}} M_{b_k})| &= \left|\left\langle K_{B_R}, K_{S_{k,T}^{\mathrm{hi}} M_{b_k}} - K_{S_k^{\mathrm{hi}} M_{b_k}}\right\rangle_{L^2(\mathbb{R}^d\times D)}\right| \\ &\le \|K_{B_R}\|_{L^2(\mathbb{R}^d\times D)} \|K_{S_{k,T}^{\mathrm{hi}} M_{b_k}} - K_{S_k^{\mathrm{hi}} M_{b_k}}\|_{L^2(\mathbb{R}^d\times D)} \to 0\end{aligned}$$

On the frequency side, the rapid decay equation 43 of G_z established next, together with boundedness of q_k, permits dominated convergence as $T \to \infty$. This removes the cut-off and yields the trace formula equation 39, with an absolutely convergent inner integral.

To understand G_z from equation 38, use the same change of variables ($t = x_1 + v^\top x' - a$, $y = x'$) to define

$$F_z(t) := \int_{\mathbb{R}^{d-1}} \chi(t + a - v^\top y, y)^2\, dy, \qquad G_z(h) = \int_{\mathbb{R}} F_z(t) e^{2\pi i h t}\, dt.$$

Here $F_z(t)$ measures the weight of the data cut-off on the parallel slice $\phi(x,z) = t$. In particular, $F_z(0) = H_\chi(z)$ is the weight on the jump hyperplane itself. Since F_z is smooth and compactly supported, one-dimensional Fourier inversion at $t = 0$ gives

$$\int_{\mathbb{R}} G_z(h)\, dh = F_z(0) = H_\chi(z). \tag{42}$$

Moreover, $G_z(h)$ is small when the frequency difference $|h|$ is large. Indeed, since $\partial_{x_1}\phi = 1$, integration by parts m times gives, for $h \neq 0$,

$$G_z(h) = \frac{(-1)^m}{(2\pi i h)^m} \int_{\mathbb{R}^d} \partial_{x_1}^m(\chi^2)(x) e^{2\pi i h \phi(x,z)}\, dx.$$

There are no boundary terms because χ has compact support. Taking absolute values, we can show that $|G_z(h)| \leq C'_m |h|^{-m}$ for $h \neq 0$, where $C'_m = (2\pi)^{-m} \|\partial_{x_1}^m(\chi^2)\|_{L^1}$. For small $|h|$, we use the simple bound $|G_z(h)| \leq \|\chi\|_{L^2}^2$. Both these bounds are independent of z and so for every integer $m \geq 0$,

$$|G_z(h)| \leq C_m (1 + |h|)^{-m}, \tag{43}$$

with C_m independent of $z \in D$, and so $G_z(h)$ decays faster than every inverse power of $|h|$. In particular, $\int |h|\, |G_z(h)|\, dh$ is uniformly bounded.

We can now approximate the inner integral in the trace formula equation 39. For ease of notation, define

$$I_k(\tau, z) := \int_{\mathbb{R}} G_z(h) q_k(\tau + h)\, dh$$

Since $G_z(h)$ decays rapidly as $|h| \to \infty$, any large contribution to the integral, for large $|\tau|$, will occur for $|h| \ll |\tau|$. And so we approximate $q_k(\tau + h) \approx q_k(\tau)$ in the integral, giving the approximation

$$I_k(\tau, z) \approx q_k(\tau) \int G_z(h)\, dh = q_k(\tau) H_\chi(z)$$

To quantify the error in this approximation, write

$$\begin{aligned} E_k(\tau, z) &:= \int_{\mathbb{R}} G_z(h) q_k(\tau + h)\, dh - q_k(\tau) H_\chi(z) \\ &= \int_{\mathbb{R}} G_z(h) \big[q_k(\tau + h) - q_k(\tau)\big]\, dh. \end{aligned} \tag{44}$$

Split the integral in equation 44 into nearby and distant frequencies. For $|\tau| \geq 4$ and $|h| \leq |\tau|/2$, every point between τ and $\tau + h$ has absolute value at least $|\tau|/2 \geq 2$. On this interval $q_k(u) = (2\pi i u)^{-(k+1)}$, so its derivative is bounded by $C|\tau|^{-k-2}$. The mean value theorem therefore gives

$$\left| \int_{|h| \leq |\tau|/2} G_z(h) \big[q_k(\tau + h) - q_k(\tau)\big]\, dh \right| \leq C|\tau|^{-k-2} \int_{\mathbb{R}} |h|\, |G_z(h)|\, dh \leq C'|\tau|^{-k-2}. \tag{45}$$

On $|h| > |\tau|/2$, the argument $\tau + h$ can be close to zero, so we instead use the global boundedness of q_k, $\|q_k\|_\infty < \infty$, supplied by its low-frequency cut-off. The decay estimate equation 43 for G_z gives

$$\left| \int_{|h| > |\tau|/2} G_z(h) \big[q_k(\tau + h) - q_k(\tau)\big]\, dh \right| \leq 2\|q_k\|_\infty C_m \int_{|h| > |\tau|/2} (1 + |h|)^{-m}\, dh = O(|\tau|^{1-m}). \tag{46}$$

We are free to take $m \geq k + 3$, bounding this contribution by the same $|\tau|$ order as the first. Combining equation 45 and equation 46, uniformly over $z \in D$,

$$\int_{\mathbb{R}} G_z(h) q_k(\tau + h)\, dh = q_k(\tau) H_\chi(z) + O(|\tau|^{-k-2}). \tag{47}$$

Thus, we have the leading response at each large frequency as $|\tau|^{(d-1)/2} \overline{p_s(\tau)} q_k(\tau) H_\chi(z)$. Its error is bounded by $C|\tau|^{(d-1)/2-k-2}$, one inverse power smaller than the $|\tau|^{(d-1)/2-k-1}$ scale of the leading response.

4. The leading jump gives a divergent trace. Put

$$\gamma_s = s + \frac{d+1}{2}, \qquad L_R := \int_1^R t^{d-\gamma_s - 1}\, dt.$$

By a prescient choice of phase $p_s(\tau)$, for $k = s$, the phases in equation 31 cancel: $\overline{p_s(\tau)}q_s(\tau) = (2\pi)^{-(s+1)}|\tau|^{-s-1}$ for $|\tau| \geq 2$. The trace formula equation 39 and approximation equation 47 therefore give

$$\operatorname{tr}(B_R^* S_s^{\mathrm{hi}} M_{b_s}) = c_s L_R + o(L_R), \tag{48}$$

where

$$c_s = 2(2\pi)^{-(s+1)} \int_D b_s(z) H_\chi(z)\, dz > 0. \tag{49}$$

The factor two accounts for the two signs of τ. The error integrates one lower power of $|\tau|$ and is thus $o(L_R)$. For $k > s$, we bound the absolute value of the trace using equation 39. Since $|p_s(\tau)| = 1$ and $|q_k(\tau)| = (2\pi)^{-(k+1)}|\tau|^{-k-1}$ for $|\tau| \geq 2$, equation 47 gives, for $|\tau| \geq 4$,

$$\left| \int_{\mathbb{R}} G_z(h) q_k(\tau + h)\, dh \right| \leq C_k \big(|\tau|^{-k-1} + |\tau|^{-k-2} \big) \leq C_k' |\tau|^{-k-1},$$

uniformly over $z \in D$. Here H_χ is uniformly bounded because χ is bounded and compactly supported. Multiplying by the weight $|\tau|^{(d-1)/2}$ and integrating over z therefore gives an upper bound proportional to $|\tau|^{(d-1)/2-k-1}$, since $\int_D |b_k(z)|\, dz \leq |D| \|b_k\|_\infty$. The bounded frequency range $1 \leq |\tau| \leq 4$ contributes only a constant independent of R, and the two signs of τ are absorbed into the constant as well. Thus

$$\begin{aligned} |\operatorname{tr}(B_R^* S_k^{\mathrm{hi}} M_{b_k})| &\leq C_k \left(1 + \int_1^R t^{(d-1)/2-k-1}\, dt \right) \\ &= C_k \left(1 + \int_1^R t^{d-\gamma_s-1-(k-s)}\, dt \right) = o(L_R). \end{aligned} \tag{50}$$

So, compared with the integrand defining L_R, a jump of order $k > s$ supplies the extra decay factor $t^{-(k-s)}$. To justify the little-o estimates in equation 48 and equation 50, suppose $L_R \to \infty$ and let $\delta > 0$ denote the extra decay exponent, that is, $\delta = 1$ for the error in the leading jump response, and $\delta = k - s$ for a higher-order jump. We need to show that

$$\frac{1 + \int_1^R t^{d-\gamma_s-1} t^{-\delta}\, dt}{L_R} \longrightarrow 0.$$

Fix any $\varepsilon > 0$ and choose a cut-off $T \geq 1$ large enough that $T^{-\delta} \leq \varepsilon$. For $R \geq T$, split the integral at T. On $[1, T]$ the integral is a fixed finite number, independent of R, whereas on $[T, R]$ we have $t^{-\delta} \leq \varepsilon$, so

$$0 \leq \frac{1 + \int_1^R t^{d-\gamma_s-1-\delta}\, dt}{L_R} \leq \frac{1 + \int_1^T t^{d-\gamma_s-1-\delta}\, dt}{L_R} + \varepsilon \frac{\int_T^R t^{d-\gamma_s-1}\, dt}{L_R} \leq \frac{1 + \int_1^T t^{d-\gamma_s-1-\delta}\, dt}{L_R} + \varepsilon. \tag{51}$$

Keeping T fixed and letting $R \to \infty$, the first term on the right of equation 51 vanishes because $L_R \to \infty$. The limit superior of the ratio is therefore at most ε, but since ε was arbitrary, the ratio must tend to zero. Thus the lower-order contribution can itself diverge, but its ratio to L_R still vanishes.

The dimensional assumption $d \geq 2s + 1$ is exactly what makes L_R diverge:

$$d - \gamma_s = \frac{d - 2s - 1}{2} \geq 0, \qquad L_R = \begin{cases} \log R, & d = 2s + 1, \\ (R^{d-\gamma_s} - 1)/(d - \gamma_s), & d > 2s + 1. \end{cases}$$

We now wire all the pieces we have developed together. The key object from the lemma statement is $\mathcal{A} = \sum_{k=s}^N \Delta_k S_k M_{b_k} + \mathcal{R}$. In part 1, we separated S_k into high and low frequencies, showing that the low-frequency part was trace class. Therefore, we can rewrite $\mathcal{A}$ as $\mathcal{A} = \sum_{k=s}^N \Delta_k S_k^{\mathrm{hi}} M_{b_k} + \widetilde{\mathcal{R}}$ where $\widetilde{\mathcal{R}} := \mathcal{R} + \sum_{k=s}^N \Delta_k S_k^{\mathrm{lo}} M_{b_k}$ is still trace class.

We then plug our $\mathcal{A}$ into the bounded trace test from Lemma 10.

$$\operatorname{tr}(B_R^* \mathcal{A}) = \sum_{k=s}^N \Delta_k \operatorname{tr}(B_R^* S_k^{\mathrm{hi}} M_{b_k}) + \operatorname{tr}(B_R^* \widetilde{\mathcal{R}})$$

We know that B_R is Hilbert–Schmidt and $S_k^{\mathrm{hi}} M_{b_k}$ is Hilbert–Schmidt, so $B_R^* S_k^{\mathrm{hi}} M_{b_k}$ is trace class. Moreover, since $\widetilde{\mathcal{R}}$ is trace class and B_R is bounded, we have that $B_R^* \widetilde{\mathcal{R}}$ is trace class too, so all the traces above genuinely exist. Importantly, the trace-class remainder $\widetilde{\mathcal{R}}$ has a uniformly bounded trace against B_R. This follows from applying Lemma 10 with $A = \widetilde{\mathcal{R}}$ and $B = B_R$, and recalling from the construction of B_R that $\|B_R\| \le \|\chi\|_\infty$. Overall this gives $|\operatorname{tr}(B_R^* \widetilde{\mathcal{R}})| \le \|\chi\|_\infty \|\widetilde{\mathcal{R}}\|_{\mathcal{S}_1}$. In particular, since $L_R \to \infty$, $\operatorname{tr}(B_R^* \widetilde{\mathcal{R}}) = O(1) = o(L_R)$.

Combining equation 48 with our higher-order jump estimates we see

$$\operatorname{tr}(B_R^* \mathcal{A}) = \Delta_s c_s L_R + o(L_R).$$

If $\mathcal{A}$ were trace class, then Lemma 10 and equation 34 would imply that $|\operatorname{tr}(B_R^* \mathcal{A})|$ is bounded uniformly in R by $\|\chi\|_\infty \|\mathcal{A}\|_{\mathcal{S}_1}$. However, $|\operatorname{tr}(B_R^* \mathcal{A})|$ diverges as $\Delta_s c_s \neq 0$ and $L_R \to \infty$, and so $\mathcal{A}$ cannot be trace class. This proves the lemma.

□

D.3 Local Finite-Jump Spectral Obstruction

Lemma 13 (Local finite-jump spectral obstruction)**.** *Under the assumptions of Theorem 4, define, for measurable $E \subseteq \Theta$,*

$$T_{\sigma,E} := \int_E \sigma_\theta \otimes \sigma_\theta \, \pi(d\theta).$$

If $\pi(E) > 0$ and $d \ge 2s + 1$, then

$$\operatorname{tr}(T_{\sigma,E}^{1/2}) = \infty. \tag{52}$$

Proof of Lemma 13. We first localise the features near a common jump, then apply Lemma 12. The main task is to put the feature operator into the form required by that lemma: its parameter $z = (v, a)$ describes a hyperplane, whereas our original parameter (w, b) also specifies a scale. We will localise in both input and parameter space, then combine scales with the same hyperplane. These operations suffice because, for a trace-class operator A and bounded operators B, C,

$$\|BAC\|_{\mathcal{S}_1} \le \|B\| \, \|A\|_{\mathcal{S}_1} \, \|C\|. \tag{53}$$

Thus, if the resulting operator is not trace class, neither was the original one. No openness or regularity of the measurable set E will be needed.

The principal task of the proof here is to massage the actual ridge synthesis operator until we can apply Lemma 12.

1. Write the covariance as a synthesis operator. In a similar vein to the proof of Lemma 9, we begin by defining, or recalling, the synthesis operator A_E, as

$$(A_E f)(x) := \int_E f(\theta) \sigma_\theta(x) \, \pi(d\theta).$$

We wish to check that this integral defines a Hilbert–Schmidt operator. Since the activation function σ has at most polynomial growth, we can choose an integer $m \ge 0$ and $C < \infty$ such that $|\sigma(t)| \le C(1 + |t|^m)$. Since $|w^\top x + b| \le \|w\| \, \|x\| + |b|$, we have, after increasing the constant,

$$|\sigma_\theta(x)|^2 \le C'(1 + \|w\|^{2m} \|x\|^{2m} + |b|^{2m}).$$

Both π and ν are non-degenerate Gaussian measures and so have finite moments of every order. Consequently, this bound is integrable against $\nu(dx)\pi(d\theta)$. In particular,

$$\|A_E\|_{\mathcal{S}_2}^2 = \int_E \|\sigma_\theta\|_{L^2(\nu)}^2 \, \pi(d\theta) < \infty.$$

Now that $A_E \in \mathcal{S}_2$ is established, as in the proof of Lemma 9, it is easy to show that $A_E A_E^* = T_{\sigma,E}$ and $\operatorname{tr}(T_{\sigma,E}^{1/2}) = \|A_E\|_{\mathcal{S}_1}$. Thus the required divergence of $\operatorname{tr}(T_{\sigma,E}^{1/2})$ is equivalent to A_E not being trace class.

2. Use coordinates describing the jump hyperplane. The set $w = 0$ has probability zero, and the $2d$ sets $\{w_j > 0\}$ and $\{w_j < 0\}$ cover its complement. At least one of their intersections with E therefore has positive measure, and so we restrict to that intersection. The corresponding synthesis operator is $A_E I$, where I extends a function on the smaller set by zero and is an isometry, so proving non-trace-class membership after this restriction is sufficient. A simultaneous permutation or reflection of the x and w coordinates lets us call the chosen set $\{w_1 > 0\}$, without loss of generality. These changes preserve $w^\top x$, induce only unitary changes of the corresponding L^2 spaces, and leave both Gaussian laws non-degenerate. We use E, ν, π for the restricted set and the laws in these coordinates.

We now introduce coordinates that separate the scale from the hyperplane:

$$r = w_1 > 0, \quad v = w'/w_1, \quad a = (t_0 - b)/w_1, \qquad z = (v, a), \quad \phi(x, z) = x_1 + v^\top x' - a. \tag{54}$$

By equation 54, $w^\top x + b = t_0 + r\phi(x, z)$, so z specifies the hyperplane on which the pre-activation $w^\top x + b$ equals t_0, and r controls the scale normal to it. The inverse coordinate map is

$$(r, v, a) \longmapsto (w, b) = (r, rv, t_0 - ra).$$

Its absolute Jacobian determinant is r^d, as after the first coordinate r is fixed, the $d-1$ coordinates v each contribute a factor r, and a contributes a factor r as well. If p_π is the Gaussian density in these coordinates, the transformed density is consequently

$$p(r, z) = p_\pi(r, rv, t_0 - ra)r^d.$$

This is a smooth and strictly positive transformation for $r > 0$. Under this change of variables, integration with respect to the original parameter measure π becomes integration with respect to $p(r, z)\, dr\, dz$. Thus, we may work equivalently in the (r, z) coordinates with L^2 space $L^2(p(r, z)\, dr\, dz)$. For notational ease, we continue to denote the image of E under this coordinate change by E, and we will denote the measure after the coordinate change by $d\tilde{\pi}(r, z) = p(r, z)\, dr\, dz$. Thus, we now work with our synthesis operator acting on $L^2(E, \tilde{\pi})$ as

$$(A_E f)(x) = \int_E f(r, z)\sigma(t_0 + r\phi(x, z))d\tilde{\pi}(r, z)$$

Since p is strictly positive, E has positive Lebesgue measure as well as positive parameter probability. The Lebesgue density theorem therefore provides a density point (r_0, v_0, a_0) of E, at which the proportion of E in sufficiently small balls tends to one. That is,

$$\lim_{\delta \downarrow 0} \frac{\mathrm{Leb}(E \cap B_\delta(r_0, v_0, a_0))}{\mathrm{Leb}(B_\delta(r_0, v_0, a_0))} = 1,$$

In particular, every sufficiently small neighbourhood of this point has a positive-measure intersection with E. Choose a small bounded neighbourhood P with compact closure in $r > 0$. Choose a bump function $\zeta \in C_c^\infty(P)$, $0 \le \zeta \le 1$, which equals one in a smaller neighbourhood of that point. Thus ζ is non-zero on a positive-measure part of E. Note that we need never assume E contains an open set.

We next choose an input patch that meets all the nearby jump hyperplanes. Set $z_0 = (v_0, a_0)$ and $x_0 = (a_0, 0)$, so that $\phi(x_0, z_0) = 0$, and choose a bump function $\chi \in C_c^\infty(\mathbb{R}^d)$, $0 \le \chi \le 1$, equal to one near x_0. Make its support and then P small enough that

$$|r\phi(x, z)| \le \epsilon' < \epsilon \quad \text{on } \operatorname{supp}\chi \times \operatorname{supp}\zeta, \tag{55}$$

where the one-sided regularity assumption holds on $[t_0 - \epsilon, t_0 + \epsilon]$ (the assumption that the two one-sided restrictions of σ around t_0 admit C^N extensions to t_0). This is possible because $r_0\phi(x_0, z_0) = 0$ and $(x, r, z) \mapsto r\phi(x, z)$ is continuous. Whenever P is shrunk, choose ζ inside the smaller neighbourhood, still equal to one near the density point, so that its intersection with E retains positive measure.

We must also check the lower bound on H_χ required by Lemma 12. For $d > 1$, choose $\delta > 0$ sufficiently small so that $\chi = 1$ on the ball $B(x_0, \delta)$. Since P is a bounded patch of (r, v, a)-space, its projection onto $z = (v, a)$-space is also bounded. So there exists some finite V such that $\|v\| \le V$ for all points in P. We then shrink the patch P even further, rechoosing ζ if necessary, so that $|a - a_0| < \delta/2$. Further, choose $\eta > 0$ with $(V + 1)\eta < \delta/2$. Then, for all $\|y\| < \eta$,

$$\|(a - v^\top y, y) - x_0\| \le |a - a_0| + (\|v\| + 1)\|y\| < \delta.$$

Thus $\chi(a - v^\top y, y) = 1$ for every such y and every z in the patch P, and we have our lower bound as

$$H_\chi(z) = \int \chi(a - v^\top y, y)^2 \, dy \geq |B_{\mathbb{R}^{d-1}}(0, \eta)| =: h_- > 0. \tag{56}$$

For $d = 1$, restricting a to the region where $\chi = 1$ gives $H_\chi(a) = \chi(a)^2 = 1$ there.

Currently, our synthesis operator A_E maps between the L^2 spaces $L^2(E, \tilde{\pi})$ and $L^2(\nu)$, whereas Lemma 12 is formulated in terms of operators on Lebesgue L^2 spaces. So, to cleanly apply the lemma, we wish to move to these Lebesgue spaces. Write $n(x)$ for the strictly positive Gaussian density of ν, and let $K = \operatorname{supp} \chi$. Compactness gives $n_- := \inf_K n > 0$. The map

$$L_\chi : L^2(\nu) \to L^2(dx), \qquad L_\chi g = \chi g,$$

is bounded, because

$$\|L_\chi g\|^2_{L^2(dx)} \leq \frac{\|\chi\|^2_\infty}{n_-} \|g\|^2_{L^2(\nu)}.$$

Multiplication by ζ on the parameter space, $M_\zeta : L^2(E, \tilde{\pi}) \to L^2(E, \tilde{\pi})$, is also bounded, with norm at most one, as

$$\|M_\zeta f\|^2_{L^2(E,\tilde{\pi})} \leq \|\zeta\|^2_\infty \|f\|^2_{L^2(E,\tilde{\pi})} = \|f\|^2_{L^2(E,\tilde{\pi})}$$

Define the localised operator $A_0 : L^2(E, \tilde{\pi}) \to L^2(dx)$

$$A_0 := L_\chi A_E M_\zeta, \qquad (A_0 f)(x) = \chi(x) \int_E \zeta(r, z) f(r, z) \sigma(t_0 + r\phi(x, z)) \, \tilde{\pi}(dr, dz). \tag{57}$$

By the ideal property of trace-class operators, the conditions $L_\chi, M_\zeta \in \mathcal{B}$ and $A_E \in \mathcal{S}_1$ imply that $A_0 = L_\chi A_E M_\zeta \in \mathcal{S}_1$, where $\mathcal{B}$ is the space of bounded operators. And so, it suffices to show that A_0 is not trace class.

At this point, we have only fixed the output space of the relevant synthesis operator to be Lebesgue space. To apply Lemma 12, we still need to fix its input space to be Lebesgue, as well as integrate out the scale parameter r, which the lemma does not depend on. Both of these tasks are achieved in the next section of the proof.

3. Combine scales that share the same hyperplane. For fixed z, all scales r have the same jump hyperplane $\phi(x, z) = 0$. We combine these scales using a normalised function of r, leaving an operator whose only parameter is z. To motivate how to choose this function, note that the jump in the s-th x_1-derivative of $\sigma(t_0 + r\phi(x, z))$ is $r^s \Delta_s$, that is

$$\begin{aligned}[\partial^s_{x_1} \sigma(t_0 + r\phi)]_{\phi=0^+} - [\partial^s_{x_1} \sigma(t_0 + r\phi)]_{\phi=0^-} &= r^s(\sigma^{(s)}(t_0+) - \sigma^{(s)}(t_0-)) \\ &= r^s \Delta_s\end{aligned}$$

Including the cut-off, its scale-dependent factor is $\zeta(r, z) r^s$, where $\Delta_s = \sigma^{(s)}(t_0+) - \sigma^{(s)}(t_0-) \neq 0$. This factor is telling us how strongly the (r, z)-ridge function is carrying the singularity we care about. Now, defining the scale fibre for a given hyperplane z as $E_z := \{r : (r, z) \in E\}$, we define the squared norm along this scale fibre by

$$Q(z) := \int_{E_z} \zeta(r, z)^2 r^{2s} p(r, z) \, dr. \tag{58}$$

The function Q is measurable by Tonelli's theorem, which also allows the interchange of integrals in

$$\int Q(z) \, dz = \int_E \zeta(r, z)^2 r^{2s} p(r, z) \, dr \, dz.$$

This integral is positive because ζ equals one on a positive-measure part of E, and r, p are positive there. It is finite because all the factors are bounded on the compact support of ζ. Hence $\{Q > 0\}$ has positive measure, and Q is finite almost everywhere on $\mathbb{R}^d$. Consequently, some bounded measurable set D of positive measure satisfies

$$0 < q_- \leq Q(z) \leq q_+ < \infty \qquad (z \in D). \tag{59}$$

For example, choose a positive-measure set from $\{0 < Q < \infty\} = \bigcup_{j\geq 1}\{1/j \leq Q \leq j\}$, and take $q_- = 1/j$ and $q_+ = j$. This set lies within the projected parameter patch, $\mathrm{proj}_z(P)$, as ζ is supported inside P.

Now, define $J : L^2(D) \to L^2(E, \tilde{\pi})$ by

$$(Jg)(r,z) = \frac{\zeta(r,z)r^s}{\sqrt{Q(z)}} g(z) \quad (z \in D), \qquad (Jg)(r,z) = 0 \quad (z \notin D). \tag{60}$$

The lower bound in equation 59 makes this formula well-defined. Using the definition of $Q(z)$, equation 58, its normalising factor gives

$$\|Jg\|^2_{L^2(E,p\,dr\,dz)} = \int_D \frac{|g(z)|^2}{Q(z)} \left(\int_{E_z} \zeta(r,z)^2 r^{2s} p(r,z)\, dr \right) dz = \|g\|^2_{L^2(D)}.$$

Thus J is an isometry. J is a lifting operation, taking a function over z and lifting it to a function over (r,z), by weighting by the natural leading jump amplitude $\zeta(r,z)r^s$, with an extra factor of $1/\sqrt{Q(z)}$, to ensure that J is an isometry. In particular, composing with J is a bounded operation, and the chosen weighting will give a strictly positive coefficient for the leading jump, as we will see in the next part of the proof.

Considering the operator $A_0 J : L^2(D) \to L^2(dx)$, we now have an operator between two Lebesgue spaces, as desired. Moreover, since J is bounded, it suffices to show that $A_0 J$ is not trace class.

4. Separate the first jump from smoother terms. Choose the integer $N > d/2$ supplied by the one-sided regularity assumption of Theorem 4. Since $d \geq 2s+1$, necessarily $N > s$. For $s \leq k \leq N$, define the finite one-sided jumps

$$\Delta_k := \sigma^{(k)}(t_0+) - \sigma^{(k)}(t_0-).$$

Recall from Lemma 12 that $h_k(t) = t_+^k/k!$, with $h_0(t) = \mathbf{1}_{\{t>0\}}$. Its one-sided j-th derivatives at zero have difference one when $j = k$ and zero otherwise. For $j < k$, the right derivative is $t^{k-j}/(k-j)!$ and tends to zero as $t \to 0$, while for $j > k$ it vanishes on each side. Thus subtracting $\Delta_k h_k$ cancels precisely the jump at order k, without changing the other one-sided derivative jumps. This is easily checked by explicit computation of the jumps for $\tilde{\sigma}(t) = \sigma(t_0+t) - \Delta_k h_k(t)$. The jumps below order s already vanish by assumption. Consequently,

$$\rho_N(t) := \sigma(t_0+t) - \sum_{k=s}^{N} \Delta_k h_k(t) \in C^N(-\epsilon, \epsilon). \tag{61}$$

Indeed, the two restrictions of ρ_N are C^N up to zero, and all their derivatives through order N have matching limits at zero. Therefore, defining $\rho_N(0)$ to be the common order 0 limit yields a C^N function on $(-\epsilon, \epsilon)$, whose derivatives at zero are the common one-sided limits. This redefinition of the value at zero does not change any L^2 ridge feature $\rho_N(\phi(x,z))$, since the value only differs on the hyperplane $\phi(x,z) = 0$ which has Lebesgue measure zero. Because $r > 0$, $h_k(r\phi) = r^k h_k(\phi)$, including for $k = 0$. Combining our integral operator from equation 57 with our isometry J from equation 60 gives

$$(A_0 Jg)(x) = \chi(x) \int_D \frac{g(z)}{\sqrt{Q(z)}} \int_{E_z} \zeta(r,z)^2 r^s \sigma(t_0 + r\phi(x,z)) p(r,z)\, dr\, dz.$$

The localisation bound equation 55 ensures that the argument $r\phi$ lies in $[-\epsilon', \epsilon']$, where equation 61 applies. Substituting the identity $\sigma(t_0+t) = \sum_{k=s}^{N} \Delta_k h_k(t) + \rho_N(t)$ into this expression and collecting each $h_k(\phi)$ therefore gives

$$A_0 J = \sum_{k=s}^{N} \Delta_k S_k M_{b_k} + R_N, \tag{62}$$

with S_k as in Lemma 12 and

$$b_k(z) := \frac{1}{\sqrt{Q(z)}} \int_{E_z} \zeta(r,z)^2 r^{s+k} p(r,z)\, dr. \tag{63}$$

Comparing equation 63 with equation 58 gives $b_s = \sqrt{Q} \geq \sqrt{q_-}$ on D. More explicitly, let r_+ be an upper bound for r on $\operatorname{supp}\zeta$. For $k \geq s$, we have $r^{s+k} = r^{2s}r^{k-s} \leq r_+^{k-s}r^{2s}$ there, so

$$0 \leq b_k(z) \leq r_+^{k-s}\frac{Q(z)}{\sqrt{Q(z)}} \leq r_+^{k-s}\sqrt{q_+}. \tag{64}$$

These coefficients are measurable by the same parameter-integration argument as for Q. Thus all the b_k are real and essentially bounded, and the leading one is bounded away from zero.

The final thing to check before applying Lemma 12 is that the remainder R_N is trace class. It is convenient first to consider the corresponding operator on the whole transformed parameter space. Let

$$\widetilde{\Theta} := (0,\infty)\times\mathbb{R}^d, \qquad \widetilde{\pi}(dr,dz) := p(r,z)\,dr\,dz,$$

and define

$$(\widetilde{R}_N f)(x) := \int_{\widetilde{\Theta}} K_N(x;r,z) f(r,z)\,\widetilde{\pi}(dr,dz),$$

where

$$K_N(x;r,z) := \chi(x)\zeta(r,z)\rho_N(r\phi(x,z)).$$

Since ρ_N is currently defined only on $(-\epsilon,\epsilon)$, choose $\eta \in C_c^\infty((-\epsilon,\epsilon))$ with $\eta = 1$ on $[-\epsilon',\epsilon']$, replace ρ_N by $\eta\rho_N$ on $(-\epsilon,\epsilon)$, and extend the result by zero to the whole line. This gives a $C^N(\mathbb{R})$ function. By the bound in equation 55, $r\phi(x,z) \in [-\epsilon',\epsilon']$ whenever $(x,r,z) \in \operatorname{supp}\chi \times \operatorname{supp}\zeta$, so this extension does not change the kernel K_N.

We now verify the hypotheses of Lemma 11. For a multi-index α with $|\alpha| \leq N$, the product and chain rules give, writing $\beta = (\beta_1,\beta')$,

$$\partial_x^\alpha K_N(x;r,z) = \zeta(r,z)\sum_{\beta\leq\alpha}\binom{\alpha}{\beta}(\partial_x^{\alpha-\beta}\chi)(x) r^{|\beta|}v^{\beta'}\rho_N^{(|\beta|)}(r\phi(x,z)). \tag{65}$$

Every term in equation 65 is uniformly bounded on its support: the derivatives of χ are bounded, r and v range over the bounded set $\operatorname{supp}\zeta$, and the derivatives of ρ_N through order N are bounded on the relevant compact interval. Moreover, these derivatives vanish outside $\operatorname{supp}\chi \times \operatorname{supp}\zeta$.

Choose a bounded cube $Q_x \subset \mathbb{R}^d$ whose interior contains $\operatorname{supp}\chi$. Since Q_x has finite Lebesgue measure and $\operatorname{supp}\zeta$ has finite $\widetilde{\pi}$-measure, it follows that

$$\sum_{|\alpha|\leq N}\int_{\widetilde{\Theta}}\int_{Q_x}\left|\partial_x^\alpha K_N(x;r,z)\right|^2\,dx\,\widetilde{\pi}(dr,dz) < \infty.$$

Since $N > d/2$, Lemma 11 therefore shows that $\widetilde{R}_N$, regarded as an operator into $L^2(Q_x)$, is trace class. Extending its output by zero outside Q_x is an isometry, so $\widetilde{R}_N$ is also trace class as an operator into $L^2(\mathbb{R}^d,dx)$.

It remains only to recover the remainder appearing in equation 62. Let

$$I_E : L^2(E,\widetilde{\pi}) \longrightarrow L^2(\widetilde{\Theta},\widetilde{\pi})$$

denote extension by zero outside E. As mentioned, this is an isometry. The remainder operator before composition with J is $\widetilde{R}_N I_E$, and hence the remainder in equation 62 is

$$R_N = \widetilde{R}_N I_E J.$$

Since I_E and J are bounded and trace-class operators form an operator ideal, R_N is trace class.

5. Apply the trace test. The decomposition equation 62 has exactly the form required by Lemma 12. The set D is bounded with positive measure. Equation 59 gives $b_s = \sqrt{Q} \geq \sqrt{q_-} > 0$. Equation 56 supplies the lower bound on H_χ (with the stated one-dimensional interpretation). The remaining coefficients b_k are bounded by equation 64, and so $b_k \in L^\infty(D)$, as needed. R_N is trace class by part 4, and $\Delta_s \neq 0$ by assumption. Hence Lemma 12 shows that $A_0 J$ is not trace class when $d \geq 2s+1$. If A_E were trace class, then $A_0 J = L_\chi A_E M_\zeta J$ would be trace class by the bounded-composition inequality equation 53, a contradiction. The earlier restriction and coordinate changes were also bounded or unitary operations, so this rules out trace-class membership for the original synthesis operator on the original set E. Finally, the identity $\operatorname{tr}(T_{\sigma,E}^{1/2}) = \|A_E\|_{\mathcal{S}_1}$ gives $\operatorname{tr}(T_{\sigma,E}^{1/2}) = \infty$, as required. □

D.4 Proof of Theorem 4

Proof of Theorem 4. If $T = 0$, then $\mathcal{H}_T = \{0\}$, whereas σ_θ is non-zero in $L^2(\nu)$ whenever $w \neq 0$, as its one-dimensional profile has a non-zero derivative jump and the Gaussian input has full support. Since $\pi\{w = 0\} = 0$, the conclusion follows in this case.

Fix now a non-zero positive trace-class T and let

$$C_T(\theta) := \operatorname{tr}(T)\|\sigma_\theta\|^2_{\mathcal{H}_T}.$$

If the set $B := \{\theta : \sigma_\theta \in \mathcal{H}_T\}$ had positive measure, then so would some bounded-cost subset, because

$$B = \bigcup_{M=1}^{\infty} E_M, \qquad E_M := \{\theta : C_T(\theta) \leq M\}.$$

Since $\pi(B) > 0$, at least one of these E_M has positive measure. Applying equation SR-E to such an E_M gives

$$M\pi(E_M) \geq \int_{E_M} C_T(\theta)\, \pi(d\theta) \geq \left[\operatorname{tr}(T^{1/2}_{\sigma,E_M})\right]^2.$$

Lemma 13 makes the right-hand side infinite. This contradiction proves $\pi(B) = 0$. □

E High-Dimensional Cost for Smooth Activations

Proof overview

For Theorem 5, fix a proposed polynomial cost scale d^p and choose an integer Hermite degree $n > p$. The Gaussian-chaos decomposition identifies a harmonic subspace $Y_{d,n}$ whose dimension satisfies $m_{d,n} \sim d^n/n!$. Two lemmas control the neuron's projection onto this growing subspace.

Lemma 14 shows that $c_n(R_d, b)^2$ converges in probability to $c_n(r_*, b)^2 > 0$ almost surely. Gaussian smoothing makes the limiting coefficient analytic in the bias. If it vanished identically, the smoothed activation would be polynomial, and the lemma's Fourier argument would force σ itself to be polynomial, contradicting the assumption. Lemma 15 shows that the projected ridge vectors $v_{d,n}(U_d)$ spread isotropically across $Y_{d,n}$, with squared norm $a_{d,n} \to 1$.

Normalise the candidate covariance to $S_d = T_d/\operatorname{tr}(T_d)$. Its trace is one, so the angular identity bounds its average variance along these projected ridge vectors by $a_{d,n}/m_{d,n}$. RKHS Cauchy–Schwarz says that a neuron with a non-small degree-n coefficient can have low cost only if the candidate assigns a sufficiently large variance to its projected direction. Markov's inequality therefore gives equation 94:

$$\pi_d\{C_d \leq Md^p\} \leq \mathbb{P}\{c_n(R_d, b)^2 \leq \eta\} + \frac{Md^p}{\eta a_{d,n} m_{d,n}}.$$

For fixed $\eta > 0$, the second term vanishes because $n > p$. Taking $d \to \infty$ and then $\eta \downarrow 0$ removes the first term using Lemma 14. This proves the theorem directly for the general radial setting. The final step verifies that Gaussian fan-in weights satisfy that setting's concentration and moment assumptions.

E.1 General Radial Setting and Gaussian-Chaos Decomposition

We prove the conclusions of Theorem 5 under the following more general radial assumptions, and then verify the Gaussian fan-in specialisation. Retain $\nu_d = \mathcal{N}(0, I_d)$ and a smooth, non-polynomial activation σ of at most polynomial growth. Choose $m \geq 0$ and $C < \infty$ such that $|\sigma(t)| \leq C(1 + |t|)^m$. Let

$$\begin{aligned} w &= R_d U_d, \qquad R_d \geq 0, \\ R_d &\longrightarrow r_* > 0 \text{ in probability}, \qquad \sup_{d\geq 1} \mathbb{E}[R_d^{2m}] < \infty. \end{aligned} \tag{66}$$

where U_d is uniform on $\mathbb{S}^{d-1}$ and independent of (R_d, b). The bias b is independent of R_d and has a fixed law, either non-degenerate Gaussian or uniform on a bounded interval of positive length. Both bias laws have the required finite $2m$-th moment.

As in the theorem, $\theta = (w, b)$ has law π_d and $\sigma_\theta(x) = \sigma(w^\top x + b)$. Throughout the proof we use the representation $w = R_d U_d$. We write Θ_d for the parameter space and retain the theorem's definitions of T_d and $C_d(\theta; T_d)$. The moment assumptions ensure that the neuron covariance on $L^2(\nu_d)$ is trace class. As usual, we define the neuron covariance by

$$T_{\sigma,d} := \mathbb{E}_{\theta\sim\pi_d}[\sigma_\theta \otimes \sigma_\theta]. \tag{67}$$

We begin by reviewing the necessary mathematical machinery we need for the proof of Theorem 5. We use the standard Hermite expansion and Wiener (Gaussian) chaos decompositions described in Janson (1997). Let h_n denote the normalised probabilists' Hermite polynomial of degree n, so that $h_n(t) = \mathrm{He}_n(t)/\sqrt{n!}$. For fixed radius r and bias b, expand the one-dimensional neuron profile as

$$\sigma(rt + b) = \sum_{n\geq 0} c_n(r, b) h_n(t) \tag{68}$$

in $L^2(\mathcal{N}(0, 1))$, where

$$c_n(r, b) := \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[\sigma(rZ + b)h_n(Z)]. \tag{69}$$

For a unit vector u and $X \sim \mathcal{N}(0, I_d)$, we have $u^\top X \sim \mathcal{N}(0, 1)$. The corresponding neuron therefore has the expansion

$$\sigma_\theta(x) = \sigma(R_d U_d^\top x + b) = \sum_{n\geq 0} c_n(R_d, b) h_n(U_d^\top x) \tag{70}$$

in $L^2(\nu_d)$. Orthogonality between distinct Hermite degrees gives

$$\|\sigma_\theta\|^2_{L^2(\nu_d)} = \sum_{n\geq 0} c_n(R_d, b)^2, \tag{71}$$

$$\mathbb{E}\|\sigma_\theta\|^2_{L^2(\nu_d)} = \sum_{n\geq 0} q_{d,n}, \qquad q_{d,n} := \mathbb{E}_{R_d,b}[c_n(R_d, b)^2]. \tag{72}$$

Thus $q_{d,n}$ measures the expected neuron covariance energy in the n-th Gaussian chaos.

Under the stated bias law, the limiting coefficient $c_n(r_*, b)$ is non-zero for almost every bias in every fixed degree $n \geq 1$, although it may vanish at particular biases. This fact and the convergence of $c_n(R_d, b)$ are established in Lemma 14.

For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$, write $|\alpha| := \sum_{i=1}^d \alpha_i$, $\alpha! := \prod_{i=1}^d \alpha_i!$, and $u^\alpha := \prod_{i=1}^d u_i^{\alpha_i}$. The corresponding normalised multivariate Hermite polynomial is the product

$$h_\alpha(x) := \prod_{i=1}^d h_{\alpha_i}(x_i).$$

Thus α_i specifies the Hermite degree in coordinate i, and $|\alpha|$ is the total degree. These products form an orthonormal basis of $L^2(\nu_d)$, and the n-th Gaussian chaos is defined as

$$\mathcal{C}_n = \mathrm{span}\{h_\alpha : |\alpha| = n\}. \tag{73}$$

Every $f \in L^2(\nu_d)$ decomposes as $f = \sum_{n\geq 0} f_n$, where $f_n \in \mathcal{C}_n$, with convergence in $L^2(\nu_d)$.

We use the same Hermite generating identity to identify both the ridge features and the harmonic polynomials needed below within this chaos:

$$e^{tz - t^2/2} = \sum_{k\geq 0} \mathrm{He}_k(z) \frac{t^k}{k!}.$$

For a unit vector u, factorising the generating function across coordinates gives $e^{tu^\top x - t^2/2} = \prod_{i=1}^d e^{tu_i x_i - t^2 u_i^2/2}$, since $\sum_i u_i^2 = 1$. Comparing coefficients of t^n yields

$$h_n(u^\top x) = \sum_{|\alpha|=n} \sqrt{\frac{n!}{\alpha!}}\, u^\alpha h_\alpha(x).$$

Every term in this expansion has total Hermite degree n, so $h_n(u^\top x) \in \mathcal{C}_n$. Taking $u = U_d$ and multiplying by the scalar $c_n(R_d, b)$ therefore gives $c_n(R_d, b)h_n(U_d^\top x) \in \mathcal{C}_n$.

The proof will use the projection of these ridge features onto a smaller subspace of $\mathcal{C}_n$: the space $Y_{d,n}$ of homogeneous harmonic polynomials of degree n, with the $L^2(\nu_d)$ inner product. Here homogeneous means $p(tx) = t^n p(x)$, and harmonic means

$$\Delta p = 0, \qquad \Delta := \sum_{i=1}^{d} \partial_{x_i}^2.$$

This is simply a condition on polynomial coefficients. Equivalently, these polynomials correspond to symmetric coefficient tensors with every contraction of two indices equal to zero (traceless). They describe the degree-n dependence on the input direction.

To justify this inclusion, use the same generating identity to express Hermite polynomials as corrected monomials. Comparing coefficients in $e^{-\partial_z^2/2}e^{tz} = e^{tz-t^2/2}$ gives $e^{-\partial_z^2/2}z^k = \mathrm{He}_k(z)$. Since coordinate derivatives commute, its multivariate version is

$$e^{-\Delta/2}x^\alpha = \prod_{i=1}^{d} \mathrm{He}_{\alpha_i}(x_i) = \sqrt{\alpha!}\, h_\alpha(x).$$

Thus applying this operation to any homogeneous polynomial p of degree n produces a linear combination of the same degree-n Hermite basis functions as in the ridge expansion:

$$:p: := e^{-\Delta/2}p = \sum_{j=0}^{\lfloor n/2 \rfloor} \frac{(-1)^j}{2^j j!}\Delta^j p \in \mathcal{C}_n. \tag{74}$$

This differential-operator exponential is only a finite sum here, since each Laplacian lowers the polynomial degree by two. If p is harmonic, $\Delta p = 0$ and all correction terms with $j > 0$ vanish in equation 74. Hence $p = :p: \in \mathcal{C}_n$, which proves $Y_{d,n} \subset \mathcal{C}_n$.

For $d \geq 2$, the dimension is (Dai & Xu, 2013)

$$m_{d,n} := \dim(Y_{d,n}) = \binom{d+n-1}{n} - \binom{d+n-3}{n-2} \sim \frac{d^n}{n!} \tag{75}$$

for fixed n as $d \to \infty$, with the second binomial interpreted as zero for $n < 2$. Writing $\mathcal{P}_{d,n}$ for the space of homogeneous polynomials of degree n, the formula follows by rank–nullity from $Y_{d,n} = \ker(\Delta|_{\mathcal{P}_{d,n}})$ and the surjectivity of $\Delta : \mathcal{P}_{d,n} \to \mathcal{P}_{d,n-2}$ for $n \geq 2$, and for $n < 2$, the Laplacian vanishes identically. The first binomial is asymptotic to $d^n/n!$, while the second is of lower order, giving the stated dimension growth.

So, we have now decomposed a neuron across the Gaussian chaoses $\mathcal{C}_n$ as $\sigma(R_d U_d^\top x + b) = \sum_{n\geq 0} c_n(R_d, b)h_n(U_d^\top x)$, and identified a subspace $Y_{d,n} \subset \mathcal{C}_n$ of dimension asymptotic to $d^n/n!$ for fixed n. The following lemmas show that the degree-n coefficient has an almost surely non-zero limit and that random directions fill $Y_{d,n}$ isotropically. These are the two ingredients needed for the direct cost bound in the theorem's proof.

E.2 Hermite Coefficients and Angular Isotropy

The next lemma combines classical Gaussian smoothing and Hermite identities (Janson, 1997) with the characterisation of point-supported distributions (Dyatlov, 2022, Theorem 4.19).

Lemma 14 (Non-vanishing limiting Hermite coefficients)**.** *Suppose that σ is continuous, non-polynomial and has at most polynomial growth. Let $R_d \geq 0$ converge in probability to $r_* > 0$, and let b be independent of the radii, with a fixed law that is either non-degenerate Gaussian or uniform on a bounded interval of positive length. For every fixed $n \geq 1$,*

$$c_n(r_*, b)^2 > 0 \qquad \textit{almost surely},$$

and, using the same bias b for all d,

$$c_n(R_d, b)^2 \longrightarrow c_n(r_*, b)^2 \qquad \textit{in probability}.$$

Proof of Lemma 14. We first show that the limiting coefficient $c_n(r_*, b)$ is non-zero for almost every bias. We then use continuity of the Hermite coefficients to establish convergence in probability.

1. Express Hermite coefficients through Gaussian smoothing. For $r > 0$, define

$$G_r(b) := \mathbb{E}[\sigma(rZ + b)] = \int_{\mathbb{R}} \sigma(t)\varphi_r(t - b)\, dt, \qquad \varphi_r(t) = \frac{1}{\sqrt{2\pi}r} e^{-t^2/(2r^2)}. \tag{76}$$

This representation lets us differentiate the Gaussian kernel rather than σ, which is only assumed continuous. The identity

$$\partial_b^n \varphi_r(t - b) = r^{-n}\, \mathrm{He}_n\left(\frac{t - b}{r}\right) \varphi_r(t - b)$$

follows by differentiating the Gaussian density. Each derivative is just a polynomial times a Gaussian, and so on compact sets of b and positive r, it remains integrable after multiplication by the polynomial-growth function $\sigma(t)$. Differentiation under the integral is therefore valid. Substituting $t = rZ + b$ gives

$$c_n(r, b) = \frac{r^n}{\sqrt{n!}} G_r^{(n)}(b). \tag{77}$$

For fixed $r > 0$, Gaussian smoothing also makes G_r real analytic. To see this, consider the complex extension

$$\widetilde{G}_r(z) := \frac{1}{\sqrt{2\pi}r} \int_{\mathbb{R}} \sigma(t) e^{-(t-z)^2/(2r^2)}\, dt, \qquad z \in \mathbb{C}.$$

For each fixed t, the integrand is entire in z. On a compact set $K \subset \mathbb{C}$, choose A with $|\operatorname{Re} z|, |\operatorname{Im} z| \le A$ for $z \in K$. Writing $z = a + iv$ and using $(t - a)^2 \ge t^2/2 - a^2$ gives

$$\left|\sigma(t) e^{-(t-z)^2/(2r^2)}\right| \le C e^{A^2/r^2} (1 + |t|)^m e^{-t^2/(4r^2)},$$

an integrable bound independent of $z \in K$. The holomorphic parameter-integral theorem therefore implies that $\widetilde{G}_r$ is entire (Haslinger, 2018, Section 2.7). Its restriction $G_r = \widetilde{G}_r|_{\mathbb{R}}$ is consequently real analytic, as is $c_n(r, \cdot)$ by equation 77.

2. Show that the limiting coefficient cannot vanish identically. If $c_n(r_*, \cdot)$ were identically zero, then equation 77 would give $G_{r_*}^{(n)} = 0$. Integrating this identity n times shows that G_{r_*} is a polynomial of degree at most $n-1$. We show that Gaussian smoothing cannot turn a non-polynomial σ into a polynomial.

As used extensively in our previous proofs, polynomial growth makes σ a tempered distribution, so we can Fourier transform it even when it is not integrable. Using the convention $\widehat{f}(\xi) = \int e^{-2\pi i t\xi} f(t)\, dt$, convolution with the Gaussian, $G_{r_*}(b) = (\sigma * \varphi_{r_*})(b)$, gives

$$\widehat{G}_{r_*}(\xi) = e^{-2\pi^2 r_*^2 \xi^2} \widehat{\sigma}(\xi). \tag{78}$$

The Fourier transform of a polynomial is a finite combination of Dirac delta derivatives at zero, so it vanishes away from zero. To see why the same must then hold for $\widehat{\sigma}$, take any smooth compactly supported test function φ whose support avoids zero. Since the Gaussian multiplier is nowhere zero, $\psi(\xi) = e^{2\pi^2 r_*^2 \xi^2} \varphi(\xi)$ is still a smooth compactly supported test function. Equation 78 gives

$$\langle \widehat{\sigma}, \varphi \rangle = \langle \widehat{G}_{r_*}, \psi \rangle = 0.$$

This is precisely the statement that $\widehat{\sigma}$ is supported at zero. A distribution supported at one point is a finite linear combination of derivatives of the Dirac delta, whose inverse Fourier transform is an ordinary polynomial (Dyatlov, 2022, Theorem 4.19 and Section 11.2). Thus σ agrees almost everywhere with a polynomial. Since σ is continuous, the agreement holds everywhere, contradicting the hypothesis.

Consequently, $c_n(r_*, \cdot)$ is a non-zero real-analytic function on $\mathbb{R}$. Being real analytic forces its zeros to be isolated, as an accumulation point would force all coefficients in its local power series to vanish, and hence the function would vanish identically. An isolated subset of the real line is

countable and therefore has Lebesgue measure zero. Both allowed bias laws have densities with respect to Lebesgue measure, so

$$X_* := c_n(r_*, b)^2 > 0 \qquad \text{almost surely.} \tag{79}$$

3. Establish convergence of the Hermite coefficients. Sample a single bias $b \sim \pi_b$, independently of the radii R_d, as in the lemma conditions. Now, since $\sigma(rz+b)$ is continuous in (r, b), we have that $\sigma(r_k z + b_k)h_n(z) \to \sigma(rz + b)h_n(z)$ for any sequence $(r_k, b_k) \to (r, b)$. The polynomial-growth bound on σ then gives the absolute bound

$$|\sigma(r_k z + b_k)h_n(z)| \le C'(1 + |z|)^m |h_n(z)|$$

This enables the use of dominated convergence in the integral definition of $c_n(r, b)$ to show that $c_n(r_k, b_k) \to c_n(r, b)$, and thus that $c_n(r, b)$ is continuous. Since $(R_d, b) \to (r_*, b)$ in probability, continuity gives

$$X_d := c_n(R_d, b)^2 \longrightarrow X_* \qquad \text{in probability.} \tag{80}$$

□

Let $P_{d,n}$ denote the orthogonal projection onto $Y_{d,n}$, and define

$$v_{d,n}(u) := P_{d,n} h_n(u^\top \cdot), \qquad a_{d,n} := \|v_{d,n}(u)\|^2_{L^2(\nu_d)}.$$

The next lemma computes this projection explicitly. In particular, its squared norm is independent of u and tends to one for each fixed n.

This is the classical isotropy identity for spherical harmonics, expressed in our Gaussian L^2 normalisation, see the addition formula in Bach (2017, Appendix D.1). We compute the normalising constant explicitly for the projected Hermite features.

Lemma 15 (Isotropic projection identity)**.** *Let $d \ge 2$ and $n \ge 1$. If U_d is uniform on $\mathbb{S}^{d-1}$, then*

$$\mathbb{E}_{U_d}[v_{d,n}(U_d) \otimes v_{d,n}(U_d)] = \frac{a_{d,n}}{m_{d,n}} P_{d,n}, \tag{81}$$

where

$$a_{d,n} = \frac{m_{d,n} n!}{d(d+2)\cdots(d+2n-2)} \longrightarrow 1 \quad \textit{for fixed } n \textit{ as } d \to \infty. \tag{82}$$

Consequently,

$$P_{d,n} T_{\sigma,d} P_{d,n} = \frac{q_{d,n} a_{d,n}}{m_{d,n}} P_{d,n}. \tag{83}$$

Proof of Lemma 15. We compute the averaged operator by testing it against arbitrary $p, q \in Y_{d,n}$. This will show directly that every direction in this subspace receives the same covariance, and will identify its value.

1. Compute the inner product with a projected ridge polynomial. Since $P_{d,n}$ is the orthogonal projection onto $Y_{d,n}$, $\langle P_{d,n} f, p\rangle = \langle f, p\rangle$ for $p \in Y_{d,n}$. For fixed $u \in \mathbb{S}^{d-1}$, decompose a standard Gaussian vector as $X = Zu + X_\perp$, where $Z = u^\top X \sim \mathcal{N}(0, 1)$ is independent of $X_\perp$. Conditioning on $X_\perp$, and regarding $p(Zu + X_\perp)$ as an ordinary one-dimensional function of Z, repeated one-dimensional Gaussian integration by parts gives

$$\mathbb{E}_Z[\mathrm{He}_n(Z) p(Zu + X_\perp)] = \mathbb{E}_Z[(u^\top \nabla)^n p(Zu + X_\perp)].$$

The boundary terms vanish because a polynomial times a Gaussian density, and all the derivatives appearing here, decay at infinity. Now averaging also over $X_\perp$ and dividing by $\sqrt{n!}$ yields

$$\begin{aligned} \langle v_{d,n}(u), p\rangle_{L^2(\nu_d)} &= \langle h_n(u^\top \cdot), p\rangle_{L^2(\nu_d)} \\ &= \frac{1}{\sqrt{n!}} \mathbb{E}_{X_\perp} \mathbb{E}_Z[\mathrm{He}_n(Z) p(Zu + X_\perp)] \\ &= \frac{1}{\sqrt{n!}} \mathbb{E}_{X_\perp} \mathbb{E}_Z[(u^\top \nabla)^n p(Zu + X_\perp)] \\ &= \frac{1}{\sqrt{n!}} \mathbb{E}_X[(u^\top \nabla)^n p(X)] = \sqrt{n!}\, p(u). \end{aligned} \tag{84}$$

For the last equality, $t \mapsto p(x+tu)$ is a polynomial of degree at most n whose coefficient of t^n is $p(u)$, by homogeneity. Its n-th derivative is therefore the constant $n!p(u)$.

2. Average the pairing over directions. Write $X = \rho U_d$. Rotational invariance of the standard Gaussian ν_d makes U_d uniform on the sphere and independent of $\rho = \|X\|$. Homogeneity gives $p(X)q(X) = \rho^{2n}p(U_d)q(U_d)$, recalling $p, q \in Y_{d,n}$, and hence

$$\langle p, q\rangle_{L^2(\nu_d)} = \mathbb{E}[\rho^{2n}]\mathbb{E}[p(U_d)q(U_d)].$$

Using equation 84, we obtain

$$\begin{aligned}&\mathbb{E}_{U_d}[\langle v_{d,n}(U_d), p\rangle\langle v_{d,n}(U_d), q\rangle] \\ &\quad = n!\mathbb{E}[p(U_d)q(U_d)] = \frac{n!}{\mathbb{E}[\rho^{2n}]}\langle p, q\rangle_{L^2(\nu_d)}.\end{aligned} \tag{85}$$

To evaluate the denominator, we note ρ^2 has the chi-squared density $t^{d/2-1}e^{-t/2}/(2^{d/2}\Gamma(d/2))$ on $t > 0$. Integrating t^n against this density and using $\Gamma(x+1) = x\Gamma(x)$ gives

$$\mathbb{E}[\rho^{2n}] = 2^n\frac{\Gamma(d/2+n)}{\Gamma(d/2)} = d(d+2)\cdots(d+2n-2). \tag{86}$$

Since equation 85 holds for every p, q in the subspace, the averaged operator is $n!/\mathbb{E}[\rho^{2n}]$ times the identity operator on $Y_{d,n}$. Each $v_{d,n}(u)$ lies in $Y_{d,n}$, so its rank-one operator vanishes on the orthogonal complement. This identifies the averaged operator $\mathbb{E}[v_{d,n}(U_d)\otimes v_{d,n}(U_d)]$ with $(n!/\mathbb{E}[\rho^{2n}])P_{d,n}$ on the whole of $L^2(\nu_d)$.

It remains to relate the constant $n!/\mathbb{E}[\rho^{2n}]$ to the norm $a_{d,n}$. For an orthogonal matrix O, the map $f(x) \mapsto f(O^\top x)$ is unitary on $L^2(\nu_d)$ and preserves homogeneity and harmonicity. It therefore preserves $Y_{d,n}$ and its orthogonal complement, and commutes with $P_{d,n}$. Applying it to $h_n(u^\top x)$ sends $v_{d,n}(u)$ to $v_{d,n}(Ou)$. Any two unit vectors are related by such an O, so the corresponding projections $v_{d,n}(u)$ and $v_{d,n}(Ou)$ have equal norms. Taking the trace of both sides of the operator identity

$$\mathbb{E}[v_{d,n}(U_d)\otimes v_{d,n}(U_d)] = \frac{n!}{\mathbb{E}[\rho^{2n}]}P_{d,n}$$

using an orthonormal basis of the finite-dimensional space $Y_{d,n}$, now gives

$$a_{d,n} = \mathbb{E}_{U_d}\|v_{d,n}(U_d)\|^2 = \frac{m_{d,n}n!}{d(d+2)\cdots(d+2n-2)}.$$

This follows from equation 86 and $\mathrm{tr}(P_{d,n}) = \dim(Y_{d,n}) = m_{d,n}$. This proves both equation 81 and the formula in equation 82. For fixed n, the numerator is asymptotic to d^n by equation 75, while the denominator is $d^n\prod_{j=0}^{n-1}(1+2j/d) \sim d^n$. Thus $a_{d,n} \to 1$. Also $a_{d,n} \le 1$, because orthogonal projection cannot increase the norm and $\|h_n(u^\top\cdot)\|_{L^2(\nu_d)} = 1$.

3. Identify the neuron covariance on this subspace. Since $Y_{d,n} \subset \mathcal{C}_n$ and different chaos degrees are orthogonal, applying $P_{d,n}$ to the Hermite expansion of σ_θ removes all terms except degree n. Boundedness of the projection justifies applying it to the full infinite expansion, so

$$P_{d,n}\sigma_\theta = c_n(R_d, b)v_{d,n}(U_d). \tag{87}$$

For any vector f, projection of its rank-one operator satisfies $P_{d,n}(f\otimes f)P_{d,n} = (P_{d,n}f)\otimes(P_{d,n}f)$. Consequently, independence of U_d and (R_d, b) gives

$$\begin{aligned}P_{d,n}T_{\sigma,d}P_{d,n} &= \mathbb{E}[(P_{d,n}\sigma_\theta)\otimes(P_{d,n}\sigma_\theta)] \\ &= \mathbb{E}[c_n(R_d,b)^2 v_{d,n}(U_d)\otimes v_{d,n}(U_d)] \\ &= q_{d,n}\mathbb{E}_{U_d}[v_{d,n}(U_d)\otimes v_{d,n}(U_d)] = \frac{q_{d,n}a_{d,n}}{m_{d,n}}P_{d,n}\end{aligned}$$

which is equation 83. □

E.3 Proof of Theorem 5

Proof of Theorem 5. We work first in the general radial setting introduced above, with equation 66 in place of Gaussian fan-in weights. Fix $p > 0$ and $M > 0$. Proving the assertion for positive M suffices, since an event with any smaller threshold is contained in one of these events. Choose an integer $n > p$. The angular space $Y_{d,n}$ then has dimension of order d^n, which will grow faster than the proposed cost bound Md^p.

1. Normalise the candidate covariance and bound its average variance. Set

$$S_d := T_d/\operatorname{tr}(T_d), \qquad \operatorname{tr}(S_d) = 1, \qquad C_d(\theta; T_d) = \|\sigma_\theta\|^2_{\mathcal{H}_{S_d}}. \tag{88}$$

The last identity follows from the spectral formula for the RKHS norm. In particular, if $T_d e_j = \lambda_j e_j$ with $\lambda_j > 0$, replacing λ_j by $\lambda_j/\operatorname{tr}(T_d)$ multiplies $\sum_j |\langle f, e_j\rangle|^2/\lambda_j$ by $\operatorname{tr}(T_d)$. The kernel and the range defining RKHS membership are unchanged by positive scalar multiplication, so the identity also holds with infinite cost for non-members. For this proof write $P = P_{d,n}$, $v(u) = v_{d,n}(u)$, $a = a_{d,n}$ and $m = m_{d,n}$.

Define $V_d(u) := \langle S_d v(u), v(u)\rangle \geq 0$, the candidate variance tested against the projected ridge direction $v(u)$. The angular identity equation 81 implies

$$\begin{aligned} \mathbb{E}_{U_d} V_d(U_d) &= \mathbb{E}\operatorname{tr}(S_d[v(U_d)\otimes v(U_d)]) && (89)\\ &= \operatorname{tr}(S_d \mathbb{E}[v(U_d)\otimes v(U_d)]) && (90)\\ &= \frac{a}{m}\operatorname{tr}(PS_dP) \leq \frac{a}{m}. && (91) \end{aligned}$$

To see the equalities explicitly, choose an orthonormal basis $e_1, \ldots, e_m$ of $Y_{d,n}$. The angular identity says $\mathbb{E}[\langle v, e_i\rangle\langle v, e_j\rangle] = (a/m)\mathbf{1}_{\{i=j\}}$. Expanding V_d in this basis therefore leaves $(a/m)\sum_{i=1}^m \langle S_d e_i, e_i\rangle$. Extending this basis to the ambient space and using positivity of S_d bounds this sum by $\operatorname{tr}(S_d) = 1$.

2. Turn a small RKHS cost into a large required variance. Suppose $\sigma_\theta \in \mathcal{H}_{S_d}$. There is a vector $g \perp \ker S_d$ with $\sigma_\theta = S_d^{1/2} g$ and $\|g\|^2 = C_d(\theta; T_d)$. In the eigenbasis of S_d, this is obtained by dividing each feature coefficient by the square root of its positive eigenvalue. RKHS membership is exactly what makes those coefficients square-summable. Cauchy–Schwarz gives

$$\begin{aligned} |\langle \sigma_\theta, v(U_d)\rangle|^2 &= |\langle g, S_d^{1/2} v(U_d)\rangle|^2 \\ &\leq C_d(\theta; T_d) V_d(U_d). \end{aligned} \tag{92}$$

By equation 87, $P\sigma_\theta = c_n(R_d, b) v(U_d)$. Since $v(U_d)$ belongs to the range of P, the left-hand side of equation 92 is $c_n(R_d, b)^2 a^2$. Thus, writing $X_d = c_n(R_d, b)^2$, for any $\eta > 0$ we have the event inclusion

$$\{C_d \leq Md^p\} \subseteq \{X_d \leq \eta\} \cup \left\{ V_d(U_d) \geq \frac{\eta a^2}{Md^p} \right\}. \tag{93}$$

Indeed, a feature with finite cost and $X_d > \eta$ must satisfy the second inequality by equation 92. Features outside the RKHS never belong to the event on the left. This avoids dividing by V_d, which might be zero.

For any nonnegative random variable V and $t > 0$, Markov's inequality tells us $\mathbb{E}V \geq t\mathbb{P}(V \geq t)$. Apply it to the second event in equation 93 and use equation 91 to obtain

$$\pi_d\{C_d \leq Md^p\} \leq \mathbb{P}\{c_n(R_d, b)^2 \leq \eta\} + \frac{Md^p}{\eta a_{d,n} m_{d,n}}. \tag{94}$$

The estimate holds uniformly over the candidate covariance S_d.

3. Take the dimension limit, then the coefficient threshold limit. Equations (80) and (79) give $X_d \to X_* := c_n(r_*, b)^2 > 0$ in probability. For every fixed $\eta > 0$,

$$\mathbb{P}(X_d \leq \eta) \leq \mathbb{P}(X_* \leq 2\eta) + \mathbb{P}(|X_d - X_*| > \eta).$$

The second probability tends to zero. Also $a_{d,n} \to 1$ and $m_{d,n} \sim d^n/n!$, so $Md^p/(\eta a_{d,n} m_{d,n}) \to 0$ because $n > p$. Consequently equation 94 gives

$$\limsup_{d\to\infty} \pi_d\{C_d \le Md^p\} \le \mathbb{P}(X_* \le 2\eta).$$

Now let $\eta \downarrow 0$. The right-hand side tends to $\mathbb{P}(X_* = 0) = 0$, proving the claimed divergence in probability. Keeping η fixed until after the dimension limit is what allows both terms of equation 94 to be controlled.

This proves the theorem's conclusions under the general radial assumptions.

4. Specialise to Gaussian fan-in weights. For the weight law in the theorem, write $w = (\sigma_w/\sqrt{d})G_d$ with $G_d \sim \mathcal{N}(0, I_d)$. Rotational invariance gives a uniform direction independent of the radius $R_d = \sigma_w\|G_d\|/\sqrt{d}$ and of the independent bias. Since $\mathbb{E}[\|G_d\|^2/d] = 1$ and $\mathrm{Var}(\|G_d\|^2/d) = 2/d$, we have $R_d \to \sigma_w > 0$ in probability. The Gaussian radial moment formula equation 86 also gives, for every integer $k \ge 1$,

$$\mathbb{E}[R_d^{2k}] = \sigma_w^{2k} \prod_{j=0}^{k-1} \left(1 + \frac{2j}{d}\right) \le \sigma_w^{2k} \prod_{j=0}^{k-1} (1 + 2j) < \infty.$$

These bounds imply uniform bounds for every fixed nonnegative radial moment, including the moment required by the polynomial-growth exponent. Thus equation 66 holds with $r_* = \sigma_w$, which yields the stated fan-in version of the theorem. □

F Variational Cost of Neural-Network Means

Proof overview

Theorem 8 combines a deterministic ELBO comparison with a probabilistic lower bound on network RKHS cost. Proposition 6 supplies the comparison. Setting the inducing mean to zero while retaining its covariance cancels every covariance-dependent term in the ELBO difference. The Loève isometry identifies the remaining mean KL term with $\frac{1}{2}\|h\|^2_{\mathcal{H}_{T_d}}$. Since the squared residual is nonnegative, a posterior can outperform its centred counterpart only if $v_d\|h\|^2_{\mathcal{H}_{T_d}} \le v_d\|y\|^2/\sigma_d^2$.

Proposition 7 transfers Theorem 5 from individual neurons to random networks. Conditional on hidden neurons in the RKHS, the network cost is a Gaussian quadratic form $a^\top G a$, and $\mathrm{tr}\, G$ is the average individual-neuron cost. The bound equation 98 makes a small quadratic form unlikely when this trace is large, without requiring a well-conditioned Gram matrix. The fraction-of-low-cost-neurons argument in equation 101 controls the trace uniformly in network width. If any hidden neuron is outside the RKHS, the output weights producing a member lie in a proper linear subspace and have Gaussian probability zero. Together these observations give equation 102, whose two terms vanish at every polynomial cost threshold by Theorem 5.

Finally, put $C_d = v_d\|g_d\|^2_{\mathcal{H}_{T_d}}$. For every allowed dataset and every finite-KL inducing posterior with mean g_d, Proposition 6 gives the simultaneous bound

$$v_d[\mathcal{L}(q^\circ) - \mathcal{L}(q)] \ge \tfrac{1}{2}(C_d - B_d).$$

The network cost exceeds every polynomial scale in probability, while B_d has polynomial growth. Their difference is therefore positive with high probability and itself exceeds every polynomial scale. The same event works for all allowed datasets and inducing constructions because it depends only on g_d and the fixed candidate covariance.

F.1 Proof of Proposition 6: Deterministic Mean Budget

Proposition 6 is a direct consequence of the standard Gaussian variational objective (Hensman et al., 2013) and the Loève isometry (Berlinet & Thomas-Agnan, 2004). We give the calculation to make the observed-data budget explicit for arbitrary admissible inducing variables.

Proof of Proposition 6. We first calculate the ELBO change when the inducing mean is set to zero. We then identify the mean-dependent KL cost with the squared RKHS norm, so that the comparison does not depend on the choice of inducing variables.

1. Cancel the covariance terms in the ELBO comparison. First suppose that K_{uu} is invertible, and consider finite-KL posteriors $q(u) = \mathcal{N}(m, S)$ and $q^\circ(u) = \mathcal{N}(0, S)$. Their induced function means are h and zero, respectively. Their predictive covariance matrices on the observed inputs are identical, as the formula equation 5 depends on S and the prior covariances but not on m.

Write this shared predictive covariance as S_f. For a random vector with mean h_X and covariance S_f, expanding the square gives

$$\mathbb{E}\|y - f_X\|^2 = \|y - h_X\|^2 + \mathrm{tr}(S_f),$$

because the centred cross term has expectation zero. Hence the Gaussian expected log likelihood equation 4 differs between the two posteriors only through the squared mean residual. Similarly, all terms in equation 6 except the mean quadratic form are shared. Defining $J_u := m^\top K_{uu}^{-1} m$, we have

$$\mathrm{KL}(q(u) \,\|\, p(u)) - \mathrm{KL}(q^\circ(u) \,\|\, p(u)) = \frac{1}{2} J_u. \tag{95}$$

Subtracting the two ELBOs, remembering that the KL is subtracted from the expected log likelihood, gives

$$\mathcal{L}(q^\circ) - \mathcal{L}(q) = \frac{1}{2} J_u + \frac{\|y - h_X\|^2 - \|y\|^2}{2\sigma_d^2}.$$

This is the equality in equation 16.

2. Express the mean cost intrinsically in the RKHS. Let $\psi_i = \mathcal{L} u_i \in \mathcal{H}_{T_d}$ be the representers from Appendix A.1. The Loève isometry gives $[K_{uu}]_{ij} = \langle \psi_i, \psi_j \rangle_{\mathcal{H}_{T_d}}$ and $[k_u(x)]_i = \psi_i(x)$. Thus, with $\alpha = K_{uu}^{-1} m$, $h = \sum_i \alpha_i \psi_i$. Expanding its squared RKHS norm gives

$$\|h\|^2_{\mathcal{H}_{T_d}} = \sum_{i,j} \alpha_i \alpha_j \langle \psi_i, \psi_j \rangle_{\mathcal{H}_{T_d}} = \alpha^\top K_{uu} \alpha = m^\top K_{uu}^{-1} m = J_u. \tag{96}$$

In particular, every such inducing mean belongs to the RKHS.

If K_{uu} is singular, since it is PSD, we can diagonalise it as $K_{uu} = Q \,\mathrm{diag}(\lambda_1, \ldots, \lambda_M) Q^\top$, with $\lambda_i \geq 0$, for some orthogonal transformation Q. Equivalently, we have rotated inducing variables $\tilde{u} = Q^\top u$, with $\mathrm{Cov}(\tilde{u}) = \mathrm{diag}(\lambda_1, \ldots, \lambda_M)$. We retain only the coordinates with positive eigenvalues, discarding all $\tilde{u}_i$ with $\lambda_i = 0$. Every discarded coordinate has zero variance and is therefore zero almost surely under the centred prior. Any variational distribution q with a finite KL from the prior p must likewise satisfy $\tilde{u}_i = 0$, almost surely. The representer of a discarded coordinate has squared RKHS norm zero by the isometry and so contributes nothing to h. On the retained space the prior covariance is invertible, so the preceding calculation applies without changing the induced mean h or the ELBO. This justifies removing redundant inducing variables.

Finally, $\|y - h_X\|^2 \geq 0$ and equation 96 give the inequality in equation 16. If $\mathcal{L}(q) \geq \mathcal{L}(q^\circ)$, its left-hand side is nonpositive, and therefore

$$0 \geq \frac{1}{2} \|h\|^2_{\mathcal{H}_{T_d}} - \frac{\|y\|^2}{2\sigma_d^2}.$$

Rearranging and multiplying by $v_d > 0$ proves equation 17. All comparisons used the realised (X, y), so no distributional assumption on the observed dataset is needed. □

F.2 Gaussian Lower-Tail Bound and Proof of Proposition 7

Proof of Proposition 7. A large cost for the individual features does not by itself exclude cancellation in their sum. We control this by conditioning on the hidden parameters and using the independent Gaussian output weights. The resulting bound will depend on the average feature cost, but not on the network width.

1. Bound the lower tail of a Gaussian quadratic form. Let $G \in \mathbb{R}^{r\times r}$ be a finite positive-semidefinite matrix and $a \sim \mathcal{N}(0, I_r)$ a standard Gaussian vector. An orthogonal diagonalisation $G = O\,\mathrm{diag}(\lambda_j)O^\top$ gives $a^\top Ga = \sum_j \lambda_j Z_j^2$, where $Z = O^\top a \sim \mathcal{N}(0, I_r)$ has independent standard Gaussian entries. For $t > 0$, a one-dimensional Gaussian integral gives

$$\mathbb{E}e^{-\lambda_j Z_j^2/(2t)} = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-(1+\lambda_j/t)z^2/2}\, dz = (1 + \lambda_j/t)^{-1/2}.$$

The event $a^\top Ga \le t$ is equivalent to $e^{-a^\top Ga/(2t)} \ge e^{-1/2}$. Markov's inequality for this nonnegative random variable, followed by independence of the Z_j, therefore gives

$$\begin{aligned} \Pr\{a^\top Ga \le t\} &= \Pr\{e^{-a^\top Ga/(2t)} \ge e^{-1/2}\} \qquad (97)\\ &\le e^{1/2}\mathbb{E}e^{-a^\top Ga/(2t)} = e^{1/2}\prod_j (1 + \lambda_j/t)^{-1/2} \\ &\le \sqrt{\frac{et}{t + \mathrm{tr}\, G}} \le \sqrt{\frac{et}{\mathrm{tr}\, G}}, \qquad (98)\end{aligned}$$

where the last expression is used only when $\mathrm{tr}\, G > 0$. For the penultimate inequality, expanding the product shows $\prod_j (1 + \lambda_j/t) \ge 1 + \sum_j \lambda_j/t$, as all the higher-order product terms are nonnegative.

2. Apply the bound conditional on the hidden features. Put

$$C_\ell := v_d \|\sigma_{\theta_\ell}\|^2_{\mathcal{H}_{T_d}}, \qquad A := \frac{1}{L_d}\sum_{\ell=1}^{L_d} C_\ell, \tag{99}$$

allowing the value ∞. Conditional on hidden parameters for which all features belong to the RKHS, expansion of the norm of the sum in equation 18 gives

$$\begin{aligned} v_d\|g_d\|^2_{\mathcal{H}_{T_d}} &= a^\top Ga, \\ G_{ij} &= \frac{v_d}{L_d}\langle \sigma_{\theta_i}, \sigma_{\theta_j}\rangle_{\mathcal{H}_{T_d}}, \qquad \mathrm{tr}\, G = A. \end{aligned} \tag{100}$$

This is a positive-semidefinite Gram matrix, that is, for any coefficient vector c,

$$c^\top Gc = \frac{v_d}{L_d}\left\|\sum_i c_i \sigma_{\theta_i}\right\|^2_{\mathcal{H}_{T_d}} \ge 0.$$

The output weights remain standard Gaussian after conditioning, by their assumed independence from the hidden parameters. We may therefore apply equation 98 to this fixed G.

We must separately handle hidden features outside the RKHS, since their RKHS inner products in equation 100 are undefined. For fixed hidden parameters, let

$$V := \left\{ a \in \mathbb{R}^{L_d} : \sum_\ell a_\ell \sigma_{\theta_\ell} \in \mathcal{H}_{T_d} \right\}.$$

It is a linear subspace, as both the coefficient map and the RKHS are linear. If $\sigma_{\theta_j} \notin \mathcal{H}_{T_d}$, then the coordinate vector e_j is not in V, so V is a proper subspace. In finite dimensions it is contained in a hyperplane $\{a : c^\top a = 0\}$ for some $c \ne 0$. Since $c^\top a \sim \mathcal{N}(0, \|c\|^2)$ has a continuous density, this hyperplane has probability zero. Hence the conditional network cost $v_d||g_d||^2_{\mathcal{H}_{T_d}}$ is infinite almost surely whenever any feature is a non-member. The factor $1/\sqrt{L_d}$ does not affect this membership argument.

3. Bound the probability of a small average feature cost. For $z > 0$, define the fraction of features whose costs are below z by $F_z = L_d^{-1}\sum_\ell \mathbf{1}_{\{C_\ell < z\}}$. If $F_z \le 1/2$, at least half the costs are at least z, so $A \ge z/2$. Taking the contrapositive and applying Markov's inequality therefore gives

$$\Pr\{A < z/2\} \le \Pr\{F_z > 1/2\} \le 2\mathbb{E}F_z = 2\Pr\{C_1 < z\}. \tag{101}$$

The last equality uses identical marginal distributions: $\mathbb{E}[\mathbf{1}_{\{C_\ell<z\}}] = \Pr\{C_1 < z\}$ for every ℓ. Equality of the cost expectations is not needed; those expectations may be infinite.

On $\{A \geq z/2\}$, provided all feature costs are finite, equation 98 and equation 100 give

$$\Pr\Big\{ v_d \|g_d\|^2_{\mathcal{H}_{T_d}} \leq t \,\Big|\, \theta_1, \ldots, \theta_{L_d} \Big\} \leq \sqrt{\frac{et}{A}} \leq \sqrt{\frac{2et}{z}}.$$

If any feature has infinite cost, the same conditional probability is zero by Part 2. On the remaining event $\{A < z/2\}$ we use the trivial bound one. Averaging over the hidden parameters therefore gives

$$\begin{aligned} \Pr\{v_d \|g_d\|^2_{\mathcal{H}_{T_d}} \leq t\} &\leq \Pr\{A < z/2\} + \sqrt{\frac{2et}{z}} \Pr\{A \geq z/2\} \\ &\leq \Pr\{A < z/2\} + \sqrt{\frac{2et}{z}}. \end{aligned}$$

Using equation 101, we obtain

$$\Pr\{v_d \|g_d\|^2_{\mathcal{H}_{T_d}} \leq t\} \leq 2 \Pr\{v_d \|\sigma_\theta\|^2_{\mathcal{H}_{T_d}} < z\} + \sqrt{\frac{2et}{z}}. \tag{102}$$

For fixed $M > 0$ and $p > 0$, take $t = Md^p$ and $z = d^{p+1}$. The first term on the right-hand side of equation 102 tends to zero by Theorem 5, applied with exponent $p + 1$, while the second becomes $\sqrt{2eM/d} \to 0$. Hence

$$\Pr\{v_d \|g_d\|^2_{\mathcal{H}_{T_d}} \leq Md^p\} \to 0$$

for every $M > 0$, which is exactly equation 19. Neither term on the right-hand side of equation 102 depends on L_d, so the bound is uniform in the network width and the conclusion holds for every sequence of finite widths. □

F.3 Proof of Theorem 8

Proof of Theorem 8. Write $C_d = v_d \|g_d\|^2_{\mathcal{H}_{T_d}}$, assigning infinite cost to non-members. We first make the ELBO comparison on an event involving only C_d, then show that this event has probability tending to one.

1. Make a simultaneous deterministic comparison. Fix a realised network, a dataset satisfying equation 20, and any finite-KL inducing posterior whose function mean is g_d. Applying Proposition 6 with $h = g_d$ gives

$$\begin{aligned} v_d[\mathcal{L}(q^\circ) - \mathcal{L}(q)] &\geq \frac{1}{2}\left(C_d - v_d \frac{\|y\|^2}{\sigma_d^2} \right) \\ &\geq \frac{1}{2}(C_d - B_d). \end{aligned} \tag{103}$$

This proves equation 21 whenever a representation exists. On the event $C_d > B_d$, the right-hand side is strictly positive, so the centred posterior has strictly larger ELBO. This event depends only on the network and the fixed candidate covariance. Once it occurs, equation 103 applies to every dataset satisfying the budget and every inducing representation and covariance. There is consequently no need for a union bound over those choices, even if they are selected after seeing the target network.

By equation 96 and the argument extending to singular K_{uu}, every finite-KL inducing mean belongs to $\mathcal{H}_{T_d}$. Thus if $g_d \notin \mathcal{H}_{T_d}$, no such finite-KL q exists, and the comparison over finite-KL posteriors is vacuous.

2. Show that the cost exceeds the budget by a superpolynomial amount. Fix $p > 0$ and choose an exponent $s > \max\{p, r\}$. The budget assumption gives $B_d/d^s \leq Cd^{r-s} \to 0$, while Proposition 7 gives $C_d/d^s \to \infty$ in probability. Explicitly, for any $M > 0$,

$$\Pr\left\{ \frac{C_d - B_d}{2d^p} \leq M \right\} = \Pr\left\{ \frac{C_d}{d^s} \leq \frac{B_d}{d^s} + 2Md^{p-s} \right\} \longrightarrow 0. \tag{104}$$

Indeed, the deterministic threshold on the right tends to zero, so it is eventually at most one, and the probability is bounded by $\Pr\{C_d/d^s \leq 1\} \to 0$. This also covers $C_d = \infty$, which never belongs to these lower-tail events. Equation 104 proves the claimed divergence of the lower bound divided by d^p. In particular, $\Pr\{C_d \leq B_d\} \to 0$, and the simultaneous strict comparison from part 1 holds with probability tending to one. □