Operator-Based Generalization Bounds for Multitask Deep Learning

Anonymous authors
Paper under double-blind review

Abstract

We study generalization bounds for compositions of functions through an operator-theoretic (Koopman) framework. Existing analyses in this direction are primarily restricted to scalar-valued settings and to Sobolev-type reproducing kernel Hilbert spaces (RKHSs), where the resulting bounds depend on smoothness parameters. We extend this framework to vector-valued RKHSs, enabling the analysis of multi-output function classes and making explicit how task-coupling kernels enter the resulting Rademacher complexity bounds. Within this setting, we derive bounds that depend on operator norms, singular values, and determinant-based geometric quantities associated with the underlying linear maps. We further introduce a vector-valued Brownian RKHS formulation, which replaces Sobolev smoothness assumptions by a first-order Cameron–Martin-type structure. In this regime, the resulting bounds no longer depend on Sobolev smoothness exponents and instead exhibit a milder spectral dependence involving only operator norms and determinant factors. This highlights a qualitative difference between Sobolev- and Brownian-based analyses at the level of function spaces. We additionally study a shared operator-learning formulation for multitask transfer in vector-valued RKHSs deriving an exact representer theorem, a finite-dimensional reduction of the corresponding operator-learning problem, and transfer bounds for the induced operator class. We illustrate these effects empirically on synthetic data and MNIST, comparing the behavior of Sobolev and Brownian bounds during training.

1 Introduction

Understanding the generalization properties of deep networks remains a central challenge in learning theory. Classical approaches control capacity either through parameter counts or through norm-based quantities Bartlett & Mendelson (2002); Mohri et al. (2018). While norm-based bounds avoid explicit dependence on width, they often exhibit unfavorable scaling with depth and capture only coarse geometric information about the weight matrices (Neyshabur et al., 2015; Golowich et al., 2020; Bartlett et al., 2017; Wei & Ma, 2019; Liu et al., 2024). Compression-based approaches account for low-rank structure Arora et al. (2018), but do not address generalization in full-rank regimes.

A complementary perspective models neural network layers as Koopman operators acting on function spaces. This operator-theoretic viewpoint allows generalization bounds to be expressed in terms of spectral properties of the underlying transformations, and has been shown to yield improved behavior in scalar-valued Sobolev RKHS settings Hashimoto et al. (2024). However, existing analyses are restricted in two important ways: they are formulated for scalar outputs, and they rely on Sobolev-type RKHSs with smoothness constraints $s > d/2$, which impose structural limitations on the admissible function classes.

A function-space perspective. Rather than refining existing bounds within a fixed functional setting, this paper adopts a different viewpoint: we study how *changing the underlying function space* affects operator-based generalization bounds. In particular, we extend the Koopman framework to vector-valued RKHSs (vvRKHSs), and introduce a Brownian integral RKHS formulation that differs qualitatively from Sobolev spaces. This shift leads to a different spectral structure in the resulting bounds.

Overview of results. Working in vvRKHSs, we derive Koopman-based Rademacher complexity bounds for multi-output function classes. These bounds express capacity in terms of operator norms, singular values, and determinant-based geometric quantities associated with the weight matrices. We then introduce a vector-valued Brownian RKHS setting, in which the resulting bounds no longer depend on Sobolev smoothness exponents. In this regime, the resulting bounds depend only on operator norms and determinant factors, yielding a milder spectral dependence compared to Sobolev-based analyses in well-conditioned settings. This highlights a qualitative distinction between Sobolev and Brownian formulations at the level of function spaces.

Our contributions. We summarize our contributions as follows:

- **Koopman-based bounds in vector-valued RKHSs.** We extend the Koopman operator framework from scalar-valued to vector-valued RKHSs, deriving Rademacher complexity bounds for multi-output function classes and making explicit the role of task-coupling kernels through determinant and trace terms.
- **Brownian RKHS formulation.** We introduce a vector-valued Brownian integral RKHS setting in which Koopman-based bounds no longer depend on Sobolev smoothness exponents, but instead involve only operator norms and determinant-based geometric quantities.
- **Comparison of functional regimes.** We show that the Brownian formulation yields milder spectral dependence than Sobolev-based bounds in well-conditioned regimes, highlighting a structural difference between the two function-space settings.
- **Shared operator learning and transfer.** We introduce a shared operator-learning formulation for multitask transfer in vector-valued RKHSs, deriving an exact representer theorem, a finite-dimensional reduction of the corresponding operator-learning problem, and transfer bounds for the induced operator class. In this framework, the transferable object is a shared operator acting between Hilbert function spaces, rather than merely an output kernel or a finite-dimensional latent representation.
- **Empirical evaluation.** Experiments on synthetic data and MNIST illustrate the comparative behavior of Sobolev and Brownian bounds during training.

2 Related Works

Norm-based generalization bounds. A large body of work studies generalization in deep networks through norm-based capacity measures on the weight matrices (Neyshabur et al., 2015; Golowich et al., 2020; Bartlett et al., 2017). These bounds are typically dependent on spectral or Frobenius norms and avoid explicit dependence on layer width. However they often exhibit unfavorable scaling with depth and capture only coarse geometric information about the transformations. Variants based on reference matrices or margin-based quantities have also been proposed (Wei & Ma, 2020; Ju et al., 2022), while other approaches focus primarily on spectral norm control (Li et al., 2021). In contrast, operator-based analyses characterize generalization through the action of layers on function spaces, allowing bounds to depend on finer spectral quantities such as singular values and determinant-based volume distortions.

Operator-theoretic (Koopman) approaches. Recent work has introduced Koopman operator formulations for analyzing generalization in deep networks (Hashimoto et al., 2024). In this framework, each layer is viewed as inducing a composition operator acting on a reproducing kernel Hilbert space (RKHS), and generalization bounds are expressed in terms of operator norms. Existing results are primarily developed for scalar-valued functions in Sobolev-type RKHSs, where the resulting bounds depend on smoothness parameters and associated spectral quantities. The present work builds on this perspective by extending the analysis to vector-valued RKHSs and by considering alternative functional settings beyond Sobolev spaces.

Vector-valued and multi-output settings. Vector-valued RKHSs provide a natural framework for modeling multi-output function classes through matrix-valued kernels (Argyriou et al., 2006; 2008). In statistical learning theory, they have been used to study structured prediction, multitask learning, and kernel methods with coupled outputs. Classical generalization results include Rademacher complexity bounds for linear models (Maurer, 2006), excess risk bounds under trace-norm regularization (Pontil & Maurer, 2013), and local complexity analyses for structured function classes (Yousefi et al., 2018). While these works characterize how task coupling affects capacity, they do not address operator-based formulations of deep compositions.

This paper combines these two approaches using operator-based generalization bounds in vector-valued RKHSs and by introducing a Brownian RKHS formulation as an alternative to Sobolev spaces. In contrast to prior Koopman-based analyses, which operate within Sobolev function classes, our approach highlights how changing the underlying function space leads to a different spectral structure in the resulting bounds. This allows us to compare Sobolev and Brownian regimes within a unified operator-theoretic framework.

3 Preliminaries

In this section, we assemble the foundational tools employed throughout the paper. Section 3.1 specifies the notation and measure-theoretic conventions; Section 3.2 reviews the notation of vector-valued Rademacher complexity; and Section 3.3 introduces the Koopman representation of the network.

3.1 Notations

In the following, we introduce a few notations that will be used throughout the paper. We denote the set of natural numbers by $\mathbb{N} \coloneqq \{0, 1, 2, \dots\}$, and the set of positive integers by $\mathbb{N}^* \coloneqq \{1, 2, \dots\}$. We denote $[-R, R]^d$ the d-dimensional hypercube for $R > 0$. The set of non-negative real numbers is written as $\mathbb{R}^{\geq 0}$. For a positive integer n, we define the set $[n]$ as $\{1, 2, \dots, n\}$. The Cartesian product of a family of sets $(A_i)_{i \in I}$ is denoted by $\prod_{i \in I} A_i$. In particular, if $A_1 = \cdots = A_n = A$, we write A^n for the n-fold Cartesian product of A with itself. Let $\mathcal{U}$ be a topological space with a Borel sigma-field. We denote the space of probability measures on $\mathcal{U}$ as $\mathcal{P}(\mathcal{U})$. For a linear operator $\mathbf{W}$ on a Hilbert space, its range and kernel are denoted by $\mathrm{ran}(\mathbf{W})$ and $\ker(\mathbf{W})$, respectively. Its operator norm is denoted by $\|\mathbf{W}\|$. For an injective matrix $\mathbf{W} \in \mathbb{R}^{d \times d'}$ with $d \geq d'$, we define $|\det(\mathbf{W})| \coloneqq \left(\det\left(\mathbf{W}^\top \mathbf{W}\right)\right)^{1/2}$. For a function $p \in L^\infty\left(\mathbb{R}^d\right)$, its L^∞-norm is denoted by $\|p\|_\infty$. For a function h on $\mathbb{R}^d$ and a subspace $\mathcal{S}$ of $\mathbb{R}^d$, the restriction of h on $\mathcal{S}$ is denoted by $h|_{\mathcal{S}}$. With $\mathbb{S}_+^m \subset \mathbb{R}^{m \times m}$ we denote the set of $m \times m$ symmetric and positive semi-definite (p.s.d.) matrices. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel if it is symmetric and positive semidefinite, that is, for every $n \in \mathbb{N}^*$, every collection of points $(\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathcal{X}^n$, and all coefficients $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$, $\sum_{i,j=1}^n \alpha_i \, k(\mathbf{x}_i, \mathbf{x}_j) \, \alpha_j \geq 0$. Each such kernel uniquely determines a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k \subset \mathbb{R}^{\mathcal{X}}$, consisting of real-valued functions on $\mathcal{X}$. The space $\mathcal{H}_k$ is characterized by the facts that $k_{\mathbf{x}}(\cdot) \coloneqq k(\cdot, x) \in \mathcal{H}_k$, $\forall \, \mathbf{x} \in \mathcal{X}$, and it satisfies the reproducing property $h(\mathbf{x}) = \langle h, \, k_{\mathbf{x}} \rangle_{\mathcal{H}_k}$, $\forall \, h \in \mathcal{H}_k, \mathbf{x} \in \mathcal{X}$.[1] Finally, the RKHS admits the representation $\mathcal{H}_k = \overline{\mathrm{Span}\left\{k_{\mathbf{x}} : x \in \mathcal{X}\right\}}$, where $\mathrm{span}(\cdot)$ denotes the linear span of its argument and $\overline{(\cdot)}$ indicates closure.

Let $\mathcal{X}$ be a non-empty set. A bivariate function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m \times m}$, $m \in \mathbb{N}$, is called a matrix-valued kernel if $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x})^\top$ for all $(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2$, and for all $n \in \mathbb{N}$ and any $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ we have $\sum_{i,j=1}^n \mathbf{y}_i^\top K(\mathbf{x}_i, \mathbf{x}_j) \mathbf{y}_j \geq 0$. Fix $m \in \mathbb{N}$ and let $\mathbf{M} \in \mathbb{S}_+^m$ be a symmetric positive semidefinite matrix. Let K be a matrix-valued kernel. There exists a unique Hilbert space $\mathcal{H}_K$ of functions $f : \mathcal{X} \to \mathbb{R}^m$, called the vector-valued RKHS (vvRKHS) induced by K, such that for all $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathbb{R}^m$, and $f \in \mathcal{H}_K$, the function $\mathbf{x}' \mapsto K(\mathbf{x}, \mathbf{x}') \, \mathbf{y}$ belongs to $\mathcal{H}_K$ and $\langle f, \, K(\cdot, \mathbf{x}) \, \mathbf{y} \rangle_{\mathcal{H}_K} = f(\mathbf{x})^\top \mathbf{y}$. In the separable case $K(\mathbf{x}, \mathbf{y}) = k(\mathbf{x}, \mathbf{y}) \, \mathbf{M}$ with $\mathbf{M} \in \mathbb{S}_+^m$, the associated vvRKHS $\mathcal{H}_K$ consists of functions $f = (f_1, \dots, f_m) : \mathcal{X} \to \mathbb{R}^m$ with each $f_j \in \mathcal{H}_k$, and the inner product is given by

$$\langle f, g \rangle_{\mathcal{H}_K} = \sum_{i,j=1}^m \left(\mathbf{M}^{-1}\right)_{ij} \langle f_i, \, g_j \rangle_{\mathcal{H}_k}, \qquad (\mathbf{M} \succ 0).$$

[1] Here $k(\cdot, \mathbf{x})$ denotes the function $\mathbf{x}' \mapsto k(\mathbf{x}', \mathbf{x})$.

Equivalently,

$$\|f\|_{\mathcal{H}_K}^2 = \sum_{i,j=1}^{m} \left(\mathbf{M}^{-1}\right)_{ij} \langle f_i,\, f_j\rangle_{\mathcal{H}_k}.$$

If $\mathbf{M}$ is positive semidefinite (but not invertible), the same formulas hold on the corresponding quotient space induced by $\ker(\mathbf{M})$. In particular, the norm $\|f\|_{\mathcal{H}_K}$ is equivalent to the unweighted product norm $\|f\|_{\mathcal{H}_K}^2 \asymp \sum_{j=1}^{m} \|f_j\|_{\mathcal{H}_k}^2$, where the equivalence constants depend only on the eigenvalues of $\mathbf{M}$.

Let $\mathcal{H}_{K_1}$ and $\mathcal{H}_{K_2}$ be vvRKHSs on $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, induced by the matrix-valued kernels K_1 and K_2, respectively. Given a measurable function $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$, the associated Koopman operator acts by composition from $\mathcal{H}_{K_2}$ to $\mathcal{H}_{K_1}$. We define its domain as $\mathcal{D}_f = \{g \in \mathcal{H}_{K_2} \,:\, g \circ f \in \mathcal{H}_{K_1}\}$, namely all functions in $\mathcal{H}_{K_2}$ whose pullback through f belongs to $\mathcal{H}_{K_1}$. The Koopman operator associated with f is the map $\mathcal{K}_f : \mathcal{D}_f \to \mathcal{H}_{K_1}$, $\mathcal{K}_f g = g \circ f$. For $\mathbf{b} \in \mathbb{R}^d$, the translation operator $T_{\mathbf{b}} : \mathbb{R}^d \to \mathbb{R}^d$ is $T_{\mathbf{b}}(x) \coloneqq x + \mathbf{b}$, and the associated Koopman operator is $\mathcal{K}_{\mathbf{b}} g \coloneqq g \circ T_{\mathbf{b}}$.

A multi-index is a vector $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, with length $|\alpha| = \sum_{i=1}^{d} \alpha_i$. The symbol $D^\alpha f$ denotes the weak derivative of order α. For $s \geq 0$, $H^s\left(\mathbb{R}^d\right)$ denotes the standard Sobolev space, identified with $W^{s,2}\left(\mathbb{R}^d\right)$. For a linear subspace $\mathcal{S} \subset \mathbb{R}^d$, we define $H^s\left(\mathcal{S}, \mathbb{R}^m\right)$ as the Sobolev space obtained by identifying $\mathcal{S}$ with $\mathbb{R}^{\dim \mathcal{S}}$ via an orthonormal basis. The notation $\widehat{f}$ denotes the Fourier transform of a function f, and $\operatorname{supp}(f)$ its support. For $m \in \mathbb{N}$ and $s \geq 0$, $H^s\left(\mathbb{R}^d, \mathbb{R}^m\right)$ denotes the standard vector-valued Sobolev space, equipped with its usual Hilbert norm. The associated matrix-valued Sobolev kernel is written as $K_s(x, x') = \mathbf{\Phi}_s(x - x')$, where $\mathbf{\Phi}_s : \mathbb{R}^d \to \mathbb{R}^{m \times m}$ is a translation-invariant matrix-valued function whose Fourier transform satisfies the usual Sobolev decay condition. When $\mathbf{\Phi}_s$ is radial, we write $\mathbf{\Phi}_s(x) = \phi_s(\|x\|_2)$ for a scalar profile function $\phi_s : [0, \infty) \to \mathbb{R}$.

Recall that the Brownian kernel on $\mathbb{R}$ is $k^{(\mathrm{B})}(x, x') = \frac{|x| + |x'| - |x - x'|}{2}$, $\forall x, x' \in \mathbb{R}$, which is non-negatively 1-homogeneous in the sense that $k^{(\mathrm{B})}(ax, ax') = a\, k^{(\mathrm{B})}(x, x')$ for all $a \in \mathbb{R}^{\geq 0}$.

Let $\mathcal{D} \subset \mathbb{R}^d$ be a domain, Ω a topological space equipped with a Borel probability measure $\mu \in \mathcal{P}(\Omega)$, and for each $\boldsymbol{\omega} \in \Omega$ let $k^{(\boldsymbol{\omega})} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ be a scalar kernel with RKHS $\mathcal{H}_{k^{(\boldsymbol{\omega})}}$, satisfying $\int_\Omega k^{(\boldsymbol{\omega})}(\mathbf{x}, \mathbf{x})\, \mathrm{d}\mu(\boldsymbol{\omega}) < \infty$, $\forall\, \mathbf{x} \in \mathcal{D}$. Define

$$\mathcal{H}_\oplus = \left\{ (f_{\boldsymbol{\omega}})_{\boldsymbol{\omega} \in \Omega} \in \prod_{\boldsymbol{\omega} \in \Omega} \mathcal{H}_{k^{(\boldsymbol{\omega})}} : \int_\Omega \|f_{\boldsymbol{\omega}}\|_{\mathcal{H}_{k^{(\boldsymbol{\omega})}}}^2 \, \mathrm{d}\mu(\boldsymbol{\omega}) < \infty \right\}, \tag{1}$$

with inner product $\langle f, g\rangle_{\mathcal{H}_\oplus} = \int_\Omega \langle f_{\boldsymbol{\omega}}, g_{\boldsymbol{\omega}}\rangle_{\mathcal{H}_{k^{(\boldsymbol{\omega})}}} \, \mathrm{d}\mu(\boldsymbol{\omega})$. Then, by Hotz & Telschow (2012, Theorem 3.1), the space

$$\mathcal{H}_k = \left\{ f : \mathcal{D} \to \mathbb{R} : f(\mathbf{x}) = \int_\Omega f_{\boldsymbol{\omega}}(\mathbf{x})\, \mathrm{d}\mu(\boldsymbol{\omega}) \;\; \forall \mathbf{x} \in \mathcal{D},\; (f_{\boldsymbol{\omega}}) \in \mathcal{H}_\oplus \right\} \tag{2}$$

is an RKHS (the integral RKHS) with kernel $k(\mathbf{x}, \mathbf{y}) = \int_\Omega k^{(\boldsymbol{\omega})}(\mathbf{x}, \mathbf{y})\, \mathrm{d}\mu(\boldsymbol{\omega})$, $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, and norm

$$\|f\|_{\mathcal{H}_k}^2 = \inf_{\substack{g \in \mathcal{H}_\oplus: \\ f(\mathbf{x}) = \int_\Omega g_{\boldsymbol{\omega}}(\mathbf{x}) \mathrm{d}\mu(\boldsymbol{\omega}),\ \forall \mathbf{x}}} \int_\Omega \|g_{\boldsymbol{\omega}}\|_{\mathcal{H}_{k^{(\boldsymbol{\omega})}}}^2 \, \mathrm{d}\mu(\boldsymbol{\omega}). \tag{3}$$

Let $\mathcal{X} \subset \mathbb{R}^d$ be a domain, $\mathbf{M} \in \mathbb{S}_+^m$, and let $K(\mathbf{x}, \mathbf{x}') \coloneqq k^{(\mathrm{B})}(\mathbf{x}, \mathbf{x}')\mathbf{M}$ be a separable Brownian kernel on $\mathcal{X} \times \mathcal{X}$. The induced vector-valued Brownian RKHS is denoted by $H^{(\mathrm{B})}\left(\mathcal{X}, \mathbb{R}^m\right)$.

For an activation function $f : \mathbb{R} \to \mathbb{R}$, the sup-norm is $\|f\|_\infty \coloneqq \sup_{x \in \mathbb{R}} |f(x)|$. For a function $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ with component functions $(\sigma_i)_{i=1}^d$, the uniform derivative bound is $\|\sigma'\|_\infty \coloneqq \max_{1 \leq i \leq d} \sup_{x \in \mathbb{R}^d} |\sigma_i'(x)|$, and analogously for $\left\|\left(\sigma^{-1}\right)'\right\|_\infty$.

3.2 Vector-valued Rademacher complexity

We consider a general multiple-output regression framework. Let us briefly recall the fundamental setting of supervised learning. We are given a training sample $\mathcal{D}_{XY,n} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \sim \nu_{XY}^n$, where (i) $\nu_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ denotes the joint distribution governing the relationship between the input X and the output Y; (ii) $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}^m$ denote the input and output spaces, respectively; and (iii) ν_X represents the marginal distribution of X.[2] The objective is to learn a function f from the data $\mathcal{D}_{XY,n}$ such that $f(x)$ provides an accurate prediction of the corresponding output $\mathbf{y}$ for unseen inputs $\mathbf{x}$. Based on the definition of Rademacher complexity, the vector-valued Rademacher complexity is defined as follows:

Definition 1 (empirical vector-valued Rademacher complexity)**.** *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}^m$ over an input space $\mathcal{X}$, and let*

$$\boldsymbol{\sigma}_i = (\sigma_{i1}, \ldots, \sigma_{im}) \sim \mathrm{Rad}^m, \quad i \in [n],$$

be independent Rademacher vectors, i.e. σ_{ij} are i.i.d. random variables uniformly distributed on $\{-1, +1\}$. Then, for a fixed dataset $\mathcal{D}_n = \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$, the empirical vector-valued Rademacher complexity of $\mathcal{F}$ is defined as

$$\widehat{\mathfrak{R}}_n^m(\mathcal{F}) \coloneqq \mathbb{E}_{\mathrm{Rad}^{mn}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \langle \boldsymbol{\sigma}_i, f(\mathbf{x}_i) \rangle \right| \right], \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner product in $\mathbb{R}^m$. Equivalently, one may write $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_n) \sim \mathrm{Rad}^{nm}$.

3.3 Koopman representation of deep networks

We study the generalization behaviour of an L-layer neural network

$$f = g \circ \mathbf{b}_L \circ \mathbf{W}_L \circ \sigma_{L-1} \circ \cdots \circ \sigma_1 \circ \mathbf{b}_1 \circ \mathbf{W}_1, \tag{5}$$

mapping $\mathbb{R}^{d_0}$ to $\mathbb{R}^m$. Here, $\mathbf{W}_l \in \mathbb{R}^{d_l \times d_{l-1}}$ are injective linear maps,

$$\mathbf{b}_l(x) = x + a_l, \qquad a_l \in \mathbb{R}^{d_l},$$

are bias shifts, $\sigma_l : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$ are nonlinear activations, and $g : \mathbb{R}^{d_L} \to \mathbb{R}^m$ is the terminal representation map. Using the Koopman operator viewpoint, the network can be written as the operator product

$$f = \mathcal{K}_{\mathbf{W}_1} \mathcal{K}_{\mathbf{b}_1} \mathcal{K}_{\sigma_1} \cdots \mathcal{K}_{\mathbf{W}_{L-1}} \mathcal{K}_{\mathbf{b}_{L-1}} \mathcal{K}_{\sigma_{L-1}} \mathcal{K}_{\mathbf{W}_L} \mathcal{K}_{\mathbf{b}_L}\, g. \tag{6}$$

The Koopman representation separates the contribution of each layer at the level of function spaces: $\mathcal{K}_{\mathbf{W}_l}$ encodes the action of the linear map $\mathbf{W}_l$, $\mathcal{K}_{\mathbf{b}_l}$ accounts for translations, and $\mathcal{K}_{\sigma_l}$ captures the effect of the nonlinear activations. Rather than controlling generalization directly through the layer weights, we instead bound the corresponding Koopman operators acting on suitable vector-valued RKHSs. This operator-theoretic factorization exposes spectral-geometric quantities such as singular values and determinant factors in the resulting Rademacher complexity bounds. Each layer $\mathbb{R}^{d_l}$ is equipped with a vector-valued Sobolev RKHS $H^{s_l}\left(\mathbb{R}^{d_l}, \mathbb{R}^m\right)$, generated by the separable matrix-valued kernel

$$K_{s_l}(\mathbf{x}, \mathbf{x}') = k_{s_l}(\mathbf{x}, \mathbf{x}')\, \mathbf{M}, \qquad \text{for all } \mathbf{x}, \mathbf{x}' \in \mathcal{X}$$

where k_{s_l} is a scalar Sobolev kernel of order $s_l > d_l/2$, and $\mathbf{M} \in \mathbb{S}_+^m$. Under these assumptions, the Koopman operators satisfy

$$\begin{aligned} \mathcal{K}_{\mathbf{W}_l} &: H^{s_l}\left(\mathbb{R}^{d_l}, \mathbb{R}^m\right) \to H^{s_{l-1}}\left(\mathbb{R}^{d_{l-1}}, \mathbb{R}^m\right), \\ \mathcal{K}_{\mathbf{b}_l},\ \mathcal{K}_{\sigma_l} &: H^{s_l}\left(\mathbb{R}^{d_l}, \mathbb{R}^m\right) \to H^{s_l}\left(\mathbb{R}^{d_l}, \mathbb{R}^m\right), \end{aligned}$$

which ensures that the Koopman factorization (6) is well-defined.

We impose the following assumption throughout this section.

[2]The notation ν_X will be used later when introducing Rademacher complexities.

Assumption 1. *The final nonlinear transformation $g \in H^{s_L}\left(\mathbb{R}^{d_L}, \mathbb{R}^m\right)$, and the Koopman operators $\mathcal{K}_{\sigma_l}$ are bounded for $l = 1, \ldots, L-1$.*

To ensure that the terminal representation map satisfies Assumption 1, we may choose

$$g(\mathbf{x}) = e^{-\|\mathbf{x}\|^2}\mathbf{M}\mathbf{c}^\top, \qquad \mathbf{x} \in \mathbb{R}^{d_L},$$

for coefficients $\mathbf{c} \in \mathbb{R}^m$, matrix $\mathbf{M} \in \mathbb{S}_+^m$, and Sobolev order $s_L > d_L/2$. With this choice, $g \in H^{s_L}\left(\mathbb{R}^{d_L}, \mathbb{R}^m\right)$.

Remark 1. *Let g be a smooth function which does not decay at infinity, for example the sigmoid activation. Although $H^{s_L}\left(\mathbb{R}^{d_L}, \mathbb{R}^m\right)$ does not contain such functions globally, one may construct $\widetilde{g} \in H^{s_L}\left(\mathbb{R}^{d_L}, \mathbb{R}^m\right)$ such that $\widetilde{g}(\mathbf{x}) = g(\mathbf{x})$ on a sufficiently large compact region and replace g by $\widetilde{g}$ in practical settings.*

Assumption 2. *There exists $\kappa > 0$ such that*

$$k_{s_0}(\mathbf{x}, \mathbf{x}) \leq \kappa, \qquad \forall\, \mathbf{x} \in \mathcal{X}.$$

Lemma B1 shows that under suitable smoothness and bi-Lipschitz assumptions on σ_l, the Koopman operator $\mathcal{K}_{\sigma_l}$ is bounded on $H^{s_l}\left(\mathbb{R}^{d_l}, \mathbb{R}^m\right)$, for $l \in [L-1]$.

As activation functions, one may use smooth variants of Leaky ReLU, such as those introduced in Biswas et al. (2022).

We now derive Rademacher complexity bounds for the corresponding function class.

4 Koopman Formulation in Vector-Valued Sobolev RKHSs

We first derive Koopman-based Rademacher complexity bounds in vector-valued Sobolev RKHSs. The resulting estimates decompose into three components: a kernel-dependent factor associated with the input space, a product of Koopman operator norms corresponding to the nonlinear activations, and a product of spectral-geometric terms associated with the linear layers. The spectral-geometric contribution involves both operator norms and determinant factors. The determinant terms quantify the volume distortion induced by the linear maps, while the Sobolev-symbol ratios describe how regularity is transported across layers under the associated Koopman operators.

4.1 Koopman-based Sobolev bounds for invertible deep networks

In this subsection, we consider the case $d_l = d$, $l = 0, \ldots, L$, for some fixed $d \in \mathbb{N}$. For constants $C, D > 0$, define

$$\mathsf{W}(C, D) \coloneqq \left\{\mathbf{W} \in \mathbb{R}^{d\times d} \;\middle|\; \|\mathbf{W}\| \leq C,\ |\det(\mathbf{W})| \geq D\right\},$$

and consider the invertible hypothesis class $\mathcal{F}_{\text{inv}} \coloneqq \{f \in \mathcal{F} \mid \mathbf{W}_l \in \mathsf{W}(C, D)\}$. The following theorem gives a Koopman-based Rademacher complexity bound for $\mathcal{F}_{\text{inv}}$.

Theorem 1. *The Rademacher complexity $\mathfrak{R}_n^m(\mathcal{F}_{\text{inv}})$ satisfies*

$$\widehat{\mathfrak{R}}_n^m(\mathcal{F}_{\text{inv}}) \leq \left(\sqrt{\frac{\kappa \operatorname{Tr}(\mathbf{M})}{n}}\right) \|g\|_{H^{s_L}(\mathbb{R}^d, \mathbb{R}^m)}$$
$$\cdot \sup_{\mathbf{W}_l \in \mathsf{W}(C,D)} \prod_{l=1}^{L} \sup_{\boldsymbol{\omega} \in \mathbb{R}^d} \left| \frac{\left(1 + \left\|\mathbf{W}_l^\top \boldsymbol{\omega}\right\|_2^2\right)^{s_{l-1}}}{\left(1 + \|\boldsymbol{\omega}\|_2^2\right)^{s_l}} \right|^{1/2} \frac{1}{|\det(\mathbf{W}_l)|^{1/2}} \prod_{l=1}^{L-1} \|\mathcal{K}_{\sigma_l}\|,$$

where $M \in \mathbb{S}_+^m$.

The bound in Theorem 1 follows from the Koopman factorization of the network and the corresponding layer-wise operator estimates. In particular, the determinant terms arise from the change-of-variables structure of the linear Koopman operators and quantify the volume distortion induced by the weight matrices. Applying Lemma 5 of Hashimoto et al. (2024) to Theorem 1 yields the following corollary.

Corollary 1. *Assume*

$$H^{s_l}\left(\mathbb{R}^d, \mathbb{R}^m\right) = H^s\left(\mathbb{R}^d, \mathbb{R}^m\right), \qquad l = 0, \ldots, L,$$

for some $s > d/2$. Then,

$$\widehat{\mathfrak{R}}_n^m\left(\mathcal{F}_{\text{inv}}\right) \leq \max\left\{1, C^s\right\} \left(\sqrt{\frac{\kappa \operatorname{Tr}(\mathbf{M})}{nD}}\right) \|g\|_{H^s(\mathbb{R}^d, \mathbb{R}^m)} \prod_{l=1}^{L-1} \|\mathcal{K}_{\sigma_l}\|.$$

4.2 Koopman-based Sobolev bounds for injective deep networks

We now extend the invertible-layer analysis to injective architectures with width expansion, where the linear maps $\mathbf{W}_l : \mathbb{R}^{d_{l-1}} \to \mathbb{R}^{d_l}$ are injective but not necessarily square. In this setting, the determinant of $\mathbf{W}_l$ is no longer defined. The relevant geometric quantity is instead $\det\left(\mathbf{W}_l^\top \mathbf{W}_l\right)$, which measures the volume distortion induced by $\mathbf{W}_l$ on its range. A second feature of the injective setting is the appearance of geometric factors G_l, which quantify how much of the function mass is concentrated on the lower-dimensional subspace $\operatorname{ran}(\mathbf{W}_l) \subset \mathbb{R}^{d_l}$. These factors capture the interaction between Sobolev regularity and the geometry induced by the injective linear layers. For constants $C, D > 0$, define

$$\mathsf{W}_l(C, D) \coloneqq \left\{ \mathbf{W} \in \mathbb{R}^{d_{l-1} \times d_l} \;\middle|\; d_l \geq d_{l-1},\ \|\mathbf{W}\| \leq C,\ \det\left(\mathbf{W}^\top \mathbf{W}\right)^{1/2} \geq D \right\},$$

and consider the injective hypothesis class $\mathcal{F}_{\text{inj}} \coloneqq \{f \in \mathcal{F} \mid \mathbf{W}_l \in \mathsf{W}_l(C, D)\}$. For

$$f_l \coloneqq g \circ \mathbf{b}_L \circ \mathbf{W}_L \circ \sigma_{L-1} \circ \mathbf{b}_{L-1} \circ \mathbf{W}_{L-1} \circ \cdots \circ \sigma_l \circ \mathbf{b}_l,$$

define

$$\mathrm{G}_l \coloneqq \frac{\left\| f_l|_{\operatorname{ran}(\mathbf{W}_l)} \right\|_{H^{s_{l-1}}(\operatorname{ran}(\mathbf{W}_l), \mathbb{R}^m)}}{\|f_l\|_{H^{s_l}\left(\mathbb{R}^{d_l}, \mathbb{R}^m\right)}}.$$

The following theorem gives the corresponding injective Koopman-based Rademacher complexity bound.

Theorem 2. *The Rademacher complexity $\widehat{\mathfrak{R}}_n^m\left(\mathcal{F}_{\text{inj}}\right)$ satisfies*

$$\widehat{\mathfrak{R}}_n^m\left(\mathcal{F}_{\text{inj}}\right) \leq \left(\sqrt{\frac{\kappa \operatorname{Tr}(\mathbf{M})}{n}}\right) \|g\|_{H^{s_L}\left(\mathbb{R}^{d_L}, \mathbb{R}^m\right)} \cdot \sup_{\mathbf{W}_l \in \mathsf{W}_l(C,D)} \prod_{l=1}^{L} \mathrm{G}_l \sup_{\boldsymbol{\omega} \in \operatorname{ran}(\mathbf{W}_l)} \left| \frac{1 + \left\|\mathbf{W}_l^\top \boldsymbol{\omega}\right\|_2^2}{1 + \|\boldsymbol{\omega}\|_2^2} \right|^{s_{l-1}/2} \frac{1}{\det\left(\mathbf{W}_l^\top \mathbf{W}_l\right)^{1/4}} \prod_{l=1}^{L-1} \|\mathcal{K}_{\sigma_l}\|.$$

where $M \in \mathbb{S}_+^m$.

Remark 2 (Non-injective weight matrices)**.** *The injectivity assumption in Theorem 2 guarantees boundedness of the composition operator associated with $\mathbf{W}_l$. When $\mathbf{W}_l$ is rank-deficient, directions in $\ker(\mathbf{W}_l)$ collapse, and the corresponding Koopman operator is no longer bounded on Sobolev-type RKHSs. Following the operator-regularization viewpoint introduced in Hashimoto et al. (2024, Section 4.3.1), one can replace the singular volume factor by the stabilized quantity $\det\left(I + \mathbf{W}_l^\top \mathbf{W}_l\right)$, which remains strictly positive even for rank-deficient layers. This leads to the modified estimate $1/\det\left(I + \mathbf{W}_l^\top \mathbf{W}_l\right)^{1/4}$ in place of $1/\det\left(\mathbf{W}_l^\top \mathbf{W}_l\right)^{1/4}$. The resulting bound remains finite in the non-injective setting and preserves the same operator-theoretic interpretation in terms of regularized volume distortion.*

5 Koopman Formulation in Vector-Valued Brownian RKHSs

We now consider an alternative functional setting based on vector-valued Brownian RKHSs. In contrast to the Sobolev framework, where regularity is controlled by smoothness exponents s_l, the Brownian RKHS is

governed by a first-order Cameron–Martin-type structure. This change of function space leads to a qualitatively different form of operator-based generalization bounds. In particular, Koopman operators induced by linear maps in the Brownian setting depend only on operator norms and determinant-based volume distortions, without any smoothness parameters. As a consequence, the resulting Rademacher complexity bounds exhibit a different spectral structure: each layer contributes a factor of the form $\|\mathbf{W}_l\|/|\det(\mathbf{W}_l)|^{1/2}$, rather than $\max\{1, \|\mathbf{W}_l\|^{s_l}\}/|\det(\mathbf{W}_l)|^{1/2}$ as in the Sobolev case. This yields a milder dependence on spectral norms in well-conditioned regimes. The following theorem formalizes this behavior for invertible architectures and highlights the contrast with Sobolev-based bounds.

5.1 Koopman-based Brownian bounds for invertible networks

In this subsection, for some $d \in \mathbb{N}$, we assume $d_l = d$ for all $l = 0, \ldots, L$. Let $R > 0$ and set $\mathcal{X} = [-R, R]^d$. Consider the separable matrix-valued Brownian kernel $K(\mathbf{x}, \mathbf{x}') = k^{(\mathrm{B})}(\mathbf{x}, \mathbf{x}')\,\mathbf{M}$, for $\mathbf{M} \in \mathbb{S}_+^m$ and $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, with associated vector-valued Brownian RKHS $H^{(\mathrm{B})}(\mathcal{X}, \mathbb{R}^m)$ as defined in Section 3. Throughout this subsection, we assume that all functions are supported in $[-R, R]^d$. For constants $C, D > 0$, define the class of admissible invertible weight matrices

$$\mathsf{W}(C, D) \coloneqq \left\{ \mathbf{W} \in \mathbb{R}^{d \times d} \;\middle|\; \|\mathbf{W}\| \le C,\ |\det(\mathbf{W})| \ge D \right\},$$

and consider the Brownian RKHS hypothesis class

$$\mathcal{F}_{\mathrm{inv}}^{(\mathrm{B})} \coloneqq \left\{ f = \mathcal{K}_{\mathbf{W}_1}\mathcal{K}_{\mathbf{b}_1}\mathcal{K}_{\sigma_1}\cdots\mathcal{K}_{\mathbf{W}_L}\mathcal{K}_{\mathbf{b}_L}\, g \;\middle|\; g \in H^{(\mathrm{B})}(\mathcal{X}, \mathbb{R}^m),\ \mathbf{W}_l \in \mathsf{W}(C, D) \right\},$$

where $\mathcal{K}_{\mathbf{b}_l}$ denote translation operators and σ_l satisfy the assumptions of Lemma B4. The following theorem gives the resulting Brownian RKHS Rademacher complexity bound.

Theorem 3. *The Rademacher complexity $\mathfrak{R}_n^m\left(\mathcal{F}_{\mathrm{inv}}^{(\mathrm{B})}\right)$ is bounded as*

$$\widehat{\mathfrak{R}}_n^m\left(\mathcal{F}_{\mathrm{inv}}^{(\mathrm{B})}\right) \le \sqrt{\frac{\kappa\,\mathrm{Tr}(\mathbf{M})}{n}}\,\|g\|_{H^{(\mathrm{B})}(\mathcal{X},\mathbb{R}^m)} \sup_{\mathbf{W}_l \in \mathsf{W}(C,D)} \prod_{l=1}^{L} \frac{\|\mathbf{W}_l\|}{|\det(\mathbf{W}_l)|^{1/2}} \prod_{l=1}^{L-1} \|\sigma_l'\|_\infty \left\| \left(\sigma_l^{-1}\right)' \right\|_\infty^{1/2}.$$

A corresponding extension to non-injective (rank-deficient) weight matrices can be obtained by combining the operator-regularization mechanism discussed in Theorem 2 with the Brownian Koopman estimates derived here. In that setting, the determinant factor $|\det(\mathbf{W}_l)|^{1/2}$ is replaced by the stabilized quantity $\det\left(I + \mathbf{W}_l^\top \mathbf{W}_l\right)^{1/2}$, yielding a finite Brownian RKHS complexity bound even in the non-injective regime. We omit the full derivation for brevity.

Remark 3. *The vector-valued Brownian RKHS formulation considered here corresponds to a different functional regime from the Sobolev RKHS framework of Hashimoto et al. (2024). Rather than refining the same estimates within a fixed RKHS class, the present approach changes the underlying function space itself, which modifies both the associated Koopman operator bounds and the resulting Rademacher complexity estimates. The main distinctions are summarized below.*

- *Function-space regime: Hashimoto et al. work in Sobolev RKHSs induced by Fourier multipliers $p(\boldsymbol{\omega}) = \left(1 + \|\boldsymbol{\omega}\|_2^2\right)^{-s}$, $s > d/2$, where bounded evaluation relies on Sobolev embedding. In contrast, our analysis is carried out in a vector-valued Brownian RKHS, which corresponds on compact domains to a first-order Cameron–Martin-type structure and avoids higher-order Sobolev smoothness parameters.*
- *Koopman-based bounds for activation maps: Hashimoto et al. derive bounds for activation-induced Koopman operators in Sobolev RKHSs. In the special case $s = 1$ with elementwise σ, they obtain*

  $$\|\mathcal{K}_\sigma\| \le \|\det(J_{\sigma^{-1}})\|_\infty \max\left\{1, \|\partial_j \sigma\|_\infty\right\},$$

  and note that deriving tight bounds for larger s is challenging (their Remark 3). In our vector-valued Brownian RKHS framework, the activation Koopman bounds take the form

  $$\|\mathcal{K}_\sigma\| \le C_d\,\|\sigma'\|_\infty \left\| \left(\sigma^{-1}\right)' \right\|_\infty^{1/2},$$

with dependence only on first-order derivative quantities.

- *Structure of the final Rademacher bounds: For invertible weights, the Koopman-based Sobolev bound of Hashimoto et al. (2024) (Theorem 4 and Lemma 5) yields*

$$\mathcal{O}\left(\sup_{\mathbf{W}_j\in \mathsf{W}(C,D)} \prod_{j=1}^{L} \frac{\max\left\{1, \|\mathbf{W}_j\|^{s_j}\right\}}{\left|\det\left(\mathbf{W}_j\right)\right|^{1/2}} \prod_{j=1}^{L-1} \left\|K_{\sigma_j}\right\|\right).$$

In contrast, our vector-valued Brownian RKHS result has the form

$$\mathcal{O}\left(\sup_{\mathbf{W}_l\in \mathsf{W}(C,D)} \prod_{l=1}^{L} \frac{\|\mathbf{W}_l\|}{\left|\det\left(\mathbf{W}_l\right)\right|^{1/2}} \prod_{l=1}^{L-1} \|\sigma_l'\|_\infty \left\|\left(\sigma_l^{-1}\right)'\right\|_\infty^{1/2}\right).$$

Thus, the Brownian formulation removes higher-order Sobolev exponents from the spectral factors and replaces them by first-order operator quantities. For $\|\mathbf{W}_l\| \geq 1$ and $s_l > 1$, the Sobolev formulation therefore exhibits higher-order dependence on spectral norms, although the two bounds operate on different hypothesis spaces and are not directly comparable.

- *Injective weights and geometric G_j factors: Hashimoto et al. (2024) introduce geometric factors G_j to control injective non-surjective layers through restrictions to $\operatorname{ran}(\mathbf{W}_j)$. These factors depend on Sobolev regularity and Grassmannian averaging (Appendix C in their paper). In the Brownian setting, analogous geometric factors naturally arise (cf. Lemma B5), but their dependence is tied only to the geometry of $\ker(\mathbf{W}_j)$ and the first-order Brownian structure, rather than higher-order Sobolev smoothness parameters.*

- *Non-injective weight matrices (Brownian setting): The extension of the Brownian RKHS analysis to non-injective (rank-deficient) linear layers requires additional care, since the associated Koopman operator is no longer bounded due to collapse along $\ker(\mathbf{W}_l)$. Using the same stabilized volume mechanism discussed above, one replaces the singular determinant factor by $\det\left(I + \mathbf{W}_l^\top \mathbf{W}_l\right)$, which remains strictly positive even in the rank-deficient setting. This suggests that an analogous Brownian RKHS Rademacher complexity bound can be derived by combining the same operator-regularization mechanism with the Brownian Koopman estimates developed here, although we do not pursue the full derivation in the present work.*

6 Shared Operator Learning and Transfer

In the previous sections, we studied operator-theoretic generalization bounds for fixed Koopman composition chains acting on vector-valued RKHSs. We now consider a related extension in which a shared operator is learned jointly across multiple tasks.

The purpose of this section is not to introduce a new empirical transfer-learning framework, but rather to show that the operator-theoretic viewpoint naturally admits a shared operator-learning formulation with an exact finite-dimensional representer structure. In contrast to multitask formulations based solely on output kernels or finite-dimensional latent representations, the transferable object considered here is an operator acting between Hilbert function spaces.

The results below are inspired by operator-learning representer methodologies for Koopman operators in scalar RKHS settings (Khosravi, 2023). However, the present formulation differs in that it operates in vector-valued RKHSs and studies shared multitask transfer operators motivated by the Koopman product framework developed earlier in this paper.

6.1 Shared operator-learning formulation

Let $\mathcal{H}_{\mathrm{out}}$ be a Hilbert space containing the final output-side maps g appearing in the Koopman factorization. Thus, $\mathcal{H}_{\mathrm{out}}$ does not refer to the penultimate hidden-layer activations, but to the function space in which

the terminal map $g : \mathbb{R}^{d_L} \to \mathbb{R}^m$ is assumed to lie. In the present framework, one may take for example $\mathcal{H}_{\text{out}} = H^{s_L}\left(\mathbb{R}^{d_L}, \mathbb{R}^m\right)$ or $\mathcal{H}_{\text{out}} = H^{(\mathrm{B})}\left(\mathcal{X}, \mathbb{R}^m\right)$. In the present framework, one may take for example $\mathcal{H}_{\text{out}} = H^{s_L}\left(\mathbb{R}^{d_L}, \mathbb{R}^m\right)$ or $\mathcal{H}_{\text{out}} = H^{(\mathrm{B})}\left(\mathcal{X}, \mathbb{R}^m\right)$. Let $\mathcal{H}_K$ be a vector-valued RKHS induced by a matrix-valued kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m\times m}$. We consider a shared operator $T : \mathcal{H}_{\text{out}} \to \mathcal{H}_K$. Motivated by the Koopman factorization introduced earlier, one may formally view T as being induced by products of the form $\mathcal{K}_{\mathbf{W}_1}\mathcal{K}_{\mathbf{b}_1}\mathcal{K}_{\sigma_1}\cdots\mathcal{K}_{\mathbf{W}_L}\mathcal{K}_{\mathbf{b}_L}$. Let $M \in \mathbb{N}$ denote the number of source tasks. For each $t \in [M]$, let $g_t \in \mathcal{H}_{\text{out}}$ denote a task-specific terminal representation, and let $(\mathbf{x}_{ti}, \mathbf{y}_{ti}) \in \mathcal{X} \times \mathbb{R}^m$, $i \in [n_t]$, be the corresponding training samples. We study the regularized operator-learning problem

$$\min_{T} \sum_{t=1}^{M}\sum_{i=1}^{n_t} \ell\left(\mathbf{y}_{ti}, (Tg_t)(\mathbf{x}_{ti})\right) + \lambda \|T\|_{\text{HS}}^2, \tag{7}$$

where $T : \mathcal{H}_{\text{out}} \to \mathcal{H}_K$, $\lambda > 0$, the loss ℓ is convex and continuous in its second argument, and $\|\cdot\|_{\text{HS}}$ denotes the Hilbert–Schmidt norm.

6.2 Representer theorem

The next theorem shows that every minimum-norm minimizer of (7) admits a finite-rank representation determined entirely by the training samples and the task representations.

Theorem 4. *Assume that* (7) *admits a minimizer. Then every minimum-norm minimizer* $\widehat{T} : \mathcal{H}_{\text{out}} \to \mathcal{H}_K$ *admits the representation*

$$\widehat{T} = \sum_{t=1}^{M}\sum_{i=1}^{n_t}\sum_{a=1}^{m} c_{tia}\left(K\left(\cdot, \mathbf{x}_{ti}\right)\mathbf{e}_a\right) \otimes g_t.$$

where $(\mathbf{e}_a)_{a=1}^m$ *is the standard basis of* $\mathbb{R}^m$*. Equivalently,*

$$\left(\widehat{T}g\right)(\mathbf{x}) = \sum_{t=1}^{M}\sum_{i=1}^{n_t}\sum_{a=1}^{m} c_{tia}\left\langle g_t, g\right\rangle_{\mathcal{H}_{\text{out}}} K\left(\mathbf{x}, \mathbf{x}_{ti}\right)\mathbf{e}_a.$$

6.3 Finite-dimensional reduction

The representer theorem implies that the infinite-dimensional optimization problem (7) reduces exactly to a finite-dimensional optimization problem over the coefficients (c_{tia}).

Theorem 5. *Let* $\widehat{T}$ *be a minimum-norm minimizer of* (7)*. Then the corresponding coefficients solve the finite-dimensional optimization problem*

$$\begin{aligned}\min_{c} \sum_{t=1}^{M}\sum_{i=1}^{n_t}\left\|\mathbf{y}_{ti} - \sum_{t'=1}^{M}\sum_{j=1}^{n_{t'}}\sum_{a=1}^{m} c_{t'ja}\left\langle g_{t'}, g_t\right\rangle_{\mathcal{H}_{\text{out}}} K\left(\mathbf{x}_{ti}, \mathbf{x}_{t'j}\right)\mathbf{e}_a\right\|_2^2 \\ + \lambda \sum_{\substack{t,i,a\\ t',j,b}} c_{tia}c_{t'jb}\left\langle g_t, g_{t'}\right\rangle_{\mathcal{H}_{\text{out}}} \mathbf{e}_a^\top K\left(\mathbf{x}_{ti}, \mathbf{x}_{t'j}\right)\mathbf{e}_b.\end{aligned}$$

6.4 Transfer bound

We finally record the generalization consequence of the learned shared operator. Let $K(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}')\mathbf{M}$, $\mathbf{M} \in \mathbb{S}_+^m$, with $k(\mathbf{x}, \mathbf{x}) \le \kappa$. Conditionally on the learned operator $\widehat{T}$, define $\mathcal{F}_{\widehat{T}}(B) \coloneqq \left\{\widehat{T}g : \|g\|_{\mathcal{H}_{\text{out}}} \le B\right\}$. Let

$$R_0(f) \coloneqq \mathbb{E}\left[\ell(Y, f(X))\right]$$

denote the target population risk, and let

$$f_0^\star \in \arg\min_f R_0(f).$$

Let $\widehat{f}_0$ be an empirical risk minimizer over $\mathcal{F}_{\widehat{T}}(B)$ on a target sample of size n_0, and define

$$A_{\widehat{T}}(B) \coloneqq \inf_{\|g\|_{\mathcal{H}_{\text{out}}} \le B} R_0\left(\widehat{T} g\right) - R_0\left(f_0^\star\right).$$

Proposition 1. *Assume that the loss is L_ℓ-Lipschitz and bounded by C_ℓ. Then, conditionally on $\widehat{T}$, with probability at least $1-\delta$,*

$$R_0\left(\widehat{f}_0\right) - R_0\left(f_0^\star\right) \le A_{\widehat{T}}(B) + 4L_\ell B \left\|\widehat{T}\right\| \sqrt{\frac{\kappa \operatorname{Tr}(\mathbf{M})}{n_0}} + C_\ell \sqrt{\frac{\log(1/\delta)}{2n_0}}.$$

In particular, since $\left\|\widehat{T}\right\| \le \left\|\widehat{T}\right\|_{\text{HS}}$, the same bound holds with $\left\|\widehat{T}\right\|_{\text{HS}}$ in place of $\left\|\widehat{T}\right\|$.

Remark 4.

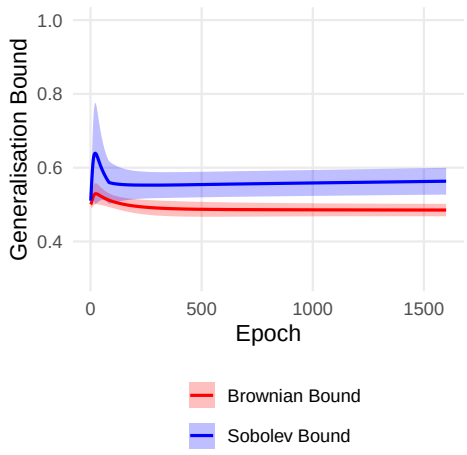
- *Theorem 1 does not imply that shared operator learning universally improves transfer performance. It shows only that whenever a useful shared operator admits controlled operator norm, the induced transfer class inherits a corresponding Rademacher complexity bound.*
- *The present viewpoint differs from multitask formulations based solely on output kernels or finite-dimensional latent representations. Here, the transferable object is an operator acting between Hilbert function spaces.*
- *When the learned operator $\widehat{T}$ is induced by Koopman composition chains satisfying the bounds of Theorems 1 to 3, the transfer complexity term $\left\|\widehat{T}\right\|$ inherits the corresponding spectral-geometric structure involving Koopman operator norms, determinant factors, and Brownian/Sobolev function-space geometry.*

7 Experimental Details

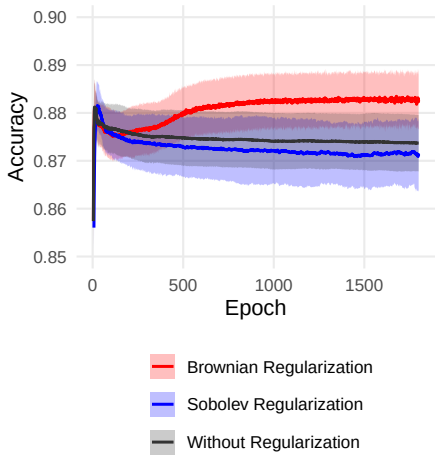
In this section, we present an empirical comparison between the Sobolev Bound and the Brownian Bound. We first conduct a numerical study of both bounds in a single regression setting. We then evaluate their effectiveness as regularizers in a multi-class classification task, analyzing their impact on model performance and generalization.

7.1 Validity of the bound

To numerically evaluate the bound, we conducted a synthetic data experiment inspired by Hashimoto et al. (2024). The synthetic data consider a regression problem in $\mathbb{R}^3$, where the target function t is defined as $t(x) \coloneqq \exp\left(-\|2x - \mathbf{1}\|^2\right)$, for all $x \in \mathbb{R}^3$. We employ a fully connected neural network of the form $f(x) \coloneqq g\left(W_2\, \sigma\left(W_1 x + b_1\right) + b_2\right)$, with parameter matrices and vectors $W_1 \in \mathbb{R}^{3\times 3}$, $b_1 \in \mathbb{R}^3$, $W_2 \in \mathbb{R}^{6\times 3}$, and $b_2 \in \mathbb{R}^6$. The matrices W_1 and W_2 are initialized using orthogonal initialization via Saxe et al. (2014), and the bias vectors b_1 and b_2 are initialized by sampling from a uniform distribution. As activation function we use the smooth Leaky ReLU from Biswas et al. (2022), and for the output mapping we take $g(x) \coloneqq \exp\left(-\|x\|^2\right)$. The network is trained for 1600 epochs using gradient-based optimization with learning rate 3×10^{-3}, and an L_2 penalty of 10^{-4} is employed. The Sobolev Bound is defined as $\text{SB} \coloneqq \prod_{l=1}^{L} \|W_l\|^{s_l} / \left(\det\left(I + W_l^\top W_l\right)\right)^{1/4}$, for $s_l \coloneqq \frac{d_l + 0.1}{2}$, while the Brownian Bound is given by $\text{BB} \coloneqq \prod_{l=1}^{L} \|W_l\| / \left(\det\left(I + W_l^\top W_l\right)\right)^{1/4}$. Figure 1(a) displays the numerical values of these two bounds, evaluated throughout training across five independent weight initializations. As training progresses, one observes that the Brownian Bound is both more stable and tighter than the Sobolev Bound. Additional experimental details are reported in Appendix D.



(a)



(b)

Figure 1: (a) Evolution of the Sobolev Bound and the Brownian Bound throughout training on the synthetic dataset. (b) Test performance on MNIST comparing models regularized via the Sobolev Bound, via the Brownian Bound, and a baseline without regularization.

8 Conclusion

We developed an operator-theoretic framework for deriving Rademacher complexity bounds for deep multi-output networks via Koopman operators acting on vector-valued RKHSs. Our results extend existing Koopman-based Sobolev analyses to vector-valued settings and to non-square architectures, and introduce a novel vector-valued Brownian RKHS formulation in which the resulting spectral factors no longer involve higher-order Sobolev exponents, but instead depend only on operator norms and determinant-based geometric quantities. This leads to a different functional regime with milder spectral dependence in well-conditioned settings, while preserving the operator-theoretic structure of the bounds. We additionally introduced a novel shared operator-learning formulation for multitask transfer in vector-valued RKHSs, deriving an exact representer theorem, a finite-dimensional reduction of the corresponding operator-learning problem, and transfer bounds for the induced operator class. More broadly, the results highlight how changing the underlying function space modifies the geometry and spectral behaviour of Koopman-based generalization estimates. Future directions include extending the Brownian framework to convolutional architectures and investigating connections between these functional regimes and practical training dynamics in deep multi-output models.

References

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *NeurIPS*, 2006.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 2008.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Compression-based framework for erm. In *ICML*, 2018.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities. *Journal of Machine Learning Research*, 2002.

Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *NeurIPS*, 2017.

Koushik Biswas, Sandeep Kumar, Shilpak Banerjee, and Ashish Pandey. Smooth maximum unit. In *CVPR*, 2022.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *Information and Inference*, 9(2):473–504, 2020.

Yuka Hashimoto, Sho Sonoda, Isao Ishikawa, Atsushi Nitanda, and Taiji Suzuki. Koopman-based generalization bound: New aspect for full-rank weights. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=JN7TcCm9LF`.

Thomas Hotz and Fabian J. E. Telschow. Representation by integrating reproducing kernels. *arXiv preprint*, 2012.

Haotian Ju, Dongyue Li, and Hongyang Zhang. Robust fine-tuning of deep neural networks. In *ICML*, 2022.

Mohammad Khosravi. Representer theorem for learning koopman operators. *IEEE Transactions on Automatic Control*, 68(5):2995–3010, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Shuai Li, Kui Jia, Yuxin Wen, Tongliang Liu, and Dacheng Tao. Orthogonal deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Fanghui Liu, Leello Dadi, and Volkan Cevher. Learning with norm constrained neural networks. *Journal of Machine Learning Research*, 2024.

Andreas Maurer. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 2006.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, Cambridge, MA, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pp. 1376–1401, 2015.

Massimiliano Pontil and Andreas Maurer. Excess risk bounds for multitask learning. In *COLT*, 2013.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions in deep linear networks. In *ICLR*, 2014.

Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. In *NeurIPS*, 2019.

Colin Wei and Tengyu Ma. Improved sample complexities for deep neural networks. In *ICLR*, 2020.

Navid Yousefi, Yunwen Lei, and Marius Kloft. Local rademacher complexity-based learning guarantees. *Journal of Machine Learning Research*, 2018.

A Additional Notations

We collect here several additional operator-theoretic and functional-analytic notations used throughout the appendices and proofs.

Let $\mathcal{H}_1, \mathcal{H}_2$ be Hilbert spaces. We denote by $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ the space of bounded linear operators from $\mathcal{H}_1$ to $\mathcal{H}_2$, equipped with the operator norm

$$\|T\| \coloneqq \sup_{\|f\|_{\mathcal{H}_1} \leq 1} \|Tf\|_{\mathcal{H}_2}.$$

The space of Hilbert–Schmidt operators from $\mathcal{H}_1$ to $\mathcal{H}_2$ is denoted by $\mathcal{L}_2(\mathcal{H}_1, \mathcal{H}_2)$, equipped with the Hilbert–Schmidt norm

$$\|T\|_{\mathrm{HS}} \coloneqq \left(\sum_{j=1}^{\infty} \|Te_j\|_{\mathcal{H}_2}^2 \right)^{1/2},$$

where $(e_j)_{j\in\mathbb{N}}$ is an arbitrary orthonormal basis of $\mathcal{H}_1$. For $u \in \mathcal{H}_2$ and $g \in \mathcal{H}_1$, the rank-one operator $u \otimes g \in \mathcal{L}_2(\mathcal{H}_1, \mathcal{H}_2)$ is defined by

$$(u \otimes g)\, h = u\, \langle g, h\rangle_{\mathcal{H}_1}, \qquad h \in \mathcal{H}_1.$$

For a subset A of a vector space, we denote by $\mathrm{Span}(A)$ its linear span. For a matrix A, the trace is denoted by $\mathrm{Tr}(A)$, and the diagonal matrix generated by a vector v is written as $\mathrm{diag}(v)$. For a closed subspace V of a Hilbert space $\mathcal{H}$, the orthogonal projection onto V is denoted by $P_V : \mathcal{H} \to V$. Finally, throughout the appendices, the notation $\widehat{f}$ denotes the Fourier transform of f.

B Proofs

All proofs of Theorems 1 to 5 and Proposition 1 appear in the main text.

B.1 Proof of Theorem 1

We first estimate the Koopman operators associated with the translation and linear layers.

Translation operators. Let $d_l = d$ for $l = 0, \ldots, L$. Let $h \in H^{s_l}(\mathbb{R}^d, \mathbb{R}^m)$. Using the shift-invariance property of the Fourier transform,

$$\widehat{h \circ \mathbf{b}_l}(\boldsymbol{\omega}) = e^{-i\mathbf{a}_l^\top \boldsymbol{\omega}} \widehat{h}(\boldsymbol{\omega}),$$

where (a) follows from the Fourier transform of translations. Consequently,

$$\begin{aligned}\|\mathcal{K}_{\mathbf{b}_l} h\|^2_{H^{s_l}(\mathbb{R}^d,\mathbb{R}^m)} &\overset{(a)}{=} \int_{\mathbb{R}^d} \left(1 + \|\boldsymbol{\omega}\|_2^2\right)^{s_l} \left\|\widehat{h \circ \mathbf{b}_l}(\boldsymbol{\omega})\right\|_2^2 \mathrm{d}\boldsymbol{\omega} \overset{(b)}{=} \int_{\mathbb{R}^d} \left(1 + \|\boldsymbol{\omega}\|_2^2\right)^{s_l} \left|e^{-i\mathbf{a}_l^\top \boldsymbol{\omega}}\right|^2 \left\|\widehat{h}(\boldsymbol{\omega})\right\|_2^2 \mathrm{d}\boldsymbol{\omega} \\ &\overset{(c)}{=} \|h\|^2_{H^{s_l}(\mathbb{R}^d,\mathbb{R}^m)},\end{aligned}$$

where (a) follows from the definition of the Sobolev norm, (b) uses the previous identity, and (c) follows from $\left|e^{-i\mathbf{a}_l^\top \boldsymbol{\omega}}\right| = 1$. Hence, $\|\mathcal{K}_{\mathbf{b}_l}\| = 1$.

Linear layers. Let $h \in H^{s_l}(\mathbb{R}^d, \mathbb{R}^m)$. Using the scaling property of the Fourier transform,

$$\widehat{h \circ \mathbf{W}_l}(\boldsymbol{\omega}) = \frac{1}{|\det(\mathbf{W}_l)|} \widehat{h}\left(\mathbf{W}_l^{-\top} \boldsymbol{\omega}\right),$$

Therefore,

$$\begin{aligned}\|\mathcal{K}_{\mathbf{W}_l} h\|^2_{H^{s_{l-1}}(\mathbb{R}^d,\mathbb{R}^m)} &\overset{(a)}{=} \int_{\mathbb{R}^d} \left(1 + \|\boldsymbol{\omega}\|_2^2\right)^{s_{l-1}} \left\|\widehat{h \circ \mathbf{W}_l}(\boldsymbol{\omega})\right\|_2^2 \mathrm{d}\boldsymbol{\omega} \\ &\overset{(b)}{=} \frac{1}{|\det(\mathbf{W}_l)|} \int_{\mathbb{R}^d} \left(1 + \|\boldsymbol{\omega}\|_2^2\right)^{s_{l-1}} \left\|\widehat{h}\left(\mathbf{W}_l^{-\top}\boldsymbol{\omega}\right)\right\|_2^2 \mathrm{d}\boldsymbol{\omega} \\ &\overset{(c)}{=} \frac{1}{|\det(\mathbf{W}_l)|} \int_{\mathbb{R}^d} \frac{\left(1 + \left\|\mathbf{W}_l^\top \boldsymbol{\xi}\right\|_2^2\right)^{s_{l-1}}}{\left(1 + \|\boldsymbol{\xi}\|_2^2\right)^{s_l}} \left(1 + \|\boldsymbol{\xi}\|_2^2\right)^{s_l} \left\|\widehat{h}(\boldsymbol{\xi})\right\|_2^2 \mathrm{d}\boldsymbol{\xi} \\ &\overset{(d)}{\leq} \sup_{\boldsymbol{\xi}\in\mathbb{R}^d} \left|\frac{\left(1 + \left\|\mathbf{W}_l^\top \boldsymbol{\xi}\right\|_2^2\right)^{s_{l-1}}}{\left(1 + \|\boldsymbol{\xi}\|_2^2\right)^{s_l}}\right| \frac{1}{|\det(\mathbf{W}_l)|} \|h\|^2_{H^{s_l}(\mathbb{R}^d,\mathbb{R}^m)},\end{aligned}$$

where (a) follows from the definition of the Sobolev norm, (b) uses the Fourier scaling identity, (c) applies the change of variables $\boldsymbol{\omega} = \mathbf{W}_l^\top \boldsymbol{\xi}$, and (d) follows by taking the supremum of the quotient. Hence,

$$\|\mathcal{K}_{\mathbf{W}_l}\| \leq \sup_{\boldsymbol{\xi}\in\mathbb{R}^d} \left|\frac{\left(1 + \left\|\mathbf{W}_l^\top \boldsymbol{\xi}\right\|_2^2\right)^{s_{l-1}}}{\left(1 + \|\boldsymbol{\xi}\|_2^2\right)^{s_l}}\right|^{1/2} \frac{1}{|\det(\mathbf{W}_l)|^{1/2}}. \tag{8}$$

Let $\mathbf{k}_0 \in \mathbb{R}^{n\times n}$ be the scalar Gram matrix of k_{s_0} and $\mathbf{K}_0 := \mathbf{k}_0 \otimes \mathbf{M}$ be the Gram matrix of K_{s_0}. Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^{d_0}$ ($d_0 = d$). Using the reproducing property of $H^{s_0}\left(\mathbb{R}^{d_0}, \mathbb{R}^m\right)$, we obtain

$$
\begin{aligned}
\widehat{\mathfrak{R}}_n^m\left(\mathcal{F}_{\text{inv}}\right) &\overset{(a)}{=} \frac{1}{n}\mathbb{E}_{\text{Rad}^{mn}}\left[\sup_{f\in\mathcal{F}_{\text{inv}}}\left|\sum_{i=1}^n \left\langle \boldsymbol{\sigma}_i, f\left(\mathbf{x}_i\right)\right\rangle_{\mathbb{R}^m}\right|\right] \\
&\overset{(b)}{=} \frac{1}{n}\mathbb{E}_{\text{Rad}^{mn}}\left[\sup_{f\in\mathcal{F}_{\text{inv}}}\left|\left\langle \sum_{i=1}^n K_{s_0}\left(\cdot, \mathbf{x}_i\right)\boldsymbol{\sigma}_i, f\right\rangle_{H^{s_0}\left(\mathbb{R}^{d_0},\mathbb{R}^m\right)}\right|\right] \\
&\overset{(c)}{\leq} \frac{1}{n}\mathbb{E}_{\text{Rad}^{mn}}\left[\left\|\sum_{i=1}^n K_{s_0}\left(\cdot, \mathbf{x}_i\right)\boldsymbol{\sigma}_i\right\|_{H^{s_0}\left(\mathbb{R}^{d_0},\mathbb{R}^m\right)} \sup_{f\in\mathcal{F}_{\text{inv}}}\|f\|_{H^{s_0}\left(\mathbb{R}^{d_0},\mathbb{R}^m\right)}\right] \\
&\overset{(d)}{=} \frac{1}{n}\sup_{f\in\mathcal{F}_{\text{inv}}}\|f\|_{H^{s_0}\left(\mathbb{R}^{d_0},\mathbb{R}^m\right)}\mathbb{E}_{\text{Rad}^{mn}}\left[\left\|\sum_{i=1}^n K_{s_0}\left(\cdot, \mathbf{x}_i\right)\boldsymbol{\sigma}_i\right\|_{H^{s_0}\left(\mathbb{R}^{d_0},\mathbb{R}^m\right)}\right] \\
&\overset{(e)}{\leq} \frac{1}{n}\sup_{f\in\mathcal{F}_{\text{inv}}}\|f\|_{H^{s_0}\left(\mathbb{R}^{d_0},\mathbb{R}^m\right)}\left(\operatorname{Tr}\left(\mathbf{K}_0\right)\right)^{1/2},
\end{aligned}
$$

where (a) follows from the definition of empirical Rademacher complexity, (b) uses the reproducing property, (c) follows from the Cauchy–Schwarz inequality, (d) pulls the supremum outside the expectation, and (e) follows from Jensen's inequality and the reproducing property. Using Assumption 2, $\operatorname{Tr}\left(\mathbf{K}_0\right) \leq \kappa n \operatorname{Tr}\left(\mathbf{M}\right)$, which implies

$$
\widehat{\mathfrak{R}}_n^m\left(\mathcal{F}_{\text{inv}}\right) \leq \sqrt{\frac{\kappa \operatorname{Tr}\left(\mathbf{M}\right)}{n}}\sup_{f\in\mathcal{F}_{\text{inv}}}\|f\|_{H^{s_0}\left(\mathbb{R}^{d_0},\mathbb{R}^m\right)}. \tag{9}
$$

Finally, using the Koopman factorization (6), together with $\|\mathcal{K}_{\mathbf{b}_l}\| = 1$, Assumption 1, and (8), we obtain

$$
\begin{aligned}
\|f\|_{H^{s_0}\left(\mathbb{R}^{d_0},\mathbb{R}^m\right)} &\overset{(a)}{\leq} \left(\prod_{l=1}^L \|\mathcal{K}_{\mathbf{W}_l}\|\right)\left(\prod_{l=1}^{L-1}\|\mathcal{K}_{\sigma_l}\|\right)\|g\|_{H^{s_L}\left(\mathbb{R}^d,\mathbb{R}^m\right)} \\
&\overset{(b)}{\leq} \|g\|_{H^{s_L}\left(\mathbb{R}^d,\mathbb{R}^m\right)}\prod_{l=1}^L \sup_{\boldsymbol{\xi}\in\mathbb{R}^d}\left|\frac{\left(1+\left\|\mathbf{W}_l^\top\boldsymbol{\xi}\right\|_2^2\right)^{s_{l-1}}}{\left(1+\|\boldsymbol{\xi}\|_2^2\right)^{s_l}}\right|^{1/2}\frac{1}{\left|\det\left(\mathbf{W}_l\right)\right|^{1/2}}\prod_{l=1}^{L-1}\|\mathcal{K}_{\sigma_l}\|,
\end{aligned}
$$

where (a) follows from submultiplicativity of operator norms, and (b) uses (8). Substituting this estimate into (9) yields the claimed result. ■

B.2 Proof of Theorem 2

For $h \in H^{s_l}\left(\mathbb{R}^{d_l}, \mathbb{R}^m\right)$, using the scaling property of the Fourier transform, viewing

$$
\mathbf{W}_l : \mathbb{R}^{d_{l-1}} \to \operatorname{ran}\left(\mathbf{W}_l\right),
$$

and taking the QR decomposition $\mathbf{W}_l = \mathbf{Q}_l\mathbf{R}_l$, we obtain

$$
\widehat{h\circ\mathbf{W}_l}\left(\boldsymbol{\omega}\right) \overset{(a)}{=} \frac{1}{\left|\det\left(\mathbf{R}_l\right)\right|}\int_{\operatorname{ran}(\mathbf{W}_l)} h\left(\mathbf{x}\right)e^{-i\mathbf{x}^\top\mathbf{W}_l^{-\top}\boldsymbol{\omega}}\,\mathrm{d}\mathbf{x} \overset{(b)}{=} \frac{1}{\det\left(\mathbf{W}_l^\top\mathbf{W}_l\right)^{1/2}}\widehat{h}\left(\mathbf{W}_l^{-\top}\boldsymbol{\omega}\right),
$$

where (a) follows from the Fourier scaling formula on the subspace $\operatorname{ran}\left(\mathbf{W}_l\right)$, and (b) uses

$$
\left|\det\left(\mathbf{R}_l\right)\right| = \det\left(\mathbf{W}_l^\top\mathbf{W}_l\right)^{1/2}.
$$

Since $\mathbf{W}_l$ is injective,

$$\dim\left(\operatorname{ran}\left(\mathbf{W}_l\right)\right) = d_{l-1}, \qquad \dim\left(\operatorname{ran}\left(\mathbf{W}_l\right)^{\perp}\right) = d_l - d_{l-1}.$$

We now estimate the corresponding Koopman operator norm:

$$\begin{aligned}
\left\|\mathcal{K}_{\mathbf{W}_l} h\right\|^2_{H^{s_{l-1}}\left(\mathbb{R}^{d_{l-1}},\mathbb{R}^m\right)} &\overset{(a)}{=} \int_{\mathbb{R}^{d_{l-1}}} \frac{\left(1+\|\boldsymbol{\omega}\|_2^2\right)^{s_{l-1}}}{\det\left(\mathbf{W}_l^\top \mathbf{W}_l\right)} \left\|\widehat{h}\left(\mathbf{W}_l^{-\top}\boldsymbol{\omega}\right)\right\|_2^2 \mathrm{d}\boldsymbol{\omega} \\
&\overset{(b)}{=} \int_{\operatorname{ran}(\mathbf{W}_l)} \frac{\left(1+\left\|\mathbf{W}_l^\top\boldsymbol{\omega}\right\|_2^2\right)^{s_{l-1}}}{\det\left(\mathbf{W}_l^\top \mathbf{W}_l\right)^{1/2}} \left\|\widehat{h}\left(\boldsymbol{\omega}\right)\right\|_2^2 \mathrm{d}\boldsymbol{\omega} \overset{(c)}{=} \int_{\operatorname{ran}(\mathbf{W}_l)} \frac{\left(1+\left\|\mathbf{W}_l^\top\boldsymbol{\omega}\right\|_2^2\right)^{s_{l-1}} \left(1+\|\boldsymbol{\omega}\|_2^2\right)^{s_{l-1}}}{\det\left(\mathbf{W}_l^\top \mathbf{W}_l\right)^{1/2} \left(1+\|\boldsymbol{\omega}\|_2^2\right)^{s_{l-1}}} \left\|\widehat{h}\left(\boldsymbol{\omega}\right)\right\|_2^2 \mathrm{d}\boldsymbol{\omega} \\
&\overset{(d)}{\leq} \left\|h|_{\operatorname{ran}(\mathbf{W}_l)}\right\|^2_{H^{s_{l-1}}(\operatorname{ran}(\mathbf{W}_l),\mathbb{R}^m)} \sup_{\boldsymbol{\omega}\in\operatorname{ran}(\mathbf{W}_l)} \left|\frac{1+\left\|\mathbf{W}_l^\top\boldsymbol{\omega}\right\|_2^2}{1+\|\boldsymbol{\omega}\|_2^2}\right|^{s_{l-1}} \frac{1}{\det\left(\mathbf{W}_l^\top \mathbf{W}_l\right)^{1/2}},
\end{aligned} \tag{10}$$

where (a) follows from the previous Fourier identity, (b) uses the change of variables induced by $\mathbf{W}_l$, (c) multiplies and divides by $\left(1+\|\boldsymbol{\omega}\|_2^2\right)^{s_{l-1}}$, and (d) follows by taking the supremum. We have

$$\begin{aligned}
\|f\|_{H^{s_0}\left(\mathbb{R}^{d_0},\mathbb{R}^m\right)} &\overset{(a)}{=} \left\|\mathcal{K}_{\mathbf{W}_1} f_1\right\|_{H^{s_0}\left(\mathbb{R}^{d_0},\mathbb{R}^m\right)} \\
&\overset{(b)}{\leq} \left\|f_1|_{\operatorname{ran}(\mathbf{W}_1)}\right\|_{H^{s_0}(\operatorname{ran}(\mathbf{W}_1),\mathbb{R}^m)} \sup_{\boldsymbol{\omega}\in\operatorname{ran}(\mathbf{W}_1)} \left|\frac{1+\left\|\mathbf{W}_1^\top\boldsymbol{\omega}\right\|_2^2}{1+\|\boldsymbol{\omega}\|_2^2}\right|^{s_0/2} \frac{1}{\det\left(\mathbf{W}_1^\top \mathbf{W}_1\right)^{1/4}} \\
&\overset{(c)}{=} \mathrm{G}_1 \left\|f_1\right\|_{H^{s_1}\left(\mathbb{R}^{d_1},\mathbb{R}^m\right)} \sup_{\boldsymbol{\omega}\in\operatorname{ran}(\mathbf{W}_1)} \left|\frac{1+\left\|\mathbf{W}_1^\top\boldsymbol{\omega}\right\|_2^2}{1+\|\boldsymbol{\omega}\|_2^2}\right|^{s_0/2} \frac{1}{\det\left(\mathbf{W}_1^\top \mathbf{W}_1\right)^{1/4}} \\
&\overset{(d)}{\leq} \mathrm{G}_1 \left\|\mathcal{K}_{\sigma_1}\right\| \left\|\mathcal{K}_{\mathbf{W}_2} f_2\right\|_{H^{s_1}\left(\mathbb{R}^{d_1},\mathbb{R}^m\right)} \sup_{\boldsymbol{\omega}\in\operatorname{ran}(\mathbf{W}_1)} \left|\frac{1+\left\|\mathbf{W}_1^\top\boldsymbol{\omega}\right\|_2^2}{1+\|\boldsymbol{\omega}\|_2^2}\right|^{s_0/2} \frac{1}{\det\left(\mathbf{W}_1^\top \mathbf{W}_1\right)^{1/4}} \\
&\quad\vdots \\
&\overset{(e)}{\leq} \left\|\mathcal{K}_{\mathbf{W}_L} f_L\right\|_{H^{s_{L-1}}\left(\mathbb{R}^{d_{L-1}},\mathbb{R}^m\right)} \prod_{l=1}^{L-1} \mathrm{G}_l \left\|\mathcal{K}_{\sigma_l}\right\| \sup_{\boldsymbol{\omega}\in\operatorname{ran}(\mathbf{W}_l)} \left|\frac{1+\left\|\mathbf{W}_l^\top\boldsymbol{\omega}\right\|_2^2}{1+\|\boldsymbol{\omega}\|_2^2}\right|^{s_{l-1}/2} \frac{1}{\det\left(\mathbf{W}_l^\top \mathbf{W}_l\right)^{1/4}} \\
&\overset{(f)}{\leq} \left\|f_L|_{\operatorname{ran}(\mathbf{W}_L)}\right\|_{H^{s_{L-1}}(\operatorname{ran}(\mathbf{W}_L),\mathbb{R}^m)} \prod_{l=1}^{L} \sup_{\boldsymbol{\omega}\in\operatorname{ran}(\mathbf{W}_l)} \left|\frac{1+\left\|\mathbf{W}_l^\top\boldsymbol{\omega}\right\|_2^2}{1+\|\boldsymbol{\omega}\|_2^2}\right|^{s_{l-1}/2} \frac{1}{\det\left(\mathbf{W}_l^\top \mathbf{W}_l\right)^{1/4}} \prod_{l=1}^{L-1} \mathrm{G}_l \left\|\mathcal{K}_{\sigma_l}\right\| \\
&\overset{(g)}{\leq} \left\|\mathcal{K}_{\mathbf{b}_L} g\right\|_{H^{s_{L-1}}(\operatorname{ran}(\mathbf{W}_L),\mathbb{R}^m)} \prod_{l=1}^{L} \mathrm{G}_l \sup_{\boldsymbol{\omega}\in\operatorname{ran}(\mathbf{W}_l)} \left|\frac{1+\left\|\mathbf{W}_l^\top\boldsymbol{\omega}\right\|_2^2}{1+\|\boldsymbol{\omega}\|_2^2}\right|^{s_{l-1}/2} \frac{1}{\det\left(\mathbf{W}_l^\top \mathbf{W}_l\right)^{1/4}} \prod_{l=1}^{L-1} \left\|\mathcal{K}_{\sigma_l}\right\|,
\end{aligned}$$

where (a) follows from the definition of f, (b) uses (10), (c) follows from the definition of G_1, (d) uses $\|\mathcal{K}_{\mathbf{b}_1}\| = 1$, (e) iterates the previous argument, (f) applies (10) at the final layer, and (g) uses the definition of f_L. Combining the above estimate with (9) yields the claimed result. ■

B.3 Proof of Theorem 3

Let $\mathcal{D}_n = \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$, and let $\boldsymbol{\sigma}_i \sim \mathrm{Rad}^m$ be independent vector-valued Rademacher variables. By definition of the empirical Rademacher complexity,

$$\widehat{\mathfrak{R}}_n^m\left(\mathcal{F}_{\mathrm{inv}}^{(\mathrm{B})}\right) = \frac{1}{n}\mathbb{E}_{\mathrm{Rad}^{mn}}\left[\sup_{f\in\mathcal{F}_{\mathrm{inv}}^{(\mathrm{B})}} \left|\sum_{i=1}^n \left\langle \boldsymbol{\sigma}_i, f\left(\mathbf{x}_i\right)\right\rangle_{\mathbb{R}^m}\right|\right], \tag{11}$$

Since $f \in \mathcal{H}_K$, the reproducing property yields

$$\widehat{\mathfrak{R}}_n^m\left(\mathcal{F}_{\text{inv}}^{(\text{B})}\right) \overset{(a)}{=} \frac{1}{n}\mathbb{E}_{\text{Rad}^{mn}}\left[\sup_{f\in\mathcal{F}_{\text{inv}}^{(\text{B})}}\left|\left\langle\sum_{i=1}^n \boldsymbol{\sigma}_i K\left(\cdot,\mathbf{x}_i\right), f\right\rangle_{\mathcal{H}_K}\right|\right]$$
$$\overset{(b)}{\leq} \frac{1}{n}\mathbb{E}_{\text{Rad}^{mn}}\left[\sup_{f\in\mathcal{F}_{\text{inv}}^{(\text{B})}}\left\|\sum_{i=1}^n \boldsymbol{\sigma}_i K\left(\cdot,\mathbf{x}_i\right)\right\|_{\mathcal{H}_K}\|f\|_{\mathcal{H}_K}\right], \tag{12}$$

where (a) follows from the reproducing property of $\mathcal{H}_K$, and (b) uses the Cauchy–Schwarz inequality in $\mathcal{H}_K$. Pulling the supremum outside the expectation gives

$$\widehat{\mathfrak{R}}_n^m\left(\mathcal{F}_{\text{inv}}^{(\text{B})}\right) \overset{(a)}{\leq} \frac{1}{n}\sup_{f\in\mathcal{F}_{\text{inv}}^{(\text{B})}}\|f\|_{\mathcal{H}_K}\,\mathbb{E}_{\text{Rad}^{mn}}\left[\left\|\sum_{i=1}^n \boldsymbol{\sigma}_i K\left(\cdot,\mathbf{x}_i\right)\right\|_{\mathcal{H}_K}\right], \tag{13}$$

where (a) follows from monotonicity of the expectation. Since $K\left(\mathbf{x},\mathbf{x}\right) = k^{(\text{B})}\left(\mathbf{x},\mathbf{x}\right)\mathbf{M}$, and

$$k^{(\text{B})}\left(\mathbf{x},\mathbf{x}\right) \leq \kappa, \qquad \mathbf{x}\in[-R,R]^d,$$

we obtain

$$\mathbb{E}_{\text{Rad}^{mn}}\left[\left\|\sum_{i=1}^n \boldsymbol{\sigma}_i K\left(\cdot,\mathbf{x}_i\right)\right\|_{\mathcal{H}_K}\right] \overset{(a)}{\leq} \left(\text{Tr}\left(\mathbf{K}_0\right)\right)^{1/2} \overset{(b)}{\leq} \sqrt{\kappa n\,\text{Tr}\left(\mathbf{M}\right)}, \tag{14}$$

where (a) follows from Jensen's inequality and the reproducing property, and (b) uses

$$\text{Tr}\left(\mathbf{K}_0\right) = \sum_{i=1}^n \text{Tr}\left(K\left(\mathbf{x}_i,\mathbf{x}_i\right)\right) \leq \kappa n\,\text{Tr}\left(\mathbf{M}\right).$$

Substituting (14) into (13) yields

$$\widehat{\mathfrak{R}}_n^m\left(\mathcal{F}_{\text{inv}}^{(\text{B})}\right) \leq \sqrt{\frac{\kappa\,\text{Tr}\left(\mathbf{M}\right)}{n}}\sup_{f\in\mathcal{F}_{\text{inv}}^{(\text{B})}}\|f\|_{\mathcal{H}_K}, \tag{15}$$

It therefore remains to estimate $\sup_{f\in\mathcal{F}_{\text{inv}}^{(\text{B})}}\|f\|_{\mathcal{H}_K}$. By definition of the hypothesis class,

$$f = \mathcal{K}_{\mathbf{W}_1}\mathcal{K}_{\mathbf{b}_1}\mathcal{K}_{\sigma_1}\cdots\mathcal{K}_{\mathbf{W}_L}\mathcal{K}_{\mathbf{b}_L}g, \qquad g\in\mathcal{H}_K.$$

By translation invariance of the Brownian Fourier seminorm, $\|\mathcal{K}_{\mathbf{b}_l}\| = 1$. Moreover, by Lemma B4,

$$\|\mathcal{K}_{\sigma_l}\| \leq C_d\,\|\sigma_l'\|_\infty\left\|\left(\sigma_l^{-1}\right)'\right\|_\infty^{1/2}, \qquad l\in[L-1], \tag{16}$$

Define

$$\Lambda_{\text{act}} \coloneqq \prod_{l=1}^{L-1}\|\sigma_l'\|_\infty\left\|\left(\sigma_l^{-1}\right)'\right\|_\infty^{1/2}. \tag{17}$$

Next, let $h\in\mathcal{H}_K$, $\text{supp}\left(h\right)\subseteq[-R,R]^d$, and let $\mathbf{W}\in\mathbb{R}^{d\times d}$ be invertible. By Lemma B3, there exist constants $0 < c_1 \leq c_2 < \infty$ such that

$$c_1\left[h\right]_{\text{B},2,\text{vv}} \leq \|h\|_{\mathcal{H}_K} \leq c_2\left[h\right]_{\text{B},2,\text{vv}}.$$

Using the scaling property of the Fourier transform and the change of variables $\boldsymbol{\omega} = \mathbf{W}^\top \boldsymbol{\zeta}$, we obtain

$$[h \circ \mathbf{W}]^2_{\mathrm{B},2,\mathrm{vv}} \leq \frac{\|\mathbf{W}\|^2}{|\det(\mathbf{W})|} [h]^2_{\mathrm{B},2,\mathrm{vv}},$$

Consequently,

$$\|\mathcal{K}_{\mathbf{W}} h\|_{\mathcal{H}_K} \overset{(a)}{=} \|h \circ \mathbf{W}\|_{\mathcal{H}_K} \overset{(b)}{\leq} c_2 [h \circ \mathbf{W}]_{\mathrm{B},2,\mathrm{vv}} \overset{(c)}{\leq} \frac{c_2}{c_1} \frac{\|\mathbf{W}\|}{|\det(\mathbf{W})|^{1/2}} \|h\|_{\mathcal{H}_K},$$

where (a) follows from the definition of $\mathcal{K}_{\mathbf{W}}$, (b) uses Lemma B3, and (c) follows from the previous estimate.

Hence,

$$\|\mathcal{K}_{\mathbf{W}}\| \overset{(a)}{\leq} C_{\mathrm{lin}} \frac{\|\mathbf{W}\|}{|\det(\mathbf{W})|^{1/2}}, \qquad C_{\mathrm{lin}} \coloneqq \frac{c_2}{c_1}, \tag{18}$$

where (a) follows from the previous inequality.

Define

$$\Lambda_{\mathrm{lin}} \coloneqq \prod_{l=1}^{L} \frac{\|\mathbf{W}_l\|}{|\det(\mathbf{W}_l)|^{1/2}}. \tag{19}$$

Combining the previous bounds yields

$$\|f\|_{\mathcal{H}_K} \overset{(a)}{\leq} \left(\prod_{l=1}^{L} \|\mathcal{K}_{\mathbf{W}_l}\| \right) \left(\prod_{l=1}^{L-1} \|\mathcal{K}_{\sigma_l}\| \right) \|g\|_{\mathcal{H}_K} \overset{(b)}{\leq} C_* \Lambda_{\mathrm{lin}} \Lambda_{\mathrm{act}} \|g\|_{\mathcal{H}_K},$$

where (a) follows from submultiplicativity of operator norms, and (b) uses (16) and (18). Taking the supremum over $\mathbf{W}_l \in \mathsf{W}(C, D)$ and substituting into (15) yields the claimed estimate. ■

B.4 Proof of Theorem 4

Define the finite-dimensional subspaces

$$\mathcal{G} \coloneqq \mathrm{Span}\{g_t : t \in [m]\} \subseteq \mathcal{H}_{\mathrm{out}},$$
$$\mathcal{V} \coloneqq \mathrm{Span}\{K(\cdot, \mathbf{x}_{ti})\, \mathbf{e}_a : t \in [m],\ i \in [n_t],\ a \in [m]\} \subseteq \mathcal{H}_K.$$

Let $P_{\mathcal{G}} : \mathcal{H}_{\mathrm{out}} \to \mathcal{G}$ and $P_{\mathcal{V}} : \mathcal{H}_K \to \mathcal{V}$ denote the orthogonal projections. Fix $T \in \mathcal{L}_2(\mathcal{H}_{\mathrm{out}}, \mathcal{H}_K)$ and define $T_0 \coloneqq P_{\mathcal{V}} T P_{\mathcal{G}}$. We first verify that T and T_0 induce identical empirical predictions. Fix $t \in [m]$, $i \in [n_t]$, and $a \in [m]$. Using the reproducing property of $\mathcal{H}_K$, we obtain

$$\begin{aligned}\langle (T_0 g_t)(\mathbf{x}_{ti}), \mathbf{e}_a \rangle_{\mathbb{R}^m} &\overset{(a)}{=} \langle T_0 g_t, K(\cdot, \mathbf{x}_{ti})\, \mathbf{e}_a \rangle_{\mathcal{H}_K} \overset{(b)}{=} \langle P_{\mathcal{V}} T P_{\mathcal{G}} g_t, K(\cdot, \mathbf{x}_{ti})\, \mathbf{e}_a \rangle_{\mathcal{H}_K} \overset{(c)}{=} \langle P_{\mathcal{V}} T g_t, K(\cdot, \mathbf{x}_{ti})\, \mathbf{e}_a \rangle_{\mathcal{H}_K} \\ &\overset{(d)}{=} \langle T g_t, K(\cdot, \mathbf{x}_{ti})\, \mathbf{e}_a \rangle_{\mathcal{H}_K} \overset{(e)}{=} \langle (T g_t)(\mathbf{x}_{ti}), \mathbf{e}_a \rangle_{\mathbb{R}^m},\end{aligned}$$

where (a) follows from the reproducing property, (b) uses the definition of T_0, (c) exploits the fact that $g_t \in \mathcal{G}$, (d) relies on $K(\cdot, \mathbf{x}_{ti})\, \mathbf{e}_a \in \mathcal{V}$ together with orthogonality of the projection, and (e) invokes the reproducing property once more. Since the previous identity holds for every coordinate $a \in [m]$, we conclude that $(T_0 g_t)(\mathbf{x}_{ti}) = (T g_t)(\mathbf{x}_{ti})$ for all training samples. Consequently,

$$\sum_{t=1}^{m} \sum_{i=1}^{n_t} \ell(y_{ti}, (T_0 g_t)(\mathbf{x}_{ti})) = \sum_{t=1}^{m} \sum_{i=1}^{n_t} \ell(y_{ti}, (T g_t)(\mathbf{x}_{ti})).$$

Since orthogonal projections are contractions and the Hilbert–Schmidt norm is an operator ideal,

$$\|T_0\|_{\mathrm{HS}} \overset{(a)}{=} \|P_{\mathcal{V}} T P_{\mathcal{G}}\|_{\mathrm{HS}} \overset{(b)}{\leq} \|P_{\mathcal{V}}\|\, \|T\|_{\mathrm{HS}}\, \|P_{\mathcal{G}}\| \overset{(c)}{\leq} \|T\|_{\mathrm{HS}},$$

where (a) comes from the definition of T_0, (b) applies the ideal property of the Hilbert–Schmidt norm, and (c) exploits $\|P_{\mathcal{V}}\| = \|P_{\mathcal{G}}\| = 1$. Therefore, T_0 achieves objective value no larger than that of T. Hence every minimum-norm minimizer may be chosen in $\mathcal{L}(\mathcal{G}, \mathcal{V})$. Since $\mathcal{V} = \operatorname{Span}\{K(\cdot, \mathbf{x}_{ti})\,\mathbf{e}_a\}$, every operator in $\mathcal{L}(\mathcal{G}, \mathcal{V})$ admits a finite expansion of the form

$$\widehat{T} = \sum_{t=1}^{m} \sum_{i=1}^{n_t} \sum_{a=1}^{m} c_{tia}\,(K(\cdot, \mathbf{x}_{ti})\,\mathbf{e}_a) \otimes g_t.$$

Finally, using the definition of rank-one operators, we obtain

$$((K(\cdot, \mathbf{x}_{ti})\,\mathbf{e}_a) \otimes g_t)\, g = (K(\cdot, \mathbf{x}_{ti})\,\mathbf{e}_a)\,\langle g_t, g\rangle_{\mathcal{H}_{\text{out}}},$$

which yields

$$\left(\widehat{T} g\right)(\cdot) = \sum_{t=1}^{m} \sum_{i=1}^{n_t} \sum_{a=1}^{m} c_{tia}\,\langle g_t, g\rangle_{\mathcal{H}_{\text{out}}}\, K(\cdot, \mathbf{x}_{ti})\,\mathbf{e}_a.$$

This completes the proof. ■

B.5 Proof of Theorem 5

By Theorem 4, a minimum-norm minimizer admits the representation

$$\widehat{T} = \sum_{s=1}^{m} \sum_{j=1}^{n_s} \sum_{a=1}^{q} c_{sja}\,(K(\cdot, \mathbf{x}_{sj})\mathbf{e}_a) \otimes g_s.$$

Therefore, for every $g \in \mathcal{H}_{\text{out}}$, we have

$$\left(\widehat{T} g\right)(\cdot) = \sum_{s=1}^{m} \sum_{j=1}^{n_s} \sum_{a=1}^{q} c_{sja}\,\langle g_s, g\rangle_{\mathcal{H}_{\text{out}}}\, K(\cdot, \mathbf{x}_{sj})\mathbf{e}_a.$$

Evaluating at $g = g_t$ and $\mathbf{x} = \mathbf{x}_{ti}$ yields

$$(\widehat{T} g_t)(\mathbf{x}_{ti}) = \sum_{s=1}^{m} \sum_{j=1}^{n_s} \sum_{a=1}^{q} c_{sja}\,\langle g_s, g_t\rangle_{\mathcal{H}_{\text{out}}}\, K(\mathbf{x}_{ti}, \mathbf{x}_{sj})\mathbf{e}_a.$$

Hence the empirical loss becomes

$$\sum_{t=1}^{m} \sum_{i=1}^{n_t} \left\| y_{ti} - (\widehat{T} g_t)(\mathbf{x}_{ti}) \right\|_2^2 = \sum_{t=1}^{m} \sum_{i=1}^{n_t} \left\| y_{ti} - \sum_{s=1}^{m} \sum_{j=1}^{n_s} \sum_{a=1}^{q} c_{sja}\,\langle g_s, g_t\rangle_{\mathcal{H}_{\text{out}}}\, K(\mathbf{x}_{ti}, \mathbf{x}_{sj})\mathbf{e}_a \right\|_2^2.$$

It remains to compute the Hilbert–Schmidt norm. For rank-one operators,

$$\langle u \otimes g, u' \otimes g' \rangle_{\text{HS}} = \langle u, u'\rangle_{\mathcal{H}_K}\,\langle g, g'\rangle_{\mathcal{H}_{\text{out}}}.$$

Using the reproducing property of $\mathcal{H}_K$,

$$\langle K(\cdot, \mathbf{x}_{sj})\mathbf{e}_a, K(\cdot, \mathbf{x}_{s'j'})\mathbf{e}_b \rangle_{\mathcal{H}_K} = \mathbf{e}_a^\top K(\mathbf{x}_{sj}, \mathbf{x}_{s'j'})\mathbf{e}_b.$$

Therefore,

$$\left\|\widehat{T}\right\|_{\text{HS}}^2 = \sum_{\substack{s,j,a \\ s',j',b}} c_{sja} c_{s'j'b}\,\langle g_s, g_{s'}\rangle_{\mathcal{H}_{\text{out}}}\, \mathbf{e}_a^\top K(\mathbf{x}_{sj}, \mathbf{x}_{s'j'})\mathbf{e}_b.$$

Substituting the finite prediction formula and this Hilbert–Schmidt norm identity into the original infinite-dimensional optimization problem yields the stated finite-dimensional objective. Conversely, every coefficient tensor c defines an operator of the displayed finite-rank form. Hence minimizing over $T \in \mathcal{L}_2(\mathcal{H}_{\text{out}}, \mathcal{H}_K)$ is equivalent to minimizing over the finite coefficient family $\{c_{sja}\}$. This completes the proof. ■

B.6 Proof of Proposition 1

The result follows from a standard Rademacher complexity argument applied to the operator-induced transfer class $\mathcal{F}_{\widehat{T}}(B)$. Using the reproducing property of $\mathcal{H}_K$, the Cauchy–Schwarz inequality, and the operator norm estimate

$$\left\|\widehat{T}g\right\|_{\mathcal{H}_K} \leq \left\|\widehat{T}\right\| \|g\|_{\mathcal{H}_{\text{out}}},$$

one obtains

$$\widehat{\mathfrak{R}}_n^m\left(\mathcal{F}_{\widehat{T}}(B)\right) \leq B\left\|\widehat{T}\right\| \sqrt{\frac{\kappa \operatorname{Tr}(\mathbf{M})}{n_0}}.$$

Combining this estimate with the standard excess-risk inequality for bounded Lipschitz losses yields the stated result. ■

C Internal Lemmas

All external lemmas used in this section are stated in Lemmas B1 to B5.

Lemma B1 (Boundedness of $\mathcal{K}_{\sigma_l}$ on vector-valued Sobolev RKHSs)**.** *Let $l \in \{1, \ldots, L-1\}$ and let $\sigma_l : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$ be bi-Lipschitz, i.e., σ_l is bijective and both σ_l and σ_l^{-1} are Lipschitz continuous. Assume further that σ_l is s_l-times differentiable and its partial derivatives satisfy*

$$\|\partial^\alpha \sigma_l\|_{L^\infty(\mathbb{R}^{d_l})} < \infty, \qquad \forall\, \alpha = (\alpha_1, \ldots, \alpha_{d_l}) \in \mathbb{N}^{d_l} \text{ with } \alpha_1 + \cdots + \alpha_{d_l} \leq s_l.$$

Suppose that $s_l \in \mathbb{N}$ and $s_l > d_l/2$, so that $H^{s_l}\left(\mathbb{R}^{d_l}, \mathbb{R}^m\right)$ is continuously embedded into $L^2\left(\mathbb{R}^{d_l}, \mathbb{R}^m\right)$. Then the Koopman operator

$$\mathcal{K}_{\sigma_l} : H^{s_l}\left(\mathbb{R}^{d_l}, \mathbb{R}^m\right) \to H^{s_l}\left(\mathbb{R}^{d_l}, \mathbb{R}^m\right), \qquad \mathcal{K}_{\sigma_l} f = f \circ \sigma_l,$$

is bounded.

Proof. Let $f = (f_1, \ldots, f_m) \in H^{s_l}\left(\mathbb{R}^{d_l}, \mathbb{R}^m\right)$. By definition,

$$\|\mathcal{K}_{\sigma_l} f\|^2_{H^{s_l}\left(\mathbb{R}^{d_l}, \mathbb{R}^m\right)} = \sum_{j=1}^m \|f_j \circ \sigma_l\|^2_{H^{s_l}\left(\mathbb{R}^{d_l}\right)}.$$

Since $s_l > d_l/2$, the scalar Sobolev norm is equivalent to

$$\|g\|^2_{H^{s_l}\left(\mathbb{R}^{d_l}\right)} = \sum_{|\alpha| \leq s_l} \|D^\alpha g\|^2_{L^2\left(\mathbb{R}^{d_l}\right)},$$

so it suffices to bound $D^\alpha(f_j \circ \sigma_l)$ in L^2 for $|\alpha| \leq s_l$. Fix $j \in \{1, \ldots, m\}$. By the multivariate Faa–di Bruno formula applied to $f_j \circ \sigma_l$, we obtain

$$D^\alpha\left(f_j \circ \sigma_l\right)(x) = \sum_{|\beta| \leq |\alpha|} D^\beta f_j\left(\sigma_l(x)\right) \sum_{i=1}^{|\alpha|} \sum_{\gamma \in p(\alpha,\beta)} \alpha! \prod_{r=1}^i \frac{\left(D^{\ell_r}\sigma_l(x)\right)^{k_r}}{k_r!\,(\ell_r!)^{|k_r|}},$$

where $p(\alpha, \beta)$ is the standard index set appearing in the Faa–di Bruno expansion. Each term is of the form

$$D^\beta f_j\left(\sigma_l(x)\right) \prod_{r=1}^i \left(D^{\gamma_r}\sigma_l(x)\right)^{\delta_r},$$

with indices bounded by $|\gamma_r|, |\delta_r| \leq |\alpha| \leq s_l$. By boundedness of all partial derivatives of σ_l up to order s_l, there exists $C > 0$ such that

$$\left| D^\beta f_j\left(\sigma_l(x)\right) \prod_{r=1}^{i} \left(D^{\gamma_r}\sigma_l(x)\right)^{\delta_r} \right| \leq C \left| D^\beta f_j\left(\sigma_l(x)\right)\right|.$$

Hence,

$$\left\|D^\alpha\left(f_j \circ \sigma_l\right)\right\|_{L^2}^2 \leq C \sum_{|\beta|\leq s_l} \int_{\mathbb{R}^{d_l}} \left|D^\beta f_j\left(\sigma_l(x)\right)\right|^2 \mathrm{d}x.$$

Since σ_l is bi-Lipschitz, the change of variables $y = \sigma_l(x)$ yields

$$\begin{aligned}\int_{\mathbb{R}^{d_l}} \left|D^\beta f_j\left(\sigma_l(x)\right)\right|^2 \mathrm{d}x &\overset{(a)}{=} \int_{\mathbb{R}^{d_l}} \left|D^\beta f_j(y)\right|^2 \left|\det D\sigma_l^{-1}(y)\right| \mathrm{d}y \overset{(b)}{\leq} \left\|\det D\sigma_l^{-1}\right\|_\infty \int_{\mathbb{R}^{d_l}} \left|D^\beta f_j(y)\right|^2 \mathrm{d}y \\ &\overset{(c)}{=} \left\|\det D\sigma_l^{-1}\right\|_\infty \left\|D^\beta f_j\right\|_{L^2\left(\mathbb{R}^{d_l}\right)}^2 \overset{(d)}{\leq} \tilde{C} \left\|D^\beta f_j\right\|_{L^2\left(\mathbb{R}^{d_l}\right)}^2,\end{aligned}$$

where (a) follows from the change of variables formula, (b) uses the L^∞ bound on the Jacobian determinant, (c) follows from the definition of the L^2 norm, and (d) uses the bi-Lipschitz property of σ_l^{-1}. for some $\tilde{C} > 0$ depending only on the Lipschitz constant of σ_l^{-1}. Combining the above bounds gives

$$\left\|f_j \circ \sigma_l\right\|_{H^{s_l}}^2 \leq C' \left\|f_j\right\|_{H^{s_l}}^2,$$

where C' depends only on σ_l and s_l. Summing over $j = 1, \ldots, m$ completes the proof:

$$\left\|\mathcal{K}_{\sigma_l} f\right\|_{H^{s_l}\left(\mathbb{R}^{d_l}, \mathbb{R}^m\right)}^2 \leq C' \left\|f\right\|_{H^{s_l}\left(\mathbb{R}^{d_l}, \mathbb{R}^m\right)}^2.$$

Thus $\mathcal{K}_{\sigma_l}$ is bounded. □

Lemma B2 (Equivalence of norms)**.** *Assume $\mathbf{M} \in \mathbb{S}_+^m$ is strictly positive definite with eigenvalues $0 < \lambda_{\min}(\mathbf{M}) \leq \lambda_{\max}(\mathbf{M}) < \infty$. Then, for all $f = (f_1, \ldots, f_m)$ with $f_j \in \mathcal{H}_k$,*

$$\lambda_{\max}(\mathbf{M})^{-1} \sum_{j=1}^m \|f_j\|_{\mathcal{H}_k}^2 \;\leq\; \|f\|_{\mathcal{H}_K}^2 \;\leq\; \lambda_{\min}(\mathbf{M})^{-1} \sum_{j=1}^m \|f_j\|_{\mathcal{H}_k}^2.$$

Proof. Let $c_{ij} = \langle f_i, f_j \rangle_{\mathcal{H}_k}$ and $\mathbf{C} = (c_{ij})_{i,j=1}^m$. Then

$$\|f\|_{\mathcal{H}_K}^2 = \sum_{i,j=1}^m \left(\mathbf{M}^{-1}\right)_{ij} c_{ij} = \mathrm{Tr}\left(\mathbf{M}^{-1}\mathbf{C}\right).$$

Since both $\mathbf{M}^{-1}$ and $\mathbf{C}$ are positive semidefinite,

$$\lambda_{\min}\left(\mathbf{M}^{-1}\right)\mathrm{Tr}\left(\mathbf{C}\right) \;\leq\; \mathrm{Tr}\left(\mathbf{M}^{-1}\mathbf{C}\right) \;\leq\; \lambda_{\max}\left(\mathbf{M}^{-1}\right)\mathrm{Tr}\left(\mathbf{C}\right).$$

Noting that $\lambda_{\min}\left(\mathbf{M}^{-1}\right) = \lambda_{\max}(\mathbf{M})^{-1}$, $\lambda_{\max}\left(\mathbf{M}^{-1}\right) = \lambda_{\min}(\mathbf{M})^{-1}$, and $\mathrm{Tr}\left(\mathbf{C}\right) = \sum_{j=1}^m \|f_j\|_{\mathcal{H}_k}^2$, we obtain the claimed bounds. □

Lemma B3 (Fourier seminorm characterization of the vector-valued Brownian RKHS)**.** *Let $\mathcal{X} = [-R, R]^d$, $R > 0$, and define the multidimensional Brownian kernel by*

$$k^{(\mathrm{B})}(\mathbf{x}, \mathbf{x}') \coloneqq \sum_{j=1}^d \frac{|x_j| + |x_j'| - |x_j - x_j'|}{2}, \qquad \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d.$$

Let $\mathbf{M} \in \mathbb{S}_+^m$ *be strictly positive definite and define the separable matrix-valued kernel*

$$K(\mathbf{x}, \mathbf{x}') \coloneqq k^{(\mathrm{B})}(\mathbf{x}, \mathbf{x}')\, \mathbf{M}, \qquad \mathbf{x}, \mathbf{x}' \in \mathcal{X}.$$

We denote by $H^{(\mathrm{B})}(\mathcal{X}, \mathbb{R}^m)$ *the associated vector-valued RKHS. For a scalar function* $g : \mathbb{R}^d \to \mathbb{R}$ *with Fourier transform* $\widehat{g}$*, define the Fourier seminorm*

$$[g]_{\mathrm{B},2}^2 \coloneqq \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|_2^2 |\widehat{g}(\boldsymbol{\omega})|^2 \,\mathrm{d}\boldsymbol{\omega}.$$

For $f = (f_1, \ldots, f_m) : \mathbb{R}^d \to \mathbb{R}^m$*, define*

$$[f]_{\mathrm{B},2,\mathrm{vv}}^2 \coloneqq \sum_{i=1}^m [f_i]_{\mathrm{B},2}^2.$$

Then there exist constants $0 < \tilde{c}_1 \le \tilde{c}_2 < \infty$*, depending only on* d*,* R*,* $\mathbf{M}$*, such that, for all* $f \in H^{(\mathrm{B})}(\mathcal{X}, \mathbb{R}^m)$ *satisfying* $\operatorname{supp}(f) \subseteq \mathcal{X}$*, we have*

$$\tilde{c}_1 [f]_{\mathrm{B},2,\mathrm{vv}} \le \|f\|_{H^{(\mathrm{B})}(\mathcal{X}, \mathbb{R}^m)} \le \tilde{c}_2 [f]_{\mathrm{B},2,\mathrm{vv}}.$$

Proof. We first consider the scalar one-dimensional case and then extend the argument to the vector-valued setting. Assume first that $d = 1$ and $m = 1$. On the interval $[-R, R]$, the scalar Brownian RKHS coincides with the Cameron–Martin space

$$H_0 \coloneqq \left\{ g : \mathbb{R} \to \mathbb{R} \;\middle|\; g(0) = 0,\ \int_{-R}^{R} |g'(t)|^2 \,\mathrm{d}t < \infty \right\},$$

equipped with the norm

$$\|g\|_{H_0}^2 \coloneqq \int_{-R}^{R} |g'(t)|^2 \,\mathrm{d}t.$$

Using a standard Sobolev extension operator on the Lipschitz domain $[-R, R]$, together with Plancherel's theorem on $\mathbb{R}$, we obtain

$$\|g\|_{H_0}^2 \overset{(a)}{=} \int_{\mathbb{R}} |\widehat{g'}(\omega)|^2 \,\mathrm{d}\omega \overset{(b)}{=} \int_{\mathbb{R}} |\omega|^2 |\widehat{g}(\omega)|^2 \,\mathrm{d}\omega \overset{(c)}{=} [g]_{\mathrm{B},2}^2,$$

where (a) follows from Plancherel's theorem, (b) uses $\widehat{g'}(\omega) = i\omega \widehat{g}(\omega)$, and (c) follows from the definition of $[g]_{\mathrm{B},2}$. Hence,

$$\|g\|_{\mathcal{H}_{k^{(\mathrm{B})}}} = [g]_{\mathrm{B},2}. \tag{20}$$

We now consider the scalar multidimensional case: $m = 1$, and $d \ge 1$. On the compact domain $[-R, R]^d$, the scalar Brownian RKHS associated with $k^{(\mathrm{B})}$ admits a norm equivalent to the first-order Sobolev seminorm. Consequently, there exist constants $0 < c_1 \le c_2 < \infty$, depending only on d and R, such that

$$c_1 \|f\|_{H^1([-R,R]^d)} \le \|f\|_{\mathcal{H}_{k^{(\mathrm{B})}}} \le c_2 \|f\|_{H^1([-R,R]^d)}. \tag{21}$$

Using a Sobolev extension operator on $[-R, R]^d$, Plancherel's theorem on $\mathbb{R}^d$ yields, for each $j \in [d]$,

$$\int_{[-R,R]^d} |\partial_{x_j} f(\mathbf{x})|^2 \,\mathrm{d}\mathbf{x} \overset{(a)}{=} \int_{\mathbb{R}^d} |\widehat{\partial_{x_j} f}(\boldsymbol{\omega})|^2 \,\mathrm{d}\boldsymbol{\omega} \overset{(b)}{=} \int_{\mathbb{R}^d} |\boldsymbol{\omega}_j|^2 |\widehat{f}(\boldsymbol{\omega})|^2 \,\mathrm{d}\boldsymbol{\omega},$$

where (a) follows from Plancherel's theorem, and (b) uses $\widehat{\partial_{x_j} f}(\boldsymbol{\omega}) = i\boldsymbol{\omega}_j \widehat{f}(\boldsymbol{\omega})$. Summing over $j = 1, \ldots, d$ gives

$$\sum_{j=1}^{d} \int_{[-R,R]^d} |\partial_{x_j} f(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x} \overset{(a)}{=} \int_{\mathbb{R}^d} \sum_{j=1}^{d} |\boldsymbol{\omega}_j|^2 |\widehat{f}(\boldsymbol{\omega})|^2 \, \mathrm{d}\boldsymbol{\omega} \overset{(b)}{=} \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|_2^2 |\widehat{f}(\boldsymbol{\omega})|^2 \, \mathrm{d}\boldsymbol{\omega} \overset{(c)}{=} [f]_{\mathrm{B},2}^2,$$

where (a) follows from linearity, (b) uses $\sum_{j=1}^{d} |\boldsymbol{\omega}_j|^2 = \|\boldsymbol{\omega}\|_2^2$, and (c) follows from the definition of $[f]_{\mathrm{B},2}$. Combining this identity with (21) yields

$$\bar{c}_1 [f]_{\mathrm{B},2} \leq \|f\|_{\mathcal{H}_{k^{(\mathrm{B})}}} \leq \bar{c}_2 [f]_{\mathrm{B},2}, \tag{22}$$

for suitable constants $0 < \bar{c}_1 \leq \bar{c}_2 < \infty$. Finally, let $f = (f_1, \ldots, f_m) : \mathbb{R}^d \to \mathbb{R}^m$ with $\mathrm{supp}(f) \subseteq [-R, R]^d$. Yielding

$$\sum_{i=1}^{m} \|f_i\|_{\mathcal{H}_{k^{(\mathrm{B})}}}^2 \overset{(a)}{\asymp} \sum_{i=1}^{m} [f_i]_{\mathrm{B},2}^2 \overset{(b)}{=} [f]_{\mathrm{B},2,\mathrm{vv}}^2,$$

where (a) follows from (22), and (b) follows from the definition of $[f]_{\mathrm{B},2,\mathrm{vv}}$. Since $K(\mathbf{x}, \mathbf{x}') = k^{(\mathrm{B})}(\mathbf{x}, \mathbf{x}')\mathbf{M}$ is separable and $\mathbf{M} \in \mathbb{S}_+^m$ is strictly positive definite, Lemma B2 implies the existence of constants $0 < \tilde{c}_1 \leq \tilde{c}_2 < \infty$ depending only on d, R, and $\mathbf{M}$ such that

$$\tilde{c}_1 [f]_{\mathrm{B},2,\mathrm{vv}} \leq \|f\|_{\mathcal{H}_K} \leq \tilde{c}_2 [f]_{\mathrm{B},2,\mathrm{vv}}.$$

This completes the proof. □

Lemma B4 (Bounds for activations in vector-valued Brownian RKHSs)**.** *Let*

$$K(\mathbf{x}, \mathbf{y}) = k^{(\mathrm{B})}(\mathbf{x}, \mathbf{y}) \, \mathbf{M}, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d,$$

be a separable matrix-valued kernel induced by the scalar Brownian kernel, where $\mathbf{M} \in \mathbb{S}_+^m$ is strictly positive definite, and let $\mathcal{H}_K$ denote the associated vector-valued RKHS. Assume that all functions under consideration satisfy $\mathrm{supp}(f) \subseteq [-R, R]^d$ *for some $R > 0$. Let $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ act componentwise:*

$$\sigma(\mathbf{x}) = (\sigma_1(x_1), \ldots, \sigma_d(x_d)), \qquad \mathbf{x} = (x_1, \ldots, x_d),$$

where each $\sigma_j : \mathbb{R} \to \mathbb{R}$ is a C^1-diffeomorphism with C^1 inverse satisfying

$$\|\sigma_j'\|_\infty < \infty, \qquad \|(\sigma_j^{-1})'\|_\infty < \infty, \qquad j \in [d].$$

Define

$$\|\sigma'\|_\infty \coloneqq \max_{j \in [d]} \|\sigma_j'\|_\infty,$$
$$\|(\sigma^{-1})'\|_\infty \coloneqq \max_{j \in [d]} \|(\sigma_j^{-1})'\|_\infty.$$

Then the Koopman operator

$$\mathcal{K}_\sigma : \mathcal{H}_K \to \mathcal{H}_K, \qquad \mathcal{K}_\sigma f = f \circ \sigma,$$

is bounded. Moreover, there exists a constant $C_{\mathrm{B}} > 0$ depending only on d, R, and $\mathbf{M}$ such that

$$\|\mathcal{K}_\sigma\| \leq C_{\mathrm{B}} \|\sigma'\|_\infty \|(\sigma^{-1})'\|_\infty^{1/2}.$$

Proof. We first consider the scalar one-dimensional case. On the interval $[-R, R]$, the Brownian RKHS coincides with the Cameron–Martin space

$$H_0 = \left\{ g : \mathbb{R} \to \mathbb{R} \;\middle|\; g(0) = 0, \; \int_{-R}^{R} |g'(t)|^2 \, \mathrm{d}t < \infty \right\},$$

equipped with the norm

$$\|g\|_{H_0}^2 = \int_{-R}^{R} |g'(t)|^2 \, \mathrm{d}t.$$

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a C^1-diffeomorphism with C^1 inverse. For $g \in H_0$, the chain rule yields

$$(g \circ \sigma)'(x) = g'(\sigma(x))\sigma'(x),$$

Consequently,

$$\|\mathcal{K}_\sigma g\|_{H_0}^2 \overset{(a)}{=} \int_{-R}^{R} |(g \circ \sigma)'(x)|^2 \, \mathrm{d}x \overset{(b)}{=} \int_{-R}^{R} |g'(\sigma(x))|^2 |\sigma'(x)|^2 \, \mathrm{d}x,$$

where (a) follows from the definition of the Cameron–Martin norm, and (b) uses the previous identity.

Applying the change of variables

$$y = \sigma(x), \qquad \mathrm{d}x = (\sigma^{-1})'(y) \, \mathrm{d}y,$$

we obtain

$$\begin{aligned}\|\mathcal{K}_\sigma g\|_{H_0}^2 &\overset{(a)}{=} \int_{\sigma([-R,R])} |g'(y)|^2 |\sigma'(\sigma^{-1}(y))|^2 |(\sigma^{-1})'(y)| \, \mathrm{d}y \overset{(b)}{\leq} \|\sigma'\|_\infty^2 \|(\sigma^{-1})'\|_\infty \int_{\sigma([-R,R])} |g'(y)|^2 \, \mathrm{d}y \\ &\overset{(c)}{\leq} \|\sigma'\|_\infty^2 \|(\sigma^{-1})'\|_\infty \|g\|_{H_0}^2,\end{aligned}$$

where (a) follows from the change of variables formula, (b) uses the uniform derivative bounds, and (c) follows since g is supported in $[-R, R]$. Hence,

$$\|\mathcal{K}_\sigma g\|_{H_0} \leq \|\sigma'\|_\infty \|(\sigma^{-1})'\|_\infty^{1/2} \|g\|_{H_0}. \tag{23}$$

We now consider the scalar multidimensional case $d \geq 1$. On compact domains, the Brownian RKHS associated with $k^{(\mathrm{B})}$ is equivalent to the first-order Sobolev seminorm:

$$\|f\|_{H^1([-R,R]^d)}^2 = \sum_{j=1}^{d} \int_{[-R,R]^d} |\partial_{x_j} f(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x}.$$

Since σ acts componentwise,

$$D\sigma(\mathbf{x}) = \mathrm{diag}\left(\sigma_1'(x_1), \ldots, \sigma_d'(x_d)\right),$$

and therefore

$$\partial_{x_j}(f \circ \sigma)(\mathbf{x}) \overset{(a)}{=} \partial_{y_j} f(\sigma(\mathbf{x}))\sigma_j'(x_j),$$

where (a) follows from the chain rule. Applying (23) coordinatewise and summing over $j \in [d]$, we obtain

$$\|f \circ \sigma\|_{H^1([-R,R]^d)} \leq C_d \|\sigma'\|_\infty \|(\sigma^{-1})'\|_\infty^{1/2} \|f\|_{H^1([-R,R]^d)}, \tag{24}$$

where $C_d > 0$ depends only on the dimension. Now let $f = (f_1, \ldots, f_m) : \mathbb{R}^d \to \mathbb{R}^m$. Using the product Sobolev norm,

$$\|f\|_{H^1([-R,R]^d, \mathbb{R}^m)}^2 = \sum_{i=1}^{m} \|f_i\|_{H^1([-R,R]^d)}^2.$$

Applying (24) componentwise yields

$$\|\mathcal{K}_\sigma f\|_{H^1([-R,R]^d,\mathbb{R}^m)} \le C_d \|\sigma'\|_\infty \|(\sigma^{-1})'\|_\infty^{1/2} \|f\|_{H^1([-R,R]^d,\mathbb{R}^m)}. \tag{25}$$

Finally, since $K(\mathbf{x},\mathbf{y}) = k^{(\mathrm{B})}(\mathbf{x},\mathbf{y})\mathbf{M}$ is separable and $\mathbf{M} \in \mathbb{S}_+^m$ is strictly positive definite, Lemma B3 implies the norm equivalence

$$c_1 [f]_{\mathrm{B},2,\mathrm{vv}} \le \|f\|_{\mathcal{H}_K} \le c_2 [f]_{\mathrm{B},2,\mathrm{vv}}.$$

Combining this estimate with (25) gives

$$\|\mathcal{K}_\sigma f\|_{\mathcal{H}_K} \overset{(a)}{\le} c_2 [\mathcal{K}_\sigma f]_{\mathrm{B},2,\mathrm{vv}} \overset{(b)}{\le} c_2 C_d \|\sigma'\|_\infty \|(\sigma^{-1})'\|_\infty^{1/2} [f]_{\mathrm{B},2,\mathrm{vv}} \overset{(c)}{\le} \frac{c_2}{c_1} C_d \|\sigma'\|_\infty \|(\sigma^{-1})'\|_\infty^{1/2} \|f\|_{\mathcal{H}_K},$$

where (a) and (c) follow from Lemma B3, and (b) uses (25). Absorbing $\frac{c_2}{c_1}$ into the constant C_d completes the proof. □

Lemma B5 (Brownian geometric factor for injective linear layers)**.** *Let $j \in [L]$, and let $\mathbf{W}_j \in \mathbb{R}^{d_j \times d_{j-1}}$ be injective. Define*

$$\mathcal{X}_j \coloneqq \mathbf{W}_j \mathcal{X}_{j-1} \subseteq \operatorname{ran}(\mathbf{W}_j),$$

where $\mathcal{X}_{j-1} \subset \mathbb{R}^{d_{j-1}}$ is a bounded Lipschitz domain, and $\operatorname{ran}(\mathbf{W}_j)$ is endowed with its induced Lebesgue measure. Assume that, for $\ell \in \{j-1, j\}$, the vector-valued Brownian RKHS $H^{(\mathrm{B})}(\mathcal{X}_\ell, \mathbb{R}^m)$ consists of restrictions to $\mathcal{X}_\ell$ of functions $f : \mathbb{R}^{d_\ell} \to \mathbb{R}^m$ satisfying $\operatorname{supp}(f) \subseteq \mathcal{X}_\ell$, and that there exist constants $0 < c_{\mathrm{low}} \le c_{\mathrm{up}} < \infty$ such that

$$c_{\mathrm{low}} \left(\|f\|_{L^2(\mathcal{X}_\ell,\mathbb{R}^m)}^2 + \|\nabla f\|_{L^2(\mathcal{X}_\ell,\mathbb{R}^m)}^2 \right) \le \|f\|_{H^{(\mathrm{B})}(\mathcal{X}_\ell,\mathbb{R}^m)}^2 \le c_{\mathrm{up}} \left(\|f\|_{L^2(\mathcal{X}_\ell,\mathbb{R}^m)}^2 + \|\nabla f\|_{L^2(\mathcal{X}_\ell,\mathbb{R}^m)}^2 \right). \tag{26}$$

Then the Koopman operator

$$\mathcal{K}_{\mathbf{W}_j} : H^{(\mathrm{B})}(\mathcal{X}_j, \mathbb{R}^m) \to H^{(\mathrm{B})}(\mathcal{X}_{j-1}, \mathbb{R}^m), \qquad \left(\mathcal{K}_{\mathbf{W}_j} f\right)(\mathbf{x}) \coloneqq f(\mathbf{W}_j \mathbf{x}),$$

is bounded. Moreover, there exists a constant $C_{\mathrm{B}} > 0$, depending only on d_{j-1}, d_j, $\mathcal{X}_{j-1}$, $\mathbf{M}$, c_{low}, and c_{up}, such that

$$\left\|\mathcal{K}_{\mathbf{W}_j}\right\| \le C_{\mathrm{B}} \frac{\left(1 + \|\mathbf{W}_j\|^2\right)^{1/2}}{\det\left(\mathbf{W}_j^\top \mathbf{W}_j\right)^{1/4}}.$$

Proof. Fix $f \in H^{(\mathrm{B})}(\mathcal{X}_j, \mathbb{R}^m)$, and define

$$g : \mathcal{X}_{j-1} \to \mathbb{R}^m, \qquad g(\mathbf{x}) \coloneqq f(\mathbf{W}_j \mathbf{x}).$$

Using the upper bound in (26), we obtain

$$\|g\|_{H^{(\mathrm{B})}(\mathcal{X}_{j-1},\mathbb{R}^m)}^2 \le c_{\mathrm{up}} \left(\|g\|_{L^2(\mathcal{X}_{j-1},\mathbb{R}^m)}^2 + \|\nabla g\|_{L^2(\mathcal{X}_{j-1},\mathbb{R}^m)}^2 \right), \tag{27}$$

Let $\mathbf{W}_j = \mathbf{Q}_j \mathbf{R}_j$ be the thin QR decomposition, where $\mathbf{Q}_j \in \mathbb{R}^{d_j \times d_{j-1}}$ has orthonormal columns and $\mathbf{R}_j \in \mathbb{R}^{d_{j-1} \times d_{j-1}}$ is invertible. Then

$$|\det(\mathbf{R}_j)| \overset{(a)}{=} \det\left(\mathbf{W}_j^\top \mathbf{W}_j\right)^{1/2}, \tag{28}$$

where (a) follows from $\mathbf{W}_j^\top \mathbf{W}_j = \mathbf{R}_j^\top \mathbf{R}_j$. We now estimate the L^2 term. Using the change of variables formula on the subspace $\operatorname{ran}(\mathbf{W}_j)$, we obtain

$$\|g\|_{L^2(\mathcal{X}_{j-1},\mathbb{R}^m)}^2 \overset{(a)}{=} \int_{\mathcal{X}_{j-1}} \|f(\mathbf{W}_j \mathbf{x})\|^2 \, \mathrm{d}\mathbf{x} \overset{(b)}{=} \frac{1}{\det\left(\mathbf{W}_j^\top \mathbf{W}_j\right)^{1/2}} \int_{\mathcal{X}_j} \|f(\mathbf{y})\|^2 \, \mathrm{d}\mathbf{y}$$

$$\overset{(c)}{=} \frac{1}{\det\left(\mathbf{W}_j^\top \mathbf{W}_j\right)^{1/2}} \|f\|^2_{L^2(\mathcal{X}_j,\mathbb{R}^m)},$$

where (a) follows from the definition of g, (b) uses the subspace change of variables formula together with (28), and (c) follows from the definition of the L^2 norm. Next, by the chain rule,

$$\nabla g(\mathbf{x}) \overset{(a)}{=} (\nabla f)(\mathbf{W}_j \mathbf{x})\, \mathbf{W}_j,$$

where (a) follows from differentiation of compositions. Therefore,

$$\begin{aligned}\|\nabla g\|^2_{L^2(\mathcal{X}_{j-1},\mathbb{R}^m)} &\overset{(a)}{\leq} \|\mathbf{W}_j\|^2 \int_{\mathcal{X}_{j-1}} \|\nabla f(\mathbf{W}_j \mathbf{x})\|^2 \,\mathrm{d}\mathbf{x} \overset{(b)}{=} \frac{\|\mathbf{W}_j\|^2}{\det\left(\mathbf{W}_j^\top \mathbf{W}_j\right)^{1/2}} \int_{\mathcal{X}_j} \|\nabla f(\mathbf{y})\|^2 \,\mathrm{d}\mathbf{y} \\ &\overset{(c)}{=} \frac{\|\mathbf{W}_j\|^2}{\det\left(\mathbf{W}_j^\top \mathbf{W}_j\right)^{1/2}} \|\nabla f\|^2_{L^2(\mathcal{X}_j,\mathbb{R}^m)},\end{aligned}$$

where (a) follows from submultiplicativity of the operator norm, (b) uses the same change of variables argument, and (c) follows from the definition of the Sobolev seminorm. Combining the previous estimates with (27) yields

$$\begin{aligned}\|g\|^2_{H^{(\mathrm{B})}(\mathcal{X}_{j-1},\mathbb{R}^m)} &\overset{(a)}{\leq} \frac{c_{\mathrm{up}}}{\det\left(\mathbf{W}_j^\top \mathbf{W}_j\right)^{1/2}} \left(\|f\|^2_{L^2(\mathcal{X}_j,\mathbb{R}^m)} + \|\mathbf{W}_j\|^2 \|\nabla f\|^2_{L^2(\mathcal{X}_j,\mathbb{R}^m)}\right) \\ &\overset{(b)}{\leq} \frac{c_{\mathrm{up}}\left(1+\|\mathbf{W}_j\|^2\right)}{\det\left(\mathbf{W}_j^\top \mathbf{W}_j\right)^{1/2}} \left(\|f\|^2_{L^2(\mathcal{X}_j,\mathbb{R}^m)} + \|\nabla f\|^2_{L^2(\mathcal{X}_j,\mathbb{R}^m)}\right),\end{aligned}$$

where (a) follows from the previous two bounds, and (b) factors out $1 + \|\mathbf{W}_j\|^2$. Finally, using the lower bound in (26), we obtain

$$\|g\|^2_{H^{(\mathrm{B})}(\mathcal{X}_{j-1},\mathbb{R}^m)} \leq \frac{c_{\mathrm{up}}}{c_{\mathrm{low}}} \frac{1+\|\mathbf{W}_j\|^2}{\det\left(\mathbf{W}_j^\top \mathbf{W}_j\right)^{1/2}} \|f\|^2_{H^{(\mathrm{B})}(\mathcal{X}_j,\mathbb{R}^m)},$$

Taking square roots yields

$$\|\mathcal{K}_{\mathbf{W}_j}\| \leq C_{\mathrm{B}} \frac{\left(1+\|\mathbf{W}_j\|^2\right)^{1/2}}{\det\left(\mathbf{W}_j^\top \mathbf{W}_j\right)^{1/4}},$$

where

$$C_{\mathrm{B}} = \left(\frac{c_{\mathrm{up}}}{c_{\mathrm{low}}}\right)^{1/2}.$$

This completes the proof. □

D Additional Experimental Details

D.1 Bound as a regularization mechanism

We assess the practical utility of the bound by incorporating it directly into the objective function as a regularization term. In particular, we add the bound to the loss and examine whether enforcing such a constraint during optimization improves stability and generalization. We employ a fully connected neural network of the form

$$f_\theta(x) \coloneqq g\left(W_4\, \sigma\left(W_3\, \sigma\left(W_2\, \sigma\left(W_1 x + b_1\right) + b_2\right) + b_3\right) + b_4\right),$$

with parameter matrices and vectors

$$W_1 \in \mathbb{R}^{1024\times 784}, \quad W_2 \in \mathbb{R}^{2048\times 1024}, \quad W_3 \in \mathbb{R}^{2048\times 2048}, \quad W_4 \in \mathbb{R}^{10\times 2048},$$
$$b_1 \in \mathbb{R}^{1024}, \quad b_2 \in \mathbb{R}^{2048}, \quad b_3 \in \mathbb{R}^{2048}, \quad b_4 \in \mathbb{R}^{10}.$$

For layers $l = 1, 2$ the matrices W_l are initialized via orthogonal initialization Saxe et al. (2014), while for layers $l = 3, 4$ they are sampled from a truncated normal distribution. The bias vectors are initialized from a uniform distribution. As activation function we use the smooth Leaky ReLU from Biswas et al. (2022). The network is trained on a subsample of 1000 points from the training set to strain its generalization capabilities and make optimization more challenging. Training is performed for 1800 epochs using the Adam optimizer Kingma & Ba (2015) with learning rate 10^{-4}, and no L^2 penalty is applied.

For regularization, we define the Sobolev Regularization as

$$\mathrm{SR} \coloneqq \sum_{l=1}^{L} \frac{\|W_l\|^{\,s_l}}{\left(\det\left(I + W_l^\top W_l\right)\right)^{1/4}}, \qquad s_l \coloneqq \frac{d_l + 0.1}{2},$$

while the Brownian Regularization is given by

$$\mathrm{BR} \coloneqq \sum_{l=1}^{L} \frac{\|W_l\|}{\left(\det\left(I + W_l^\top W_l\right)\right)^{1/4}}.$$

The regularization terms are evaluated only for layers $l = 1, 2$ and added to the loss with scaling factor $\gamma = 0.01$.

Figure 1(b) reports the test accuracy averaged over five independent random initializations. One observes that Brownian Regularization improves generalization relative to the baseline, whereas Sobolev Regularization in this setting degrades optimization and yields lower performance than the unregularized model.