Institut für Numerische Simulation
Rheinische Friedrich-Wilhelms-Universität Bonn
Friedrich-Hirzebruch-Allee 7 • 53115 Bonn • Germany
phone +49 228 73-69828 • fax +49 228 73-69847
www.ins.uni-bonn.de

B. Adcock, M. Griebel, G. Maier

Learning Lipschitz Operators with respect to Gaussian Measures with Near-Optimal Sample Complexity

INS Preprint No. 2402

October 2024

Learning Lipschitz Operators with respect to Gaussian Measures with Near-Optimal Sample Complexity

Ben Adcock[1], Michael Griebel[2,3], and Gregor Maier[2]

[1]Department of Mathematics, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada, V5A 1S6
[2]Institute for Numerical Simulation, University of Bonn, Friedrich-Hirzebruch-Allee 7, 53115 Bonn, Germany
[3]Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Abstract

Operator learning, the approximation of mappings between infinite-dimensional function spaces using ideas from machine learning, has gained increasing research attention in recent years. Approximate operators, learned from data, hold promise to serve as efficient surrogate models for problems in computational science and engineering, complementing traditional numerical methods. However, despite their empirical success, our understanding of the underpinning mathematical theory is in large part still incomplete. In this paper, we study the approximation of Lipschitz operators in expectation with respect to Gaussian measures. We prove higher Gaussian Sobolev regularity of Lipschitz operators and establish lower and upper bounds on the Hermite polynomial approximation error. We further consider the reconstruction of Lipschitz operators from m arbitrary (adaptive) linear samples. A key finding is the tight characterization of the smallest achievable error for all possible (adaptive) sampling and reconstruction maps in terms of m. It is shown that Hermite polynomial approximation is an optimal recovery strategy, but we have the following curse of sample complexity: No method to approximate Lipschitz operators based on finitely many samples can achieve algebraic convergence rates in m. On the positive side, we prove that a sufficiently fast spectral decay of the covariance operator of the Gaussian measure guarantees convergence rates which are arbitrarily close to any algebraic rate in the large data limit $m \to \infty$. Finally, we focus on the recovery of Lipschitz operators from finitely many point samples. We consider Christoffel sampling and weighted least-squares approximation, and present an algorithm which provably achieves near-optimal sample complexity.

Keywords: operator learning, high-dimensional approximation, Lipschitz operators, Gaussian measures, sample complexity, recovery

Corresponding author: maier@ins.uni-bonn.de (Gregor Maier)

1 Introduction

We study the approximation of generic Lipschitz operators which map between (infinite-)dimensional Hilbert spaces. The approximation error is measured in expectation in L^2 with input samples drawn from a Gaussian measure. We commence with a detailed literature review in Subsection 1.1, where we put our work in the context of operator learning and motivate the Gaussian setting as a natural framework for analyzing Lipschitz operators. We subsequently summarize our main contributions in Subsection 1.2 and give an overview of the organization of the remainder of the paper.

1.1 Motivation and literature review

With the rise of machine learning, in particular deep learning, in computational science and engineering (CSE), operator learning has emerged as a new paradigm for the data-driven approximation of mappings between infinite-dimensional function spaces in the past years. Multiple deep learning architectures, typically referred to as *neural operators*, such as DeepONet [56], FNO [53], non-local neural operators [43], and PCA-Net [15], have been proposed and their efficiency has been demonstrated in various practical applications. We refer to the recent reviews [42, 17] and references therein. Nevertheless, their empirical success has so far not yet been supported to large extent by a general mathematical theory. A thorough understanding of theoretical approximation guarantees, however, is important for a reliable deployment of operator learning methods in CSE applications.

A typical starting point in the theoretical analysis are universal approximation results. They guarantee the existence of a neural operator of certain type which approximates a target operator up to some arbitrarily small error [18, 51, 41, 50, 49]. Albeit being necessary for assessing the basic utility of a neural operator, mere existence results are of limited use in practical applications, where instead questions about quantitative approximation guarantees and explicit convergence rates are of greater importance.

To address the latter, two quantities are of key interest, which are based on different cost models: On the one hand, the *parametric complexity* quantifies the convergence of the approximation error in terms of the number of tunable parameters employed by the approximation method. In the context of (deep) neural network (NN) approximations, this is often referred to as *expression rates*. On the other hand, the *sample complexity* quantifies the convergence of the approximation error in terms of the number of samples used for fitting the parameters to data. Previous research efforts majorly focused on deriving expression rates for NN approximations of specific (classes of) operators whereas there has been comparably little work on sample complexity estimates.

A well-studied class of operators in the field of operator learning is the set of holomorphic operators. They arise, for example, as parameter-to-solution maps of parameterized partial differential equations (PDEs) in various contexts, such as uncertainty quantification and control problems, see, e.g., [20] and [6, Chpt. 4]. It has been shown that they can be learned with algebraic or (on finite-dimensional domains) even exponential parametric complexity with NNs [63, 37, 66]. Moreover, they can be approximated with near-optimal algebraic sample complexity with least-squares and compressed sensing methods [8, 9, 5, 12] as well as with NNs [10, 4, 3]. We mention in passing that algebraic NN expression rate estimates have also been derived for classes of (non-holomorphic) operators which arise as solution operators of certain PDEs [25]. Moreover, algebraic complexity estimates are also available for infinite-dimensional functionals (with one-dimensional codomain) with mixed regularity [28].

Another important class of operators is given by Lipschitz operators, which arise, for example, in the context of parametric elliptic variational inequalities as obstacle-to-solution operators in obstacle problems, see, e.g., [65] and [35, Chpt. 4]. Recently, it was shown in [49] that bounded Lipschitz (and C^k-Fréchet differentiable) operators cannot be approximated with algebraic parametric complexity using PCA-Net. More specifically, the number of real-valued PCA-Net parameters scales exponentially with the inverse of the approximation error. This result, termed the *curse of parametric complexity*, can be interpreted as the infinite-dimensional analogue to the classical curse of dimensionality in finite dimensions, see also [52]. It can be seemingly overcome by neural operators which use hyper-expressive activation functions or non-standard NN architectures [65]. In practical implementations, however, each real-valued parameter can only be represented by a sequence of bits of finite length. In [48], the cost model of counting real-valued parameters was therefore replaced by instead counting the number of bits used to encode each parameter to some finite accuracy. The resulting cost-accuracy scaling law reveals a curse of parametric complexity that is independent of the activation functions used in any NN approximation. It states that the number of bits required to encode each NN parameter still scales exponentially with the inverse of the approximation error. Based on the theory of widths, it was shown in [40] that bounded Lipschitz and C^k-operators also exhibit a *curse of data complexity*. That is, the error in expectation with respect to a Gaussian measure with at most algebraically decaying PCA eigenvalues converges at most logarithmically in the number of samples for any learning algorithm which is based on i.i.d. pointwise samples. In the present paper, we prove a *curse of sample complexity* which generalizes this result to arbitrary (centered, nondegenerate) Gaussian measures. In addition, we prove that for Gaussian measures with sufficiently fast spectral decay (of the covariance operator), convergence rates which are arbitrarily close to any algebraic rate are possible, even for unbounded Lipschitz operators. We mention in passing the work [55] for further results in the statistical theory of deep non-parametric estimation of Lipschitz operators. Therein, however, the authors work with probability measures with compact support. Consequently, their results are not directly applicable to Gaussian measures.

Gaussian measures are not only the typical choice of probability measure for measuring the approximation error in expectation, they also allow to draw on results from infinite-dimensional analysis [16, 23, 24, 57]. In fact, the theory of Gaussian Sobolev spaces is key in our analysis as it is well-known that Lipschitz functionals are Gaussian Sobolev functionals. This connection yields explicit control over approximation bounds in terms of the spectral properties of the covariance operator of the underlying Gaussian measure. The Gaussian setting has been considered previously to prove expression rates for NN approximations of operators, see, e.g. [66, 29]. It can also be studied within the abstract framework developed in [34], see Example 1 therein, to derive dimension-independent results for the approximation of high- and infinite-dimensional function(al)s. Its connection to Lipschitz regularity, however, has, to the best of our knowledge, not yet been made use of to derive sample complexity estimates for Lipschitz operators.

1.2 Contributions

Let $\mathcal{X}, \mathcal{Y}$ be separable Hilbert spaces with $\dim(\mathcal{X}) = \infty$ and let μ be a Gaussian measure on $\mathcal{X}$. Detailed notation and further preliminaries are introduced in Section 2. Additional notions and technical results from operator theory and infinite-dimensional analysis are discussed in the appendix. We now give an (informal) overview of our main contributions, which are are four-fold:

1. In Section 3, we extend standard results from infinite-dimensional analysis and define the (weighted) Gaussian Sobolev space $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ by means of a sequence of positive real-valued weights $\boldsymbol{b} = (b_i)_{i\in\mathbb{N}}$ with $0 < b_i \leq 1$. As our first main contribution we show that this space contains the set $\mathrm{Lip}(\mathcal{X},\mathcal{Y})$ of Lipschitz operators which map from $\mathcal{X}$ to $\mathcal{Y}$:

Result 1 (Lipschitz operators are Gaussian Sobolev operators, cf. Thm. 3.9)**.** *If $\mathcal{Y}$ is finite-dimensional, then* $\mathrm{Lip}(\mathcal{X},\mathcal{Y}) \subset W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$*. If $\mathcal{Y}$ is infinite-dimensional and if $\boldsymbol{b} \in \ell^2(\mathbb{N})$, then* $\mathrm{Lip}(\mathcal{X},\mathcal{Y}) \subset W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$*. In both cases, the space of bounded Lipschitz operators is continuously embedded in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$.*

The sequence $\boldsymbol{b}$ is essential to treat the case $\dim(\mathcal{Y}) = \infty$ and it can be interpreted as a sequence of parameters which control the degree of (weak) differentiability of the Sobolev operators. Result 1 is crucial in our subsequent analysis. Elements in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ are characterized as operators whose polynomial expansion coefficients with respect to the (infinite-dimensional) Hermite polynomials $\{H_{\boldsymbol{\gamma}}\}_{\boldsymbol{\gamma}\in\Gamma}$, with countable index set Γ, are weighted ℓ^2-summable. The corresponding weights $\boldsymbol{u} = (u_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma}\in\Gamma}$ are given in terms of the ($\boldsymbol{b}$-weighted) PCA eigenvalues $\lambda_{\boldsymbol{b},i}$ (of the covariance operator) of μ. As a result, we can study the approximation of a Lipschitz operator by considering its Hermite polynomial s-term expansions.

2. In Section 4, we give upper and lower bounds for the convergence of these expansions in terms of the PCA eigenvalues. In particular, we show the following *curse of parametric complexity*: No s-term Hermite polynomial expansion can converge with an algebraic rate uniformly for all Lipschitz operators as $s \to \infty$. This holds regardless of the decay rate of the eigenvalues. More specifically, let $S \subset \Gamma$ be a finite index set with at most s elements and let F_S denote the polynomial approximation of an operator $F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ by Hermite polynomials $H_{\boldsymbol{\gamma}}$ with $\boldsymbol{\gamma} \in S$. Moreover, let $\pi : \mathbb{N} \to \Gamma$ be a bijection such that $(u_{\boldsymbol{\pi(i)}})_{i\in\mathbb{N}}$ is a nonincreasing rearrangement of $\boldsymbol{u}$. Our second main contribution is the following result:

Result 2 (Curse of parametric complexity, cf. Theorem 4.1, Theorem 4.6, Theorem 4.7)**.** *For every $s \in \mathbb{N}$,*

$$\inf_{S\subset\Gamma, |S|\leq s} \sup_{\|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})}\leq 1} \|F - F_S\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} = \sup_{\|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})}\leq 1} \left\|F - F_{\{\boldsymbol{\pi(1)},\ldots,\boldsymbol{\pi(s)}\}}\right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} = u_{\boldsymbol{\pi(s+1)}}, \quad (1.1)$$

where the suprema are taken either over the class of all Lipschitz operators or all $W^{1,2}_{\mu,\boldsymbol{b}}$-operators. Moreover, $u_{\boldsymbol{\pi}(\boldsymbol{s}+\boldsymbol{1})}$ cannot decay algebraically fast as $s \to \infty$, regardless of the spectral properties of μ. On the positive side, the decay of $u_{\boldsymbol{\pi}(\boldsymbol{s}+\boldsymbol{1})}$ can become arbitrarily close to any algebraic rate as $s \to \infty$ if the PCA eigenvalues $\lambda_{\boldsymbol{b},i}$ decay sufficiently fast (e.g. double-exponentially).

3. Up to this point, we study best polynomial approximation of Lipschitz operators. In Section 5, we consider learning Lipschitz and $W^{1,2}_{\mu,\boldsymbol{b}}$-operators from finite samples. Using tools from information-based complexity [61], we define the adaptive m-width $\Theta_m(\mathcal{K})$ of a set of operators $\mathcal{K}$. It quantifies the best worst-case error that can be achieved by uniformly learning operators in $\mathcal{K}$ from m measurements $\mathcal{L}(F) \in \mathcal{Y}^m$ which are generated by an adaptive sampling operator $\mathcal{L}$ and used as inputs for an arbitrary (nonlinear) reconstruction map $\mathcal{T} : \mathcal{Y}^m \to L^2_\mu(\mathcal{X};\mathcal{Y})$. More specifically,

$$\Theta_m(\mathcal{K}) := \inf \left\{ \sup_{F \in \mathcal{K}} \|F - \mathcal{T}(\mathcal{L}(F))\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} : \mathcal{L}, \mathcal{T} \text{ as above} \right\}.$$

Our key result is its characterization in terms of the weights $u_{\boldsymbol{\gamma}}$ and roughly reads as follows (neglecting some technical assumptions):

Result 3 (Characterization of the adaptive m-width, cf. Thm. 5.4, Thm. 4.6, Thm. 4.7)**.** *For any number of samples $m \in \mathbb{N}$, we have*

$$\Theta_m(\mathcal{K}) = u_{\boldsymbol{\pi}(\boldsymbol{m}+\boldsymbol{1})},$$

where $\mathcal{K}$ is either the set of all Lipschitz or all $W^{1,2}_{\mu,\boldsymbol{b}}$-operators of at most unit $W^{1,2}_{\mu,\boldsymbol{b}}$-norm. Again, $u_{\boldsymbol{\pi}(\boldsymbol{m}+\boldsymbol{1})}$ cannot decay algebraically fast as $m \to \infty$, regardless of the spectral properties of μ. But its decay can become arbitrarily close to any algebraic rate in the large data limit $m \to \infty$ if the PCA eigenvalues $\lambda_{\boldsymbol{b},i}$ of μ decay sufficiently fast (e.g. double-exponentially).

This result tightly characterizes the sample complexity of learning Lipschitz operators and gives rise to the following *curse of sample complexity*: No procedure (e.g., NNs, polynomial approximation, random features, kernel methods, etc.) for uniformly learning Lipschitz operators with Sobolev norm at most one can achieve algebraic convergence rates for the L^2_μ-approximation error. This holds for general (centered, nondegenerate) Gaussian measures μ. In light of Result 3, note that Result 2 shows that Hermite polynomial approximation of Lipschitz or $W^{1,2}_{\mu,\boldsymbol{b}}$-operators (with Sobolev norm at most one) is optimal among all possible (adaptive) sampling and reconstruction maps.

4. Result 3 pertains to arbitrary linear samples and does not provide explicit algorithms. To address this lack of constructiveness, we restrict the class of sampling maps in Section 6 and study the reconstruction of operators $F : \mathcal{X} \to \mathcal{Y}$ from pointwise samples. In this case, each training datum has the form $(X, F(X)) \in \mathcal{X} \times \mathcal{Y}$. Pointwise sampling is one of the most relevant sampling methods in practical applications because it is nonintrusive. Using our results from Section 4 and Section 5, we present a sampling strategy and an algorithm for constructing for any Lipschitz and $W^{1,2}_{\mu,\boldsymbol{b}}$-operator F a weighted least-squares approximant $\widehat{F}$ with provable near-optimal sample complexity. More specifically, we use Christoffel sampling [21] to define a suitable sampling measure ν on $\mathcal{X}$ from which we draw m independent samples in order to construct $\widehat{F}$ as a linear combination of at most s Hermite polynomials. We prove the following two results which provide sample complexity bounds in probability:

Result 4 (Sample complexity for Sobolev operators, cf. Cor. 6.5)**.** *Let $0 < \epsilon < 1$ denote the failure probability and let $F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$. Suppose that m satisfies*

$$m \geq cs \log(s/\epsilon), \tag{1.2}$$

where $c > 0$ is a universal constant. Then, with ν-probability at least $1 - \epsilon$, $\widehat{F}$ is well-defined and

$$\left\|F - \widehat{F}\right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \leq \left(1 + \frac{2\sqrt{2}}{\sqrt{\epsilon}}\right) u_{\boldsymbol{\pi}(\boldsymbol{s}+\boldsymbol{1})} \|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})}. \tag{1.3}$$

The sample complexity in (1.2) is near-optimal, i.e., it is linear in s up to a log-factor. We present an algorithm (Algorithm 1) which achieves for m samples the optimal approximation error $u_{\boldsymbol{\pi}(\boldsymbol{s}+\mathbf{1})}$ (see Result 3) up to constants, where s can be chosen near-optimally as $m/(c \cdot \log(m/\epsilon))$. Note that the constant in (1.3) has poor scaling in ϵ. The next result shows that this can be avoided in the case of Lipschitz operators as follows:

Result 5 (Sample complexity for Lipschitz operators, cf. Thm. 6.6)**.** *Let $0 < \epsilon < 1$ denote the failure probability and let $F : \mathcal{X} \to \mathcal{Y}$ be Lipschitz continuous with Lipschitz constant $L > 0$. Suppose that m satisfies*

$$m \geq Csu^{-2}_{\boldsymbol{\pi}(\boldsymbol{s}+\mathbf{1})} \log(4s/\epsilon), \tag{1.4}$$

where $C > 0$ is a constant which only depends on $\|F(0)\|_{\mathcal{Y}}$ and L. Then, with ν-probability at least $1 - \epsilon$, we have

$$\left\| F - \widehat{F} \right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \leq \sqrt{2} u_{\boldsymbol{\pi}(\boldsymbol{s}+\mathbf{1})} \left(\|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} + 1 \right). \tag{1.5}$$

The bound on the approximation error in (1.5) is independent of ϵ, but the sample complexity in (1.4) shows the additional factor $u^{-2}_{\boldsymbol{\pi}(\boldsymbol{s}+\mathbf{1})}$. By Theorem 4.6, the latter grows only subalgebraically with respect to s. In this sense, the resulting algorithm (Algorithm 2) still has provable near-optimal sample complexity.

2 Preliminaries

First, we recall some standard notions and fix the notation which we use throughout the text. Further notation will be introduced in the text as needed.

2.1 Basic notation

As usual, $\mathbb{N}$ denotes the set of all positive integers, $\mathbb{N}_0$ the set of all nonnegative integers, and $\mathbb{R}$ the real numbers. We denote by $(\mathcal{X}, \langle \cdot, \cdot \rangle_{\mathcal{X}})$ and $(\mathcal{Y}, \langle \cdot, \cdot \rangle_{\mathcal{Y}})$ two separable Hilbert spaces with corresponding inner products. For simplicity, we focus on real Hilbert spaces, but all results can be readily generalized to the complex case as well. We use capital letters $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ (at times also H, K, Z) for elements of the respective Hilbert spaces. Operators which map from $\mathcal{X}$ to $\mathcal{Y}$ are typically denoted by the capital letters F or G. For functionals, i.e., in the case $\mathcal{Y} = \mathbb{R}$, we also use lower case letters at times. The set of all continuous operators from $\mathcal{X}$ to $\mathcal{Y}$ is denoted by $C(\mathcal{X}, \mathcal{Y})$ and we write $C(\mathcal{X}) := C(\mathcal{X}, \mathbb{R})$.

We equip the space $\mathcal{X}$ with a centered, nondegenerate Gaussian measure μ with covariance operator $Q := \int_{\mathcal{X}} X \otimes X d\mu(X)$. We recall that $Q : \mathcal{X} \to \mathcal{X}$ is a positive-definite, self-adjoint, trace-class operator, see, e.g., [23, Prop. 1.8]. As such, there exists an orthonormal eigenbasis (PCA basis) $\{\phi_i\}_{i \in \mathbb{N}}$ of $\mathcal{X}$ and a sequence of corresponding PCA eigenvalues $\boldsymbol{\lambda} = (\lambda_i)_{i \in \mathbb{N}}$. To be explicit, we have $Q\phi_i = \lambda_i \phi_i$ with $\lambda_i > 0$ for every $i \in \mathbb{N}$ and $\sum_{i \in \mathbb{N}} \lambda_i < \infty$. By rescaling, we may assume without loss of generality that $\sum_{i \in \mathbb{N}} \lambda_i = 1$. We denote the standard Gaussian measure on $\mathbb{R}$ by $\mu_1 := \mathcal{N}(0, 1)$, the standard Gaussian measures on $\mathbb{R}^n$, $n \in \mathbb{N}$, by $\mu_n := \bigotimes_{i=1}^n \mu_1$, and the standard Gaussian measure on the space of sequences $\mathbb{R}^{\mathbb{N}}$ by $\mu_\infty := \bigotimes_{i=1}^\infty \mu_1$.

For $N \in \mathbb{N}$, we set $[N] := \{1, 2, \ldots, N\}$ and $[\infty] := \mathbb{N}$. Sequences of real numbers with (possibly finite) index set I are denoted by lower case bold letters $\boldsymbol{x} = (x_i)_{i \in I} \in \mathbb{R}^I$. Sequences with elements in a Hilbert space $\mathcal{Z}$ are denoted by bold capital letters $\boldsymbol{Z} = (Z_i)_{i \in I} \in \mathcal{Z}^I$. We write $\mathbf{0}$ and $\mathbf{1}$ for the sequence of all zeros and all ones, respectively. Algebraic operations on a sequence $\boldsymbol{x} \in \mathbb{R}^I$ are defined componentwise: We write $\sqrt{\boldsymbol{x}} := (\sqrt{x_i})_{i \in I}$ and $1/\boldsymbol{x} := (1/x_i)_{i \in I}$, whenever these expressions make sense, and for a scalar $c \in \mathbb{R}$, we write $c\boldsymbol{x} := (cx_i)_{i \in I}$ for the scaled sequence. In a similar vein, inequalities of the form $\boldsymbol{x} \leq \boldsymbol{y}$ between sequences $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^I$ are understood componentwise, that is, $x_i \leq y_i$ for every $i \in I$. The expressions $\boldsymbol{x} \geq \boldsymbol{y}$, $\boldsymbol{x} < \boldsymbol{y}$, and $\boldsymbol{x} > \boldsymbol{y}$ are understood in a similar sense. Given an index subset $J \subset I$, we write $\boldsymbol{x}_J$ for the subsequence $(x_i)_{i \in J}$. We use standard notation

$$\delta_{i,j} := \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases}$$

for the Dirac delta function, and for two sequences $\boldsymbol{\gamma}, \boldsymbol{\gamma'} \in \mathbb{N}_0^{\mathbb{N}}$, we set

$$\delta_{\boldsymbol{\gamma},\boldsymbol{\gamma'}} := \prod_{i=1}^{\infty} \delta_{\gamma_i,\gamma_i'}.$$

We denote the Euclidean norm and inner product on $\mathbb{R}^n$ by $\|\cdot\|_{\mathbb{R}^n}$ and $\langle \cdot, \cdot \rangle_{\mathbb{R}^n}$, respectively, and the standard basis vectors by $\boldsymbol{e_i} := (\delta_{i,j})_{j\in[n]}$ for $i \in [n]$. The symbol $\mathcal{L}^n$ denotes the Lebesgue measure on $\mathbb{R}^n$. Finally, we use the notation $x \lesssim y$ for $x, y \in \mathbb{R}$ if there exists a global constant $C > 0$ independent of any parameters such that $x \leq Cy$. We write $x \gtrsim y$ if $y \lesssim x$, and $x \sim y$ if both $x \lesssim y$ and $x \gtrsim y$.

2.2 Sequence spaces

Let $1 \leq p \leq \infty$. Given an index set I, a sequence of positive weights $\boldsymbol{w} = (w_i)_{i\in I} > \boldsymbol{0}$, and a Hilbert space $\mathcal{Z}$, we define the weighted sequence space $\ell^p_{\boldsymbol{w}}(I;\mathcal{Z})$ as the set of all $\mathcal{Z}$-valued sequences $\boldsymbol{Z} = (Z_i)_{i\in I}$ whose norm $\|\boldsymbol{Z}\|_{\ell^p_{\boldsymbol{w}}(I;\mathcal{Z})}$ is finite, where

$$\|\boldsymbol{Z}\|_{\ell^p_{\boldsymbol{w}}(I;\mathcal{Z})} := \begin{cases} \left(\sum_{i\in I} w_i^{-p} \|Z_i\|_{\mathcal{Z}}^p\right)^{1/p} & \text{if } 1 \leq p < \infty, \\ \sup_{i\in I} \left\{ w_i^{-1} \|Z_i\|_{\mathcal{Z}} \right\} & \text{if } p = \infty. \end{cases}$$

We denote the closed unit ball in $\ell^p_{\boldsymbol{w}}(I;\mathcal{Z})$ by

$$B^p_{\boldsymbol{w}}(I;\mathcal{Z}) := \left\{ \boldsymbol{x} \in \ell^p_{\boldsymbol{w}}(I;\mathcal{Z}) : \|\boldsymbol{x}\|_{\ell^p_{\boldsymbol{w}}(I;\mathcal{Z})} \leq 1 \right\}.$$

If $\mathcal{Z} = \mathbb{R}$, we just write $(\ell^p_{\boldsymbol{w}}(I), \|\cdot\|_{\ell^p_{\boldsymbol{w}}(I)})$ and $B^p_{\boldsymbol{w}}(I)$, and if $\boldsymbol{w} = \boldsymbol{1}$, we write $(\ell^p(I;\mathcal{Z}), \|\cdot\|_{\ell^p(I;\mathcal{Z})})$ and $B^p(I;\mathcal{Z})$.

2.3 The weighted space $\mathcal{X}_{\boldsymbol{b}}$

Let $\boldsymbol{b} = (b_i)_{i\in\mathbb{N}}$ be a sequence of positive weights with $\boldsymbol{0} < \boldsymbol{b} \leq \boldsymbol{1}$. By means of the PCA basis $\{\phi_i\}_{i\in\mathbb{N}}$ of $\mathcal{X}$ we define the space

$$\mathcal{X}_{\boldsymbol{b}} := \left\{ X \in \mathcal{X} : \sum_{i\in\mathbb{N}} b_i^{-2} \left|\langle X, \phi_i \rangle_{\mathcal{X}}\right|^2 < \infty \right\}.$$

Note that $\mathcal{X}_{\boldsymbol{b}}$ is a Hilbert subspace of $\mathcal{X}$ with inner product

$$\langle X, Z \rangle_{\mathcal{X}_{\boldsymbol{b}}} := \sum_{i\in\mathbb{N}} b_i^{-2} \langle X, \phi_i \rangle_{\mathcal{X}} \langle Z, \phi_i \rangle_{\mathcal{X}}, \quad X, Z \in \mathcal{X}_{\boldsymbol{b}},$$

which induces the norm

$$\|X\|_{\mathcal{X}_{\boldsymbol{b}}} := \sqrt{\langle X, X \rangle_{\mathcal{X}_{\boldsymbol{b}}}}, \quad X \in \mathcal{X}_{\boldsymbol{b}}.$$

Moreover, it is easy to see that the family of vectors $\{\eta_i\}_{i\in\mathbb{N}}$ defined by

$$\eta_i := b_i \phi_i, \quad i \in \mathbb{N}, \tag{2.1}$$

is an orthonormal basis of $\mathcal{X}_{\boldsymbol{b}}$.

Remark 2.1. We highlight two important cases. If $\boldsymbol{b} = \boldsymbol{1}$, we recover the full space $\mathcal{X}$, that is, $\mathcal{X}_{\boldsymbol{1}} = \mathcal{X}$. If $\boldsymbol{b} = \sqrt{\boldsymbol{\lambda}}$, we obtain the Cameron-Martin space $\mathcal{H} \subset \mathcal{X}$ of μ (see Appendix C.1). Indeed, Theorem C.4 implies $\mathcal{X}_{\sqrt{\boldsymbol{\lambda}}} = \mathcal{H}$.

2.4 Lebesgue-Bochner spaces and Hermite polynomials

We write $L^2_\mu(\mathcal{X};\mathcal{Y})$ for the Lebesgue-Bochner space of (equivalence classes of) strongly measurable operators $F:\mathcal{X}\to\mathcal{Y}$ with finite Bochner norm

$$\|F\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} := \left(\int_{\mathcal{X}} \|F(X)\|^2_{\mathcal{Y}} d\mu(X)\right)^{1/2}.$$

If $\mathcal{Y}=\mathbb{R}$, the Lebesgue-Bochner space $L^2_\mu(\mathcal{X};\mathbb{R})$ coincides with the usual Lebesgue space and we write $L^2_\mu(\mathcal{X};\mathbb{R}) = L^2_\mu(\mathcal{X})$. More information can be found, e.g., in [38, Chpt. 1].

Next, we introduce the (infinite-dimensional) Hermite polynomials. For $n\in\mathbb{N}_0$, we define the nth normalized (probabilist's) Hermite polynomial on $\mathbb{R}$ by

$$H_n:\mathbb{R}\to\mathbb{R},\quad H_n(x) := \frac{(-1)^n}{\sqrt{n!}}\exp\left(\frac{x^2}{2}\right)\frac{d^n}{dx^n}\exp\left(-\frac{x^2}{2}\right).$$

The Hermite polynomials $\{H_n\}_{n\in\mathbb{N}}$ form an orthonormal basis of $L^2_{\mu_1}(\mathbb{R})$, see, e.g., [23, Prop. 9.4]. We define the infinite-dimensional Hermite polynomials by products of the one-dimensional ones. To this end, we introduce the set of all sequences of nonnegative integers with finite support

$$\Gamma := \left\{\boldsymbol{\gamma}\in\mathbb{N}_0^{\mathbb{N}} : \operatorname{supp}(\boldsymbol{\gamma})<\infty\right\},$$

with the support of $\boldsymbol{\gamma}$ defined by $\operatorname{supp}(\boldsymbol{\gamma}) := \{i\in\mathbb{N} : \gamma_i\neq 0\}$. It is easy to see that Γ is countable. For $\boldsymbol{\gamma}\in\Gamma$ and $d\in\mathbb{N}$, we set

$$H_{\boldsymbol{\gamma},d}:\mathbb{R}^d\to\mathbb{R},\quad H_{\boldsymbol{\gamma},d}(\boldsymbol{x}) := \prod_{i=1}^{d} H_{\gamma_i}(x_i). \tag{2.2}$$

Remark 2.2. Since $\boldsymbol{\gamma}\in\Gamma$ has finite support and $H_0=1$, each Hermite polynomial $H_{\boldsymbol{\gamma},d}$ can also be seen as a function $H_{\boldsymbol{\gamma},\infty}$ with infinite dimensional input $\boldsymbol{x}\in\mathbb{R}^{\mathbb{N}}$ (simply by ignoring all x_i with $i\notin\operatorname{supp}(\boldsymbol{\gamma})$).

Finally, we define Hermite polynomials on the infinite-dimensional space $\mathcal{X}$ by means of the PCA basis $\{\phi_i\}_{i\in\mathbb{N}}$ and the PCA eigenvalues $\boldsymbol{\lambda}=(\lambda_i)_{i\in\mathbb{N}}$:

$$H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}:\mathcal{X}\to\mathbb{R},\quad H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(X) := \prod_{i=1}^{\infty} H_{\gamma_i}\left(\frac{\langle X,\phi_i\rangle_{\mathcal{X}}}{\sqrt{\lambda_i}}\right). \tag{2.3}$$

As only finitely many factors in (2.3) are different from 1, every $H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}$ is a smooth function on $\mathcal{X}$ with polynomial growth at infinity, that is,

$$|H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(X)| \le C\left(1+\|X\|_{\mathcal{X}}^{\|\boldsymbol{\gamma}\|_{\ell^1(\mathbb{N})}}\right)$$

for some constant $C>0$. In particular, by the Fernique Theorem (Theorem C.1), we have $H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}\in L^2_\mu(\mathcal{X})$ for every $\boldsymbol{\gamma}\in\Gamma$. The Hermite polynomials $H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}$ play a distinguished role in $L^2_\mu(\mathcal{X})$ in the following sense:

Theorem 2.3 ([23, Thm. 9.7])**.** *The family $\{H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}\}_{\boldsymbol{\gamma}\in\Gamma}$ of infinite-dimensional Hermite polynomials is an orthonormal basis of $L^2_\mu(\mathcal{X})$.*

Let us recall that

$$L^2_\mu(\mathcal{X};\mathcal{Y}) = L^2_\mu(\mathcal{X})\otimes\mathcal{Y}$$

with Hilbertian tensor product. By Theorem 2.3, any $F\in L^2_\mu(\mathcal{X};\mathcal{Y})$ can be written as an unconditionally $L^2_\mu(\mathcal{X};\mathcal{Y})$-convergent expansion in Hermite polynomials, also called the *Wiener-Hermite Polynomial Chaos (PC) expansion*,

$$F = \sum_{\boldsymbol{\gamma}\in\Gamma} Y_{\boldsymbol{\gamma}} H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}, \tag{2.4}$$

with Wiener-Hermite PC coefficients

$$Y_{\boldsymbol{\gamma}} := \int_{\mathcal{X}} F(X) H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(X) d\mu(X) \in \mathcal{Y}.$$

Moreover, *Parseval's identity* holds, i.e.,

$$\|F\|^2_{L^2_\mu(\mathcal{X};\mathcal{Y})} = \sum_{\boldsymbol{\gamma}\in\Gamma} \|Y_{\boldsymbol{\gamma}}\|^2_{\mathcal{Y}}. \tag{2.5}$$

3 Gaussian Sobolev and Lipschitz operators

Let $\boldsymbol{b} = (b_i)_{i\in\mathbb{N}}$ be a sequence of weights with $\mathbf{0} < \boldsymbol{b} \leq \mathbf{1}$. We sketch the definition of the weighted Gaussian Sobolev space $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ and state its characterization as a weighted ℓ^2-sequence space. Details are provided in Appendix C.2. We then prove that all Lipschitz operators from $\mathcal{X}$ to $\mathcal{Y}$ lie in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ under some sufficient conditions on $\boldsymbol{b}$. The results in this section are the basis for the approximation theoretical analysis carried out in the remainder of this paper.

3.1 The space $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$

The definition of the space $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ is based on the operator $D_{\mathcal{X}_{\boldsymbol{b}}}$ which denotes the Fréchet differential operator along the space $\mathcal{X}_{\boldsymbol{b}}$ (see Appendix A). We first define $D_{\mathcal{X}_{\boldsymbol{b}}}$ as an operator which maps from a set of cylindrical boundedly differentiable operators $\mathcal{F}C^1_b(\mathcal{X},\mathcal{Y})$ to the space $L^2_\mu(\mathcal{X}; HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y}))$, where $HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y})$ denotes the space of Hilbert-Schmidt operators from $\mathcal{X}_{\boldsymbol{b}}$ to $\mathcal{Y}$ (see Appendix B.2). The study of the Cameron-Martin space $\mathcal{H}$ of μ allows us to show that $D_{\mathcal{X}_{\boldsymbol{b}}}$ is closable in $L^2_\mu(\mathcal{X};\mathcal{Y})$, see Proposition C.7. Details about the closure and closability of operators are recalled in Appendix B.1. With this in hand, we make the following definition:

Definition 3.1 (The Sobolev space $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$)**.** We define the space $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ as the domain of the closure of the operator $D_{\mathcal{X}_{\boldsymbol{b}}} : \mathcal{F}C^1_b(\mathcal{X},\mathcal{Y}) \to L^2_\mu(\mathcal{X}; HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y}))$ (still denoted by $D_{\mathcal{X}_{\boldsymbol{b}}}$) in $L^2_\mu(\mathcal{X};\mathcal{Y})$.

The space $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ is a Hilbert space with the graph norm

$$\|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} := \left(\int_{\mathcal{X}} \|F(X)\|^2_{\mathcal{Y}} d\mu(X) + \int_{\mathcal{X}} \|D_{\mathcal{X}_{\boldsymbol{b}}}F(X)\|^2_{HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y})} d\mu(X)\right)^{1/2},$$

which is induced by the inner product

$$\langle F,G\rangle_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} := \int_{\mathcal{X}} \langle F(X),G(X)\rangle_{\mathcal{Y}}\, d\mu(X) + \int_{\mathcal{X}} \langle D_{\mathcal{X}_{\boldsymbol{b}}}F(X), D_{\mathcal{X}_{\boldsymbol{b}}}G(X)\rangle_{HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y})}\, d\mu(X).$$

As usual, we write $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X}) := W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathbb{R})$ for the space of Sobolev functionals. A few remarks are in order.

Remark 3.2. Defining Gaussian Sobolev spaces as the domain of closure of a suitable differential operator is standard and can be found, e.g., in [19, 23, 57, 62]. For an equivalent definition via the completion of $\mathcal{F}C^1_b(\mathcal{X},\mathcal{Y})$ under an appropriate Sobolev norm, we refer to [16, Chpt. 5].

Remark 3.3. Weighted Gaussian Sobolev spaces have been considered in the literature in the study of continuous and compact Sobolev embeddings [67, 22, 19] by composing the differential operator with an additional self-adjoint nonnegative operator. Recently, in [58], the authors defined weighted Gaussian Sobolev spaces of functionals on $\ell^r(\mathbb{N})$, $r \geq 1$, by weighting the partial derivatives with elements of a weight sequence $\boldsymbol{b} \in \ell^\infty(\mathbb{N})$ and remarked that their construction unifies various definitions of Gaussian Sobolev spaces via different choices of $\boldsymbol{b}$. Our definition of $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ is equivalent to the definition in [58] in the Hilbert space case $r = 2$. However, our approach via the differential operator along the space $\mathcal{X}_{\boldsymbol{b}}$ highlights the role of the latter as the underlying differential structure of $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$. For $\boldsymbol{b} = \sqrt{\boldsymbol{\lambda}}$, we obtain the same space as defined in [16, 57, 24]. For $\boldsymbol{b} = \mathbf{1}$, we obtain a smaller space as defined in [23].

Next, we use the Wiener-Hermite PC expansion (2.4) of $L^2_\mu(\mathcal{X};\mathcal{Y})$-operators to characterize the space $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ by a weighted ℓ^2-space. To this end, we need the following definition:

Definition 3.4. The *$\boldsymbol{b}$-weighted PCA eigenvalues* are given by

$$\boldsymbol{\lambda_b} = (\lambda_{\boldsymbol{b},i})_{i\in\mathbb{N}}, \quad \lambda_{\boldsymbol{b},i} := \frac{\lambda_i}{b_i^2}. \tag{3.1}$$

The following theorem is an immediate consequence of Proposition C.9, which we prove in Appendix C.2:

Theorem 3.5 (ℓ^2-characterization of $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$)**.** *The map*

$$\ell^2_{\boldsymbol{u}}(\Gamma;\mathcal{Y}) \to W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y}), \quad \boldsymbol{Y} = (Y_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma}\in\Gamma} \mapsto \sum_{\boldsymbol{\gamma}\in\Gamma} Y_{\boldsymbol{\gamma}} H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}$$

with the family of weights

$$\boldsymbol{u} = (u_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma}\in\Gamma}, \quad u_{\boldsymbol{\gamma}} = u_{\boldsymbol{\gamma}}(\boldsymbol{\lambda_b}) := \left(1 + \sum_{i\in\mathbb{N}} \frac{\gamma_i}{\lambda_{\boldsymbol{b},i}}\right)^{-1/2}, \tag{3.2}$$

is an isometric isomorphism. In particular, using the representation (2.4)*, we have*

$$\|F\|^2_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} = \sum_{\boldsymbol{\gamma}\in\Gamma} u_{\boldsymbol{\gamma}}^{-2} \left\| \int_{\mathcal{X}} F H_{\boldsymbol{\gamma},\boldsymbol{\lambda}} d\mu \right\|^2_{\mathcal{Y}}, \quad \forall F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y}).$$

The weights $u_{\boldsymbol{\gamma}}$ in (3.2) are key in our subsequent analysis. We make the following assumption:

Assumption 3.6 (Properties of $\boldsymbol{b}$)**.** *We assume that $\boldsymbol{b} = (b_i)_{i\in\mathbb{N}}$ is a sequence of positive real numbers with $\boldsymbol{0} < \boldsymbol{b} \le \boldsymbol{1}$ such that the sequence of weighted PCA eigenvalues $\boldsymbol{\lambda_b} = (\lambda_{\boldsymbol{b},i})_{i\in\mathbb{N}}$, defined in* (3.1)*, is nonincreasing. If* $\dim(\mathcal{Y}) = \infty$*, we assume in addition that $\boldsymbol{b} \in \ell^2(\mathbb{N})$.*

By Assumption 3.6, we can order the weights $u_{\boldsymbol{\gamma}}$ in a nonincreasing way, that is, there exists a *nonincreasing rearrangement* $\pi : \mathbb{N} \to \Gamma$ of $\boldsymbol{u}$ such that

$$u_{\boldsymbol{\pi(1)}} \ge u_{\boldsymbol{\pi(2)}} \ge \cdots > 0. \tag{3.3}$$

The map π is unique up to permutations of weights of the same value. The additional requirement of ℓ^2-summability of $\boldsymbol{b}$ in Assumption 3.6 in the case where $\mathcal{Y}$ is infinite-dimensional will become clear by Theorem 3.9, which we prove in Subsection 3.2. It implies that the set of Lipschitz operators is a subset of $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$.

Remark 3.7. Assumption 3.6 implies that $\limsup_{i\to\infty} \lambda_{\boldsymbol{b},i} < \infty$. Interestingly, this condition is equivalent to the continuous embedding of $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X})$ in the Orlicz space $L^p \log^{\frac{p}{2}} L(\mathcal{X},\mu)$ for $p \in [1,\infty)$, see [58, Thm. 4.2]. The stronger condition $\lim_{i\to\infty} \lambda_{\boldsymbol{b},i} = 0$ is equivalent to the compact embedding of $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X})$ in $L^2_\mu(\mathcal{X})$, see [22], and, more generally, in the Orlicz space $L^2 \log^q L(\mathcal{X},\mu)$ for $q \in [0,1)$, see [58, Thm. 5.2].

3.2 Lipschitz operators

We now turn to Lipschitz continuous operators and recall their definition.

Definition 3.8 (Lipschitz operators)**.** An operator $F : \mathcal{X} \to \mathcal{Y}$ is called *(L-)Lipschitz (continuous)* if there exists a constant $L > 0$ such that

$$\|F(X) - F(Z)\|_{\mathcal{Y}} \le L\|X - Z\|_{\mathcal{X}}, \quad \forall X, Z \in \mathcal{X}.$$

The number L is called a *Lipschitz constant* of F. The smallest Lipschitz constant of F is given by

$$[F]_{\mathrm{Lip}(\mathcal{X},\mathcal{Y})} := \sup_{\substack{X,Z\in\mathcal{X}\\ X\neq Z}} \frac{\|F(X)-F(Z)\|_{\mathcal{Y}}}{\|X-Z\|_{\mathcal{X}}}.$$

We denote the space of all Lipschitz operators from $\mathcal{X}$ to $\mathcal{Y}$ by $\mathrm{Lip}(\mathcal{X},\mathcal{Y})$ and write $\mathrm{Lip}(\mathcal{X}) := \mathrm{Lip}(\mathcal{X};\mathbb{R})$. We further define the space of all *bounded* Lipschitz operators $C^{0,1}(\mathcal{X};\mathcal{Y})$ as the set of all Lipschitz operators $F \in \mathrm{Lip}(\mathcal{X},\mathcal{Y})$ with finite norm

$$\|F\|_{C^{0,1}(\mathcal{X},\mathcal{Y})} := \sup_{X\in\mathcal{X}} \|F(X)\|_{\mathcal{Y}} + [F]_{\mathrm{Lip}(\mathcal{X},\mathcal{Y})}.$$

Note that $C^{0,1}(\mathcal{X},\mathcal{Y})$ is a strict subset of $\mathrm{Lip}(\mathcal{X},\mathcal{Y})$ as operators in $\mathrm{Lip}(\mathcal{X},\mathcal{Y})$ do not need to be bounded.

The next result, which is the main result of this section, motivates Gaussian Sobolev spaces as a natural setting for the study of Lipschitz operators. We present a sketch of the proof, highlighting the main ideas. A detailed proof is given in Appendix D.

Theorem 3.9 (Lipschitz operators are Gaussian Sobolev operators)**.** *Let $\boldsymbol{b} = (b_i)_{i\in\mathbb{N}}$ be a sequence of positive numbers with $\mathbf{0} < \boldsymbol{b} \leq \mathbf{1}$.*

(i) If $\mathcal{Y}$ is finite-dimensional, then $\mathrm{Lip}(\mathcal{X},\mathcal{Y}) \subset W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ and the embedding $C^{0,1}(\mathcal{X},\mathcal{Y}) \hookrightarrow W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ is continuous with

$$\|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} \leq \sqrt{\dim(\mathcal{Y})} \cdot \|F\|_{C^{0,1}(\mathcal{X},\mathcal{Y})}, \quad \forall F \in C^{0,1}(\mathcal{X},\mathcal{Y}).$$

(ii) If $\mathcal{Y}$ is infinite-dimensional and if $\boldsymbol{b} \in \ell^2(\mathbb{N})$, then $\mathrm{Lip}(\mathcal{X},\mathcal{Y}) \subset W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ and the embedding $C^{0,1}(\mathcal{X},\mathcal{Y}) \hookrightarrow W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ is continuous with

$$\|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} \leq \max\left\{1, \|\boldsymbol{b}\|_{\ell^2(\mathbb{N})}\right\} \|F\|_{C^{0,1}(\mathcal{X},\mathcal{Y})}, \quad \forall F \in C^{0,1}(\mathcal{X},\mathcal{Y}).$$

Proof (Sketch). Let $F \in \mathrm{Lip}(\mathcal{X},\mathcal{Y})$. In order to show that F lies in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$, it suffices to find a sequence $(F_n)_{n\in\mathbb{N}}$ of operators which converge to F in $L^2_\mu(\mathcal{X};\mathcal{Y})$ and which are uniformly bounded in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ (see Lemma C.8). To this end, we construct a specific sequence of operators of the form $F_n = V_n \circ T_n$ with $V_n : \mathbb{R}^n \to \mathcal{Y}$ and $T_n : \mathcal{X} \to \mathbb{R}^n$ which converge to F in $L^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ and, in fact, in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$. As we will show, the operator V_n inherits Lipschitz continuity of F. Hence, by Rademacher's theorem, V_n is differentiable $\mathcal{L}^n$-almost everywhere. We are now left with showing that F_n is differentiable μ-almost everywhere and that there exists a constant $C > 0$ such that

$$\|D_{\mathcal{X}_{\boldsymbol{b}}} F_n(X)\|_{HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y})} \leq C \tag{3.4}$$

for all $n \in \mathbb{N}$ and μ-almost every $X \in \mathcal{X}$.

In case (i), let $m := \dim(\mathcal{Y}) \in \mathbb{N}$. We fix an orthonormal basis $\{\psi_j\}_{j\in[m]}$ of $\mathcal{Y}$ and consider the (Lipschitz continuous) coordinate functions $V_n^{(j)} := \langle V_n, \psi_j\rangle_{\mathcal{Y}} : \mathbb{R}^n \to \mathbb{R}$. Unwinding definitions, a straight-forward calculation then leads to (3.4) for any weight sequence $\mathbf{0} < \boldsymbol{b} \leq \mathbf{1}$. The resulting constant C depends linearly on $\sqrt{\dim(\mathcal{Y})}$. The same reasoning thus does not work in case (ii) where $\mathcal{Y}$ has infinite dimension. At this point, we additionally require that $\boldsymbol{b} \in \ell^2(\mathbb{N})$. A slight modification of the argument in (i) then yields (3.4) with a finite constant C which depends linearly on $\min\{1, \|\boldsymbol{b}\|_{\ell^2(\mathbb{N})}\}$ and is independent of the dimension of $\mathcal{Y}$. The continuous embedding of $C^{0,1}(\mathcal{X},\mathcal{Y})$ in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ follows in both cases from (3.4) and the fact that $F_n \to F$ in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$. □

Remark 3.10. In light of Theorem 3.9, note that Assumption 3.6 implies that $\mathrm{Lip}(\mathcal{X},\mathcal{Y}) \subset W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ regardless of whether $\mathcal{Y}$ is finite- or infinite-dimensional. Moreover, it implies that the weighted PCA eigenvalues $\lambda_{\boldsymbol{b},i}$ decay more slowly than the unweighted λ_i. If $\mathcal{Y}$ is finite-dimensional, we can choose $\boldsymbol{b} = \mathbf{1}$, which gives $\boldsymbol{\lambda_b} = \boldsymbol{\lambda}$. This is not possible if $\mathcal{Y}$ is infinite-dimensional. However, since $\boldsymbol{\lambda} \in \ell^1(\mathbb{N})$, a valid choice for $\boldsymbol{b}$ in any case is $\boldsymbol{b} = \sqrt{\boldsymbol{\lambda}}$, which leads to $\boldsymbol{\lambda_b} = \mathbf{1}$ and thus to no decay of the $\lambda_{\boldsymbol{b},i}$ at all.

Remark 3.11. For functionals, i.e., in the case $\mathcal{Y} = \mathbb{R}$, it is well-known that Lipschitz continuity implies Gaussian Sobolev regularity. We refer to [23, Prop. 10.11] and [24, Prop. 3.18], where it is shown that $\mathrm{Lip}(\mathcal{X}) \subset W^{1,2}_{\mu,\mathbf{1}}(\mathcal{X})$ and $\mathrm{Lip}(\mathcal{X}) \subset W^{1,2}_{\mu,\sqrt{\boldsymbol{\lambda}}}(\mathcal{X})$, respectively. One can define Gaussian Sobolev spaces $W^{1,p}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ for any order $1 \le p < \infty$ and a proof analogous to the one of Theorem 3.9 shows that they contain $\mathrm{Lip}(\mathcal{X},\mathcal{Y})$ as a subset. For the case $\mathcal{Y} = \mathbb{R}$, we mention [16, Ex. 5.4.10(i)] and [57, Prop. 10.1.4]. However, only in the case $p = 2$ there is a simple characterization of the Gaussian Sobolev norm in terms of Hermite polynomial coefficients, as given by Theorem 3.5.

Finally, we introduce the following notation.

Definition 3.12 (Sobolev unit (Lipschitz) ball)**.** We define the *Sobolev unit ball*

$$B_1(W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})) := \left\{ F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y}) : \|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} \le 1 \right\}$$

and the *Sobolev unit Lipschitz ball*

$$B_1^{\boldsymbol{b}}(\mathrm{Lip}(\mathcal{X},\mathcal{Y})) := \left\{ F \in \mathrm{Lip}(\mathcal{X},\mathcal{Y}) : \|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} \le 1 \right\}.$$

4 Polynomial s-term approximation

The ℓ^2-characterization of $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ via Wiener-Hermite PC expansions (Theorem 3.5) motivates studying polynomial s-term approximations of $W^{1,2}_{\mu,\boldsymbol{b}}$-operators and quantifying the smallest achievable s-term error. To this end, for any index set $S \subset \Gamma$, we define the space of $\mathcal{Y}$-valued polynomials

$$\mathcal{P}_{S;\mathcal{Y}} := \left\{ \sum_{\boldsymbol{\gamma} \in S} Y_{\boldsymbol{\gamma}} H_{\boldsymbol{\gamma},\boldsymbol{\lambda}} : Y_{\boldsymbol{\gamma}} \in \mathcal{Y} \right\}$$

and the corresponding orthogonal L^2_μ-projection

$$(\cdot)_S : L^2_\mu(\mathcal{X};\mathcal{Y}) \to \mathcal{P}_{S;\mathcal{Y}}, \quad F \mapsto F_S := \sum_{\boldsymbol{\gamma} \in S} \left(\int_{\mathcal{X}} F H_{\boldsymbol{\gamma},\boldsymbol{\lambda}} d\mu \right) H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}.$$

Next, let $F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ and let $S \subset \Gamma$ be finite with $|S| \le s$. Setting $Y_{\boldsymbol{\gamma}} := \int_{\mathcal{X}} F H_{\boldsymbol{\gamma},\boldsymbol{\lambda}} d\mu$, it follows from Parseval's identity (2.5) and Theorem 3.5 that

$$\|F - F_S\|^2_{L^2_\mu(\mathcal{X};\mathcal{Y})} = \sum_{\boldsymbol{\gamma} \in \Gamma \setminus S} \|Y_{\boldsymbol{\gamma}}\|^2_{\mathcal{Y}} \le \left(\max_{\boldsymbol{\gamma} \in \Gamma \setminus S} u^2_{\boldsymbol{\gamma}} \right) \sum_{\boldsymbol{\gamma} \in \Gamma} u^{-2}_{\boldsymbol{\gamma}} \|Y_{\boldsymbol{\gamma}}\|^2_{\mathcal{Y}} = \left(\max_{\boldsymbol{\gamma} \in \Gamma \setminus S} u^2_{\boldsymbol{\gamma}} \right) \|F\|^2_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})}. \tag{4.1}$$

Let us recall from (3.3) the nonincreasing rearrangement $\pi : \mathbb{N} \to \Gamma$ of $\boldsymbol{u} = (u_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma} \in \Gamma}$. We set $S = \pi([s]) = \{\boldsymbol{\pi(1)}, \ldots, \boldsymbol{\pi(s)}\}$ and conclude for $\mathcal{K} \in \{B_1(W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})), B_1^{\boldsymbol{b}}(\mathrm{Lip}(\mathcal{X},\mathcal{Y}))\}$ that

$$\inf_{S \subset \Gamma, |S| \le s} \sup_{F \in \mathcal{K}} \|F - F_S\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \le \sup_{F \in \mathcal{K}} \left\| F - F_{\pi([s])} \right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \le \max_{\boldsymbol{\gamma} \in \Gamma \setminus \pi([s])} u_{\boldsymbol{\gamma}} = u_{\boldsymbol{\pi(s+1)}}, \quad \forall s \in \mathbb{N}. \tag{4.2}$$

Our first main result in this section shows that this chain of inequalities can, in fact, be improved to equality and hence gives a tight characterization of the best polynomial s-term error. The proof is an immediate consequence of Theorem 5.4, which we prove in Section 5, in the special case $\mathcal{V} = L^2_\mu(\mathcal{X};\mathcal{Y})$.

Theorem 4.1 (Best polynomial s-term error)**.** *For* $\mathcal{K} \in \{B_1(W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})),\ B_1^{\boldsymbol{b}}(\mathrm{Lip}(\mathcal{X},\mathcal{Y}))\}$ *and every* $s \in \mathbb{N}$*, we have*

$$\inf_{S\subset\Gamma,|S|\le s}\sup_{F\in\mathcal{K}}\|F-F_S\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} = \sup_{F\in\mathcal{K}}\left\|F-F_{\pi([s])}\right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} = u_{\boldsymbol{\pi(s+1)}}.$$

Motivated by Theorem 4.1, we study in the rest of this section the decay of $u_{\boldsymbol{\pi(s+1)}}$ for $s\to\infty$. The proofs are based on the relation of the set $\pi([s])$ to an anisotropic total degree index set. We discuss this relation in Subsection 4.1 and subsequently prove lower and upper bounds for $u_{\boldsymbol{\pi(s+1)}}$ in Subsections 4.2 and 4.3.

4.1 Relation to anisotropic total degree index sets

We first recall the notion of anisotropic total degree (TD) index sets and provide lower and upper size bounds. We then identify a specific anisotropic TD index set to which we can relate $\pi([s])$.

Definition 4.2 (Anisotropic TD index set)**.** For $d\in\mathbb{N}$ and $\boldsymbol{a}=(a_1,\dots,a_d)\in\mathbb{R}^d$, $\boldsymbol{a}>\boldsymbol{0}$, we define the anisotropic TD index set in d dimensions with weight $\boldsymbol{a}$ by

$$\Lambda^{\mathrm{TD}}_{d,\boldsymbol{a}} := \left\{\nu\in\mathbb{N}_0^d : \sum_{i=1}^d a_i\nu_i \le 1\right\}.$$

Lemma 4.3 (Lower and upper size bounds for anisotropic TD index sets)**.** *Let* $d\in\mathbb{N}$ *and* $\boldsymbol{a}=(a_1,\dots,a_d)\in\mathbb{R}^d$ *with* $0<a_1\le\cdots\le a_d$*. We have*

$$\left|\Lambda^{\mathrm{TD}}_{d,\boldsymbol{a}}\right| \le \prod_{i=1}^d\left(\frac{1}{a_i i}+1\right), \tag{4.3}$$

and if in addition $\min_{i\in[d]} a_i\le 1$*, then*

$$\prod_{i=1}^d\frac{1}{a_i i}\le\left|\Lambda^{\mathrm{TD}}_{d,\boldsymbol{a}}\right|. \tag{4.4}$$

Proof. The upper bound (4.3) is Lemma 5.3 in [36, Lemma 5.3]. The lower bound (4.4) is proved in [13]. We refer to [33] for further discussion. □

For any $\varepsilon>0$, we now define the set

$$S(\varepsilon) := \left\{\boldsymbol{\gamma}\in\Gamma : u_{\boldsymbol{\gamma}}^{-2}\le 1+\frac{1}{\varepsilon^2}\right\} = \left\{\boldsymbol{\gamma}\in\Gamma : \sum_{i\in\mathbb{N}}\frac{\gamma_i}{\lambda_{\boldsymbol{b},i}}\le\frac{1}{\varepsilon^2}\right\} \tag{4.5}$$

as well as the quantity

$$d(\varepsilon) := \min\left\{l\in\mathbb{N} : \lambda_{\boldsymbol{b},l+1}<\varepsilon^2\right\}\in\mathbb{N}\cup\{\infty\} \tag{4.6}$$

with the convention $\min(\emptyset)=\infty$. By definition, $S(\varepsilon)=\pi([|S(\varepsilon)|])$, and if $d(\varepsilon)<\infty$, then $S(\varepsilon)$ is isomorphic to an anisotropic TD index set,

$$S(\varepsilon)\cong\Lambda^{\mathrm{TD}}_{d(\varepsilon),\boldsymbol{a}}, \tag{4.7}$$

with weights $a_i := \varepsilon^2/\lambda_{\boldsymbol{b},i}$, $i\in[d(\varepsilon)]$, under the isomorphism

$$\{\boldsymbol{\gamma}\in\Gamma : \mathrm{supp}(\boldsymbol{\gamma})\subset[d(\varepsilon)]\}\to\mathbb{N}_0^{d(\varepsilon)},\quad \boldsymbol{\gamma}\mapsto(\gamma_1,\dots,\gamma_{d(\varepsilon)}).$$

Observe that, by Assumption 3.6, we have $0<a_1\le a_2\le\cdots\le a_{d(\varepsilon)}$.

Remark 4.4 (Effective dimension). The number $d(\varepsilon)$, defined in (4.6), can be interpreted as the *effective dimension* of the approximation problem. For $i \geq d(\varepsilon)+1$, we have $\lambda_{\boldsymbol{b},i} < \varepsilon^2$. Hence, for very small $0 < \varepsilon < 1$, the variance of μ in the ith coordinate direction essentially vanishes for $i \geq d(\varepsilon)+1$ and therefore $\mathcal{N}(0, \lambda_{\boldsymbol{b},i}) \approx \delta_0$, where $\mathcal{N}(0, \lambda_{\boldsymbol{b},i})$ denotes the one-dimensional Gaussian measure on $\mathbb{R}$ with mean 0 and variance $\lambda_{\boldsymbol{b},i}$ and δ_0 is the Dirac delta measure centered at 0. Thus, measuring an operator F on $\mathcal{X}$ with respect to μ essentially reduces to measuring F in its first $d(\varepsilon)$ coordinates (with respect to the PCA basis $\{\phi_i\}_{i\in\mathbb{N}}$) on $\mathbb{R}^{d(\varepsilon)}$ with respect to the Gaussian product measure $\bigotimes_{i=1}^{d(\varepsilon)} \mathcal{N}(0, \lambda_{\boldsymbol{b},i})$.

Remark 4.5 (Finiteness of $d(\varepsilon)$). Note that requiring $d(\varepsilon) < \infty$ for every $\varepsilon > 0$ together with Assumption 3.6 implies $\lim_{i\to\infty} \lambda_{\boldsymbol{b},i} = 0$. On the other hand, the limit condition $\lim_{i\to\infty} \lambda_{\boldsymbol{b},i} = 0$ implies Assumption 3.6 after a suitable reordering of the $\lambda_{\boldsymbol{b},i}$ and the property $d(\varepsilon) < \infty$ for every $\varepsilon > 0$. In this context, we also recall Remark 3.7.

4.2 Lower bound

We now prove the second main result of this section which constitutes a lower bound for $u_{\boldsymbol{\pi}(\boldsymbol{s}+\boldsymbol{1})}$. It states that, regardless of the (unweighted) PCA eigenvalues $\boldsymbol{\lambda}$ and choice of $\boldsymbol{b}$, one cannot achieve an algebraic decay for $s \to \infty$. In light of Remark 4.5, we emphasize that we do not assume that $\lim_{i\to\infty} \lambda_{\boldsymbol{b},i} = 0$, but only that the $\lambda_{\boldsymbol{b},i}$ are nonincreasing (Assumption 3.6). In particular, the effective dimension $d(\varepsilon)$ in (4.6) might be infinite for a given $\varepsilon > 0$.

Theorem 4.6 (Impossibility of algebraic decay of $u_{\boldsymbol{\pi}(\boldsymbol{s}+\boldsymbol{1})}$). *For any $p \in \mathbb{N}$, there exists $\bar{s} \in \mathbb{N}$, depending on $\lambda_{\boldsymbol{b},1}, \ldots, \lambda_{\boldsymbol{b},p+1}, p$, such that*

$$u_{\boldsymbol{\pi}(\boldsymbol{s}+\boldsymbol{1})} \geq C s^{-\frac{1}{2p}}, \quad \forall s \geq \bar{s}, \tag{4.8}$$

with constant

$$C = C(\lambda_{\boldsymbol{b},1}, \ldots, \lambda_{\boldsymbol{b},p}, p) := \frac{1}{2} \left(\prod_{i=1}^{p} \frac{\lambda_{\boldsymbol{b},i}}{i} \right)^{\frac{1}{2p}}. \tag{4.9}$$

Proof. We fix some arbitrary $0 < \varepsilon \leq \sqrt{\lambda_{\boldsymbol{b},1}}$, whose exact value will be chosen later, and define for $n \in \mathbb{N}$,

$$S(\varepsilon, n) := \left\{ \boldsymbol{\gamma} \in \mathbb{N}_0^{\mathbb{N}} : \sum_{i\in\mathbb{N}} \frac{\gamma_i}{\lambda_{\boldsymbol{b},i}} \leq \frac{1}{\varepsilon^2},\ \operatorname{supp}(\boldsymbol{\gamma}) \subset [n] \right\}.$$

We make a couple of simple but important observations. First note that $S(\varepsilon) = S(\varepsilon, d(\varepsilon))$, where $S(\varepsilon)$ and $d(\varepsilon)$ are defined in (4.5) and (4.6), respectively. Second, we have $S(\varepsilon, n') \subset S(\varepsilon, n)$ for every $1 \leq n' \leq n$. Third, $S(\varepsilon, n)$ is isomorphic to the anisotropic TD index set $\Lambda^{\mathrm{TD}}_{\boldsymbol{a},n}$ with weight $\boldsymbol{a} = (a_1, \ldots, a_n)$, $a_i := \varepsilon^2/\lambda_{\boldsymbol{b},i}$, under the isomorphism

$$\{\boldsymbol{\gamma} \in \mathbb{N}_0^{\mathbb{N}} : \operatorname{supp}(\boldsymbol{\gamma}) \subset [n]\} \to \mathbb{N}_0^n, \quad \boldsymbol{\gamma} \mapsto (\gamma_1, \ldots, \gamma_n).$$

Moreover, since we chose $\varepsilon \leq \sqrt{\lambda_{\boldsymbol{b},1}}$, we have $\min_{i\in[n]} a_i = \varepsilon^2/\lambda_{\boldsymbol{b},1} \leq 1$. We can thus combine the preceding observations with the lower size bound (4.4) to conclude

$$|S(\varepsilon)| = |S(\varepsilon, d(\varepsilon))| \geq |S_{\varepsilon,d'}| \geq \prod_{i=1}^{d'} \frac{\lambda_{\boldsymbol{b},i}}{\varepsilon^2 i}, \quad \forall 1 \leq d' \leq d(\varepsilon). \tag{4.10}$$

Analogous to the definition of $d(\varepsilon)$ in (4.6), we set

$$\widetilde{d}(\varepsilon) := \min \left\{ l \in \mathbb{N} : \frac{\lambda_{\boldsymbol{b},l+1}}{l+1} < \varepsilon^2 \right\} \in \mathbb{N}.$$

Note that $\widetilde{d}(\varepsilon)$ is well-defined because $\lambda_{\boldsymbol{b},i} \leq \lambda_{\boldsymbol{b},1}$ for every $i \in \mathbb{N}$. Moreover, we have $\widetilde{d}(\varepsilon) \leq d(\varepsilon)$ as well as $\widetilde{d}(\varepsilon) \to \infty$ as $\varepsilon \to 0$.

Next, let us fix some arbitrary $p \in \mathbb{N}$. Then there exists $0 < \bar{\varepsilon} = \bar{\varepsilon}(\lambda_{\boldsymbol{b},1}, \dots, \lambda_{\boldsymbol{b},p+1}, p) \leq \min\{\sqrt{\lambda_{\boldsymbol{b},1}}, 1\}$ such that $\widetilde{d}(\varepsilon) \geq p$ for every $0 < \varepsilon \leq \bar{\varepsilon}$. Since $\lambda_{\boldsymbol{b},i}/(\varepsilon^2 i) \geq 1$ for every $1 \leq i \leq \widetilde{d}(\varepsilon)$, it follows from (4.10) that

$$|S(\varepsilon)| \geq \prod_{i=1}^{\widetilde{d}(\varepsilon)} \frac{\lambda_{\boldsymbol{b},i}}{\varepsilon^2 i} \geq \prod_{i=1}^{p} \frac{\lambda_{\boldsymbol{b},i}}{\varepsilon^2 i} = \widetilde{C}\varepsilon^{-2p}, \quad \forall 0 < \varepsilon \leq \bar{\varepsilon},$$

with constant

$$\widetilde{C} = \widetilde{C}(\lambda_{\boldsymbol{b},1}, \dots, \lambda_{\boldsymbol{b},p}, p) := \prod_{i=1}^{p} \frac{\lambda_{\boldsymbol{b},i}}{i}.$$

We choose $\bar{s} = \bar{s}(\lambda_{\boldsymbol{b},1}, \dots, \lambda_{\boldsymbol{b},p+1}, p) \in \mathbb{N}$ sufficiently large such that $s + 1 \geq \lceil \widetilde{C}\bar{\varepsilon}^{-2p} \rceil$ for every $s \geq \bar{s}$. We fix some arbitrary $s \geq \bar{s}$ and pick $0 < \widetilde{\varepsilon} \leq \bar{\varepsilon}$ such that

$$\widetilde{C}\widetilde{\varepsilon}^{-2p} = s + 1.$$

Solving for $\widetilde{\varepsilon}^2$ yields

$$\widetilde{\varepsilon}^2 = \widetilde{\varepsilon}^2(\lambda_{\boldsymbol{b},1}, \dots, \lambda_{\boldsymbol{b},p}, p, s) = \widetilde{C}^{1/p}(s+1)^{-1/p}.$$

By our choice of $\widetilde{\varepsilon}$, we have $|S(\widetilde{\varepsilon})| \geq s+1$. Since $S(\widetilde{\varepsilon}) = \pi(|S(\widetilde{\varepsilon})|)$, we conclude $\boldsymbol{\pi(s+1)} \in S(\widetilde{\varepsilon})$ and therefore

$$u^2_{\boldsymbol{\pi(s+1)}} \geq \left(\frac{1}{\widetilde{\varepsilon}^2} + 1\right)^{-1} \geq \frac{1}{2}\widetilde{\varepsilon}^2 = \frac{1}{2}\widetilde{C}^{1/p}(s+1)^{-1/p} \geq \frac{1}{4}\widetilde{C}^{1/p}s^{-1/p}, \quad \forall s \geq \bar{s},$$

where the second inequality holds because $\widetilde{\varepsilon} \leq \bar{\varepsilon} \leq 1$. This completes the proof. □

4.3 Upper bounds

We have now seen that $u_{\boldsymbol{\pi(s+1)}}$ cannot decay algebraically in for $s \to \infty$, regardless of the decay of the $\boldsymbol{b}$-weighted PCA eigenvalues $\lambda_{\boldsymbol{b},i}$. We now study the decay of $u_{\boldsymbol{\pi(s+1)}}$ for three different characteristic decay rates of these eigenvalues. In all three cases, we have $\lim_{i\to\infty} \lambda_{\boldsymbol{b},i} = 0$ so that the effective dimension $d(\varepsilon)$ is finite for every $\varepsilon > 0$, see Remark 4.5. In principle, the proof of the following result can be adapted to any other spectral decay rate as well, as long as this limit condition is satisfied.

Theorem 4.7 (Specific decays of $u_{\boldsymbol{\pi(s+1)}}$). *Let $\alpha > 0$.*

(a) Algebraic spectral decay: *Let $\lambda_{\boldsymbol{b},i} \lesssim i^{-\alpha}$ for every $i \in \mathbb{N}$. Then, for every $\delta, \eta > 0$, there exists $\bar{s} = \bar{s}(\delta, \eta) \in \mathbb{N}$ such that for every $s \geq \bar{s}$,*

$$u_{\boldsymbol{\pi(s+1)}} \leq (2(\alpha+\eta))^{\frac{1}{2(1/\alpha+\delta)}} \log(s)^{-\frac{1}{2(1/\alpha+\delta)}}. \tag{4.11}$$

(b) Exponential-algebraic spectral decay: *Let $\lambda_{\boldsymbol{b},i} \lesssim e^{-i^\alpha}$ for every $i \in \mathbb{N}$. Then, for every $\eta > 1$, there exists $\bar{s} = \bar{s}(\alpha, \eta) \in \mathbb{N}$ such that for every $s \geq \bar{s}$,*

$$u_{\boldsymbol{\pi(s+1)}} \lesssim e^{-\frac{1}{2}\left(1-\frac{1}{\eta(\alpha+1)}\right)^{-\frac{1}{1+1/\alpha}} \log(s)^{\frac{1}{1+1/\alpha}}}. \tag{4.12}$$

(c) Double exponential spectral decay: *Let $\lambda_{\boldsymbol{b},i} \lesssim e^{-e^{\alpha i}}$ for every $i \in \mathbb{N}$. Then, for every $\delta, \eta > 0$, there exists $\bar{s} = \bar{s}(\alpha, \delta, \eta) \in \mathbb{N}$ such that for every $s \geq \bar{s}$,*

$$u_{\boldsymbol{\pi(s+1)}} \lesssim e^{-\frac{1}{2}\left(\frac{\alpha}{\eta}\log(s)\right)^{\frac{1}{1+\delta}}}. \tag{4.13}$$

Proof. All rates can be derived by suitably choosing the parameter ε in the set $S(\varepsilon)$, defined in (4.5). By (4.7), the upper size bound (4.3) implies

$$|S(\varepsilon)| \leq \prod_{i=1}^{d(\varepsilon)} \left(\frac{\lambda_{\boldsymbol{b},i}}{i\varepsilon^2} + 1 \right) \leq \prod_{i=1}^{d(\varepsilon)} \frac{\lambda_{\boldsymbol{b},i}}{\varepsilon^2} \left(\frac{1}{i} + 1 \right) \leq (2\varepsilon^{-2})^{d(\varepsilon)} \prod_{i=1}^{d(\varepsilon)} \lambda_{\boldsymbol{b},i}, \tag{4.14}$$

where we used the fact that $\lambda_{\boldsymbol{b},i} \geq \varepsilon^2$ for $i \in [d(\varepsilon)]$ by definition of $d(\varepsilon)$, see (4.6). For brevity, we write in the following $d = d(\varepsilon)$.

Case (a). Let $\lambda_{\boldsymbol{b},i} \leq Ci^{-\alpha}$ for every $i \in \mathbb{N}$ and some constant $C > 0$. Plugging this into (4.14) and using the Stirling type estimate $d^d \leq e^d d!$ as well as the inequality $d^{-d} \leq e^{1/e}$ yields

$$|S_\varepsilon| \leq (2\varepsilon^{-2})^d C^d \prod_{i=1}^{d} i^{-\alpha} = (2C\varepsilon^{-2})^d (d!)^{-\alpha} \leq (2C\varepsilon^{-2})^d e^{\alpha d} d^{-\alpha d} \leq e^{\alpha d} e^{\alpha/e} (2C\varepsilon^{-2})^d. \tag{4.15}$$

Let $0 < \varepsilon \leq 1$. Then, by definition, $d = \min\{l \in \mathbb{N} : (l+1)^{-\alpha} < \varepsilon^2\} \leq \varepsilon^{-2/\alpha}$. We take the logarithm on both sides in (4.15) and deduce

$$\log(|S(\varepsilon)|) \leq \alpha d + \frac{\alpha}{e} + d \log(2C\varepsilon^{-2}) \leq \frac{\alpha}{e} + \alpha\varepsilon^{-2/\alpha} + \varepsilon^{-2/\alpha} \log(2C\varepsilon^{-2}).$$

Next, let $\delta, \eta > 0$ be arbitrary. There exists $0 < \bar{\varepsilon} = \bar{\varepsilon}(\delta, \eta) \leq 1$ such that $\log(2C\varepsilon^{-2}) \leq \eta\varepsilon^{-2\delta}$ for every $0 < \varepsilon \leq \bar{\varepsilon}$, which implies

$$\log(|S_\varepsilon|) \leq \frac{\alpha}{e} + \alpha\varepsilon^{-2/\alpha} + \eta\varepsilon^{-2/\alpha - 2\delta} \leq \frac{\alpha}{e} + (\alpha + \eta)\varepsilon^{-2/\alpha - 2\delta}, \quad \forall 0 < \varepsilon \leq \bar{\varepsilon}. \tag{4.16}$$

We set

$$\frac{\alpha}{e} + (\alpha + \eta)\varepsilon^{-2/\alpha - 2\delta} = \log(s)$$

and solve for ε^2, which yields

$$\varepsilon^2 = \varepsilon(s)^2 = \left(\log(s) - \frac{\alpha}{e} \right)^{-\frac{1}{1/\alpha + \delta}} (\alpha + \eta)^{\frac{1}{1/\alpha + \delta}}. \tag{4.17}$$

We can now choose $\bar{s} = \bar{s}(\delta, \eta) \in \mathbb{N}$ sufficiently large such that the right-hand side in (4.17) is smaller than $\bar{\varepsilon}$ and $\log(s)/2 \geq \alpha/e$ for every $s \geq \bar{s}$. Then, (4.16) holds with $\varepsilon = \varepsilon(s)$ and therefore

$$|S_{\varepsilon(s)}| \leq s, \quad \forall s \geq \bar{s}.$$

Since $S(\varepsilon) = \pi(|S(\varepsilon)|)$, it follows $u_{\boldsymbol{\pi(s+1)}} \notin S(\varepsilon)$ and consequently,

$$\begin{aligned} u_{\boldsymbol{\pi(s+1)}}^2 &\leq (\varepsilon(s)^{-2} + 1)^{-1} \leq \varepsilon(s)^2 \leq \left(\log(s) - \frac{\alpha}{e} \right)^{-\frac{1}{1/\alpha + \delta}} (\alpha + \eta)^{\frac{1}{1/\alpha + \delta}} \\ &\leq \log(s)^{-\frac{1}{1/\alpha + \delta}} (2(\alpha + \eta))^{\frac{1}{1/\alpha + \delta}} \end{aligned}$$

for every $s \geq \bar{s}$.

Case (b). Let $\lambda_{\boldsymbol{b},i} \leq Ce^{-i^\alpha}$ for every $i \in \mathbb{N}$ and some constant $C > 0$. With (4.14) we find

$$|S(\varepsilon)| \leq (2\varepsilon^{-2})^d C^d \prod_{i=1}^{d} e^{-i^\alpha} = (2\varepsilon^{-2})^d C^d e^{-\sum_{i=1}^d i^\alpha}. \tag{4.18}$$

Let $0 < \varepsilon \leq \min\{1, \sqrt{\lambda_{\boldsymbol{b},2}}\}$. We then have $2 \leq d = \min\{l \in \mathbb{N} : e^{-(l+1)^\alpha} < \varepsilon^2\} \leq \log(\varepsilon^{-2})^{1/\alpha} < d+1$. Taking the logarithm on both sides in (4.18) gives

$$\begin{aligned} \log(|S_\varepsilon|) &\leq d \log(2C\varepsilon^{-2}) - \sum_{i=1}^{d} i^\alpha \leq d \log(2C\varepsilon^{-2}) - \int_0^{d-1} t^\alpha dt \\ &\leq \log(\varepsilon^{-2})^{1/\alpha} \log(2C\varepsilon^{-2}) - \frac{1}{\alpha + 1} \left(\log(\varepsilon^{-2})^{1/\alpha} - 2 \right)^{\alpha + 1}. \end{aligned}$$

Next, let $\eta > 1$ be arbitrary. There exists $0 < \bar{\varepsilon} = \bar{\varepsilon}(\alpha, \eta) \leq \sqrt{\lambda_{\boldsymbol{b},2}}$ such that for every $0 < \varepsilon \leq \bar{\varepsilon}$,

$$\log(\varepsilon^{-2})^{1/\alpha} - 2 \geq \frac{1}{\eta^{1/(\alpha+1)}} \log(\varepsilon^{-2})^{1/\alpha}$$

and therefore

$$\log(|S_\varepsilon|) \leq \log(\varepsilon^{-2})^{1/\alpha} \log(2C\varepsilon^{-2}) - \frac{1}{\eta(\alpha+1)} \log(\varepsilon^{-2})^{1+1/\alpha} \leq \left(1 - \frac{1}{\eta(\alpha+1)}\right) \log(2C\varepsilon^{-2})^{1+1/\alpha}, \quad (4.19)$$

where $C' = \max\{C, 1\}$. We set

$$\left(1 - \frac{1}{\eta(\alpha+1)}\right) \log(2C'\varepsilon^{-2})^{1+1/\alpha} = \log(s)$$

and solve for ε^2, which yields

$$\varepsilon^2 = \varepsilon(s)^2 = 2C' e^{-\left(1 - \frac{1}{\eta(\alpha+1)}\right)^{-\frac{1}{1+1/\alpha}} \log(s)^{\frac{1}{1+1/\alpha}}}. \quad (4.20)$$

We can now choose $\bar{s} = \bar{s}(\alpha, \eta) \in \mathbb{N}$ sufficiently large such that the right-hand side of (4.20) is smaller than $\bar{\varepsilon}$ for every $s \geq \bar{s}$. Consequently, (4.19) holds with $\varepsilon = \varepsilon(s)$, and therefore

$$|S_{\varepsilon(s)}| \leq s \quad \forall s \geq \bar{s}.$$

Similar as in case (a), we conclude

$$u^2_{\boldsymbol{\pi(s+1)}} \leq (\varepsilon(s)^{-2} + 1)^{-1} \leq 2C' e^{-\left(1 - \frac{1}{\eta(\alpha+1)}\right)^{-\frac{1}{1+1/\alpha}} \log(s)^{\frac{1}{1+1/\alpha}}}$$

for every $s \geq \bar{s}$.

Case (c). Let $\lambda_{\boldsymbol{b},i} \leq Ce^{-e^{\alpha i}}$ for every $i \in \mathbb{N}$ and some constant $C > 0$. Plugging this into (4.14) yields

$$|S(\varepsilon)| \leq (2\varepsilon^{-2})^d C^d \prod_{i=1}^{d} e^{-e^{\alpha i}} = (2C\varepsilon^{-2})^d e^{-\sum_{i=1}^d e^{\alpha i}}. \quad (4.21)$$

Let $0 < \varepsilon \leq e^{-1/2}$. Then, by definition, $d = \min\{l \in \mathbb{N} : e^{-e^{\alpha(l+1)}} < \varepsilon^2\} \leq \log(\log(\varepsilon^{-2}))/\alpha < d + 1$. Taking the logarithm on both sides in (4.21) gives

$$\begin{aligned} \log(|S(\varepsilon)|) &\leq d \log(2C\varepsilon^{-2}) - \sum_{i=1}^{d} e^{\alpha i} = d \log(2C\varepsilon^{-2}) - \frac{e^{\alpha(d+1)} - 1}{e^\alpha - 1} \\ &\leq \frac{1}{\alpha} \log(\log(\varepsilon^{-2})) \log(2C\varepsilon^{-2}) - \frac{1}{e^\alpha - 1} \big(\log(\varepsilon^{-2}) - 1\big) \\ &\leq \frac{1}{\alpha} \log(\log(2C'\varepsilon^{-2})) \log(2C'\varepsilon^{-2}) + \frac{1}{e^\alpha - 1} \end{aligned}$$

with $C' = \max\{C, 1\}$. Next, let $\delta, \eta > 0$ be arbitrary. We argue similarly as in case (a). There exists $0 < \bar{\varepsilon} = \bar{\varepsilon}(\alpha, \delta, \eta) \leq e^{-1/2}$ such that $\log(\log(2C'\varepsilon^{-2})) \leq \frac{\eta}{2} \log(2C'\varepsilon^{-2})^\delta$ as well as $\frac{\eta}{2\alpha} \log(2C'\varepsilon^{-2})^{1+\delta} \geq \frac{1}{e^\alpha - 1}$ for every $0 < \varepsilon \leq \bar{\varepsilon}$. This implies

$$\log(|S(\varepsilon)|) \leq \frac{\eta}{2\alpha} \log(2C'\varepsilon^{-2})^{1+\delta} + \frac{1}{e^\alpha - 1} \leq \frac{\eta}{\alpha} \log(2C'\varepsilon^{-2})^{1+\delta}.$$

We set

$$\frac{\eta}{\alpha} \log(2C'\varepsilon^{-2})^{1+\delta} = \log(s)$$

and solve for ε^2, which yields

$$\varepsilon^2 = \varepsilon(s)^2 = 2C' e^{-\left(\frac{\alpha}{\eta}\log(s)\right)^{\frac{1}{1+\delta}}}. \tag{4.22}$$

We can now choose $\bar{s} = \bar{s}(\alpha, \delta, \eta) > 0$ sufficiently large such that the right hand-side of (4.22) is smaller than $\bar{\varepsilon}^2$ for every $s \geq \bar{s}$. Hence, (4.21) holds with $\varepsilon = \varepsilon(s)$, and therefore

$$|S_{\varepsilon(s)}| \leq s, \quad \forall s \geq \bar{s}.$$

We conclude

$$u^2_{\boldsymbol{\pi(s+1)}} \leq (\varepsilon(s)^{-2} + 1)^{-1} \leq \varepsilon(s)^2 \leq 2C' e^{-\left(\frac{\alpha}{\eta}\log(s)\right)^{\frac{1}{1+\delta}}}$$

for every $s \geq \bar{s}$. This completes the proof. □

Remark 4.8 (Asymptotics of the upper bounds).

(a) If $\lambda_{\boldsymbol{b},i} \lesssim i^{-\alpha}$ decays at least algebraically, then (4.11) shows that for fixed $\delta > 0$, $u_{\boldsymbol{\pi(s+1)}}$ decays at least logarithmically of order $\frac{1}{2(1/\alpha+\delta)}$. In particular, taking the limit $\delta \to 0$, we see that any logarithmic decay rate of order smaller than $\alpha/2$ can be attained asymptotically in the limit $s \to \infty$.

(b) If $\lambda_{\boldsymbol{b},i} \lesssim e^{-i^{\alpha}}$ has at least exponential-algebraic decay, then (4.12) yields a decay of $u_{\boldsymbol{\pi(s+1)}}$ which is faster than logarithmic for any $\eta > 1$. Taking the limit $\eta \to 1$, we see that any decay slower than

$$\exp\left(-\frac{1}{2}\left(1 + \frac{1}{\alpha}\right)^{\frac{1}{1+1/\alpha}} \log(s)^{\frac{1}{1+1/\alpha}}\right)$$

can be attained asymptotically in the limit $s \to \infty$.

(c) If $\lambda_{\boldsymbol{b},i} \lesssim e^{-e^{\alpha i}}$ decays at least double exponentially, then (4.13) shows that $u_{\boldsymbol{\pi(s+1)}}$ decays super-logarithmically but still subalgebraically (in accordance with Theorem 4.6). However, for fixed $\alpha > 0$, taking the limit $\delta \to 0$ and suitably choosing η, decay rates arbitrarily close to any algebraic order can be attained asymptotically in the limit $s \to \infty$.

4.4 Discussion

Theorem 4.6 shows that $u_{\boldsymbol{\pi(s+1)}}$ decays subalgebraically in s. In particular, as $p \in \mathbb{N}$ can be chosen arbitrarily large, we conclude that the decay is slower than *any* algebraic decay rate for all sufficiently large s. In combination with Theorem 4.1, we deduce the following **curse of parametric complexity**: *No s-term Wiener-Hermite PC expansion can converge with an algebraic rate uniformly for all operators in the Sobolev unit (Lipschitz) ball as $s \to \infty$. This holds regardless of the decay rate of the PCA eigenvalues.*

Note that in this context, the parameters, that is, the polynomial coefficients in the truncated Wiener-Hermite PC expansion, are elements of $\mathcal{Y}$. In [49], a related curse of (scalar) parametric complexity for learning (bounded) Lipschitz operators with PCA-Net was proved. It relates the learnability of Lipschitz operators by PCA-Nets to the size, i.e., the number of (scalar) neural network parameters, of the latter. More specifically, it implies the following (informal) result:

Theorem 4.9 (Curse of (scalar) parametric complexity for PCA-Net, cf. [49, Thm. 9])**.** *For any $\alpha > 0$, there exists a bounded Lipschitz operator $F \in C^{0,1}(\mathcal{X}, \mathcal{Y})$ and a constant $c_\alpha > 0$ such that*

$$\|F - \Psi\|^2_{L^2_\mu(\mathcal{X};\mathcal{Y})} \geq c_\alpha \left(\mathrm{size}(\psi)\right)^{-\alpha} \tag{4.23}$$

for every PCA-Net $\Psi = \mathcal{D}_\mathcal{Y} \circ \psi \circ \mathcal{E}_\mathcal{X}$. Here, $\psi : \mathbb{R}^{d_\mathcal{X}} \to \mathbb{R}^{d_\mathcal{Y}}$ is a (ReLU) neural network and $\mathcal{E}_\mathcal{X} : \mathcal{X} \to \mathbb{R}^{d_\mathcal{X}}$ and $\mathcal{D}_\mathcal{Y} : \mathbb{R}^{d_\mathcal{Y}} \to \mathcal{Y}$ are encoder and decoder maps, respectively, which are defined in terms of the empirical PCA eigenvalues based on finitely many fixed sample points.

Remark 4.10. Inspection of the proof of [49, Thm. 9] shows that the empirical PCA eigenvalues used in the definition of $\mathcal{E}_{\mathcal{X}}, \mathcal{D}_{\mathcal{Y}}$ can be replaced by the exact PCA eigenvalues which puts Theorem 4.9 in the setting of the present paper.

We now argue that the curse of ($\mathcal{Y}$-)parametric complexity described by Theorem 4.1 and Theorem 4.6 is consistent with the curse of (scalar) parametric complexity described by Theorem 4.9. To this end, suppose that Hermite polynomial s-term approximations of Lipschitz operators in the Sobolev unit Lipschitz ball are possible with some algebraic rate of order $\beta > 0$:

Assumption 4.11. *There exists $\beta > 0$ such that for all $F \in B_1^{\boldsymbol{b}}(\mathrm{Lip}(\mathcal{X}, \mathcal{Y}))$, there are constants $C(F) > 0$ and $\bar{s}(F) \in \mathbb{N}$ such that*

$$\left\| F - F_{\pi([s])} \right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \le C(F) s^{-\beta}, \quad \forall s \ge \bar{s}(F). \tag{4.24}$$

We then have the following result:

Proposition 4.12. *Suppose that Assumption 4.11 is true. Then there exists $\alpha > 0$ with the following property: For all $F \in C^{0,1}(\mathcal{X}, \mathcal{Y})$, there exist constants $C = C(F) > 0$ and $\bar{\epsilon} = \bar{\epsilon}(F) > 0$ such that for all $0 < \epsilon \le \bar{\epsilon}$, there is a PCA-Net $\Psi = \widetilde{\mathcal{D}}_{\mathcal{Y}} \circ \psi \circ \widetilde{\mathcal{E}}_{\mathcal{X}}$ such that*

$$\|F - \Psi\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \le \epsilon \quad \text{and} \quad \mathrm{size}(\psi) \le C \epsilon^{-1/\alpha}.$$

It is easy to see that Proposition 4.12 leads to a contradiction to Theorem 4.9, hence implying Assumption 4.11 to be false. With the notion introduced in [49], it asserts that α is an algebraic convergence rate for the class of $C^{0,1}(\mathcal{X}, \mathcal{Y})$-operators. The proof is based on [66, Thm. 3.9] which provides expression rate bounds for the approximation of multivariate Hermite polynomials by ReLU neural networks. As a preliminary step, we make the following observation: Let S be a downward closed subset of $\mathbb{N}_0^{\mathbb{N}}$, that is, $\boldsymbol{\gamma} \in S$ implies $\boldsymbol{\gamma'} \in S$ for every $\boldsymbol{\gamma'} \le \boldsymbol{\gamma}$. It is then easy to see by induction on the size of S that

$$\max_{\boldsymbol{\gamma} \in S} \|\boldsymbol{\gamma}\|_{\ell^1(\mathbb{N})} \le |S| - 1 \quad \text{and} \quad \max_{\boldsymbol{\gamma} \in S} |\mathrm{supp}(\boldsymbol{\gamma})| \le |S| - 1. \tag{4.25}$$

Proof of Proposition 4.12. Suppose that Assumption 4.11 is true. Let $F \in C^{0,1}(\mathcal{X}, \mathcal{Y})$ and define the rescaled operator $\widetilde{F} := r^{-1} \|F\|^{-1}_{C^{0,1}(\mathcal{X},\mathcal{Y})} F$, where $r := \sqrt{\dim(\mathcal{Y})}$ if $\mathcal{Y}$ is finite-dimensional and $r := \min\{1, \|\boldsymbol{b}\|_{\ell^2(\mathbb{N})}\}$ if $\mathcal{Y}$ is infinite-dimensional. Then, by Theorem 3.9, we have $\widetilde{F} \in B_1^{\boldsymbol{b}}(\mathrm{Lip}(\mathcal{X}, \mathcal{Y}))$. Next, let $\bar{s} = \bar{s}(F)$ be the constant in Assumption 4.11 and set $\bar{\epsilon} := \min\{\bar{s}^{-\beta/2}, e^{-\beta/(\beta+1)}\}$ so that $\bar{\epsilon}^{(\beta+1)/\beta} = \min\{\bar{s}^{-(\beta+1)/2}, e^{-1}\}$. Let $0 < \epsilon \le \bar{\epsilon}$ be arbitrary and choose $s \ge \bar{s}$ with $s^{-\beta/2} \sim \epsilon$.

We proceed with several observations: First, recall from (2.2) that the truncated Wiener-Hermite PC expansion $\widetilde{F}_{\pi([s])}$ is defined via the Hermite polynomials $H_{\boldsymbol{\pi(1)},d_1}, \ldots, H_{\boldsymbol{\pi(s)},d_s}$ with $d_i := \max\{j : j \in \mathrm{supp}(\boldsymbol{\pi(i)})\}$. By Remark 2.2, we can interpret each $H_{\boldsymbol{\pi(i)},d_i}$ as a function $H_{\boldsymbol{\pi(i)},\infty}$ on $\mathbb{R}^{\mathbb{N}}$. Second, observe that the set $\pi([s])$ is a downward closed subset of $\mathbb{N}_0^{\mathbb{N}}$ of size s. Hence, by (4.25), we have

$$\max_{\boldsymbol{\gamma} \in \pi([s])} \|\boldsymbol{\gamma}\|_{\ell^1(\mathbb{N})} \le s - 1 \quad \text{and} \quad \max_{\boldsymbol{\gamma} \in \pi([s])} |\mathrm{supp}(\boldsymbol{\gamma})| \le s - 1, \quad \forall s \in \mathbb{N}.$$

With these facts in hand, we can now directly apply [66, Thm. 3.9] to conclude that there exists a ReLU neural network $\psi = (\psi_1, \ldots, \psi_s) : \mathbb{R}^d \to \mathbb{R}^s$ with $d := \max\{d_1, \ldots, d_s\}$ and with each ψ_i depending solely on the variables $(x_i)_{i \in [d]}$ such that

$$\max_{i \in [s]} \left\| H_{\boldsymbol{\pi(i)},\infty} - \psi_i \right\|_{L^2_{\mu_\infty}(\mathbb{R}^{\mathbb{N}})} \le \epsilon^{(\beta+1)/\beta}. \tag{4.26}$$

Here, we interpret the ψ_i as functionals on $\mathbb{R}^{\mathbb{N}}$ by ignoring all variables x_i with $i > d$. Moreover, we have

$$\mathrm{size}(\psi) \lesssim s^6 \log(s) \log(\epsilon^{-(\beta+1)/\beta}) \le \left(1 + \frac{1}{\beta}\right) \epsilon^{-14/\beta - 1}. \tag{4.27}$$

We now define the encoder and decoder maps

$$\widetilde{\mathcal{E}}_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}^d, \quad \widetilde{\mathcal{E}}_{\mathcal{X}}(X) := \left(\frac{\langle X, \phi_1 \rangle_{\mathcal{X}}}{\sqrt{\lambda_1}}, \dots, \frac{\langle X, \phi_d \rangle_{\mathcal{X}}}{\sqrt{\lambda_d}} \right),$$

$$\widetilde{\mathcal{D}}_{\mathcal{Y}} : \mathbb{R}^s \to \mathcal{Y}, \quad \widetilde{\mathcal{D}}_{\mathcal{Y}}(\boldsymbol{x}) := \sum_{i=1}^{s} \left(\int_{\mathcal{X}} \widetilde{F} H_{\boldsymbol{\pi(i)},\boldsymbol{\lambda}} d\mu \right) x_i.$$

The corresponding (rescaled) PCA-Net $\Psi := r\|F\|^{-1}_{C^{0,1}(\mathcal{X},\mathcal{Y})} \left(\widetilde{\mathcal{D}}_{\mathcal{Y}} \circ \psi \circ \widetilde{\mathcal{E}}_{\mathcal{X}} \right)$ satisfies

$$\begin{aligned} \|F - \Psi\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} &\le r\|F\|^{-1}_{C^{0,1}(\mathcal{X},\mathcal{Y})} \left(\left\| \widetilde{F} - \widetilde{F}_{\pi([s])} \right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} + \left\| \widetilde{F}_{\pi([s])} - \widetilde{\mathcal{D}}_{\mathcal{Y}} \circ \psi \circ \widetilde{\mathcal{E}}_{\mathcal{X}} \right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \right) \\ &=: r\|F\|^{-1}_{C^{0,1}(\mathcal{X},\mathcal{Y})} (T_1(F) + T_2(F)). \end{aligned}$$

By Assumption 4.11, we can bound $T_1(F)$ by

$$T_1(F) \le C(F) s^{-\beta} \sim C(F)\epsilon^2 \le C(F)\epsilon,$$

where $C(F)$ is the constant in (4.24). For $T_2(F)$, we find by (4.26) and Parseval's identity that

$$\begin{aligned} T_2(F) &\le \sum_{i=1}^{s} \left\| \int_{\mathcal{X}} \widetilde{F} H_{\boldsymbol{\pi(i)},\boldsymbol{\lambda}} d\mu \right\|_{\mathcal{Y}} \|H_{\boldsymbol{\pi(i)},\infty} - \psi_i\|_{L^2_{\mu_\infty}(\mathbb{R}^{\mathbb{N}})} \le \epsilon^{(\beta+1)/\beta} s^{1/2} \left(\sum_{i=1}^{s} \left\| \int_{\mathcal{X}} \widetilde{F} H_{\boldsymbol{\pi(i)},\boldsymbol{\lambda}} d\mu \right\|^2_{\mathcal{Y}} \right)^{1/2} \\ &\lesssim \epsilon \|\widetilde{F}\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \le \epsilon. \end{aligned}$$

Altogether, we conclude

$$\|F - \Psi\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \le C' r \|F\|^{-1}_{C^{0,1}(\mathcal{X},\mathcal{Y})} (C(F) + 1)\epsilon$$

for some global constant $C' > 0$. Upon rescaling ϵ and $\bar{\epsilon}$ by the factor $C' r \|F\|^{-1}_{C^{0,1}(\mathcal{X},\mathcal{Y})}(C(F)+1)$, the claim follows in light of (4.27) with $\alpha = \beta/(14+\beta)$. □

We have now seen that the approximation of Lipschitz operators by Wiener-Hermite PC expansions cannot be done efficiently with algebraic convergence. On the other hand, we highlight that Theorem 4.7 shows the connection between the decay of the eigenvalues $\lambda_{\boldsymbol{b},i}$ and the decay of $u_{\boldsymbol{\pi(s+1)}}$. It implies that the curse of $\mathcal{Y}$-parametric (and hence also scalar parametric) complexity can be overcome at least asymptotically in the sense that decay rates arbitrarily close to any algebraic rate can be attained in the limit $s \to \infty$. This, however, requires the decay of the $\lambda_{\boldsymbol{b},i}$ to be faster than exponential, see case (b) in Theorem 4.7. As can be seen by case (c) therein, a double-exponential decay is sufficient.

5 Optimal sampling and (adaptive) m-widths

Up to this point, we have studied the approximation of $W^{1,2}_{\mu,\boldsymbol{b}}$- and Lipschitz operators by finite linear combinations of Hermite polynomials. This is a certain type of what in the field of information-based complexity is referred to as *linear information* [61, Chpt. 4.1.1]. In this section, we consider more general sampling and reconstruction schemes based on *adaptive information*, that is, (nonlinear) reconstruction from m adaptively chosen samples. We define adaptive sampling operators and the adaptive m-width and characterize the latter in terms of the weights $u_{\boldsymbol{\gamma}}$. We follow in parts ideas from [8] where recovery strategies based on adaptive information were studied for holomorphic operators. It is known that adaptive methods can only be better than nonadaptive methods by a factor of at most 2 and there are examples where adaptive methods perform slightly better than nonadaptive ones. We refer to Theorem 2 in [60] and references therein. For this reason, we consider adaptive sampling operators instead of nonadaptive ones. However, we will prove that for $W^{1,2}_{\mu,\boldsymbol{b}}$- and Lipschitz operators, linear approximation based on nonadaptive information is, in fact, optimal, see Theorem 5.4.

5.1 Adaptive sampling operators

We subsequently introduce in detail scalar-valued and Hilbert-valued adaptive sampling operators and discuss their definitions.

Definition 5.1 (Adaptive sampling operator; scalar-valued case)**.** Let $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ be a normed vector space and $m \in \mathbb{N}$. A (scalar-valued) *adaptive sampling operator* is a map of the form

$$\mathcal{L} : \mathcal{V} \to \mathbb{R}^m, \quad \mathcal{L}(F) = \begin{pmatrix} \mathcal{L}_1(F) \\ \mathcal{L}_2(F; \mathcal{L}_1(F)) \\ \vdots \\ \mathcal{L}_m(F; \mathcal{L}_1(F), \dots, \mathcal{L}_{m-1}(F)) \end{pmatrix},$$

where $\mathcal{L}_1 : \mathcal{V} \to \mathbb{R}$ is a bounded linear functional and $\mathcal{L}_i : \mathcal{V} \times \mathbb{R}^{i-1} \to \mathbb{R}$ is bounded and linear in its first component for $i = 2, \dots, m$.

Trivially, any bounded linear map $\mathcal{L} : \mathcal{V} \to \mathbb{R}^m$ is an adaptive sampling operator. Different choices for $\mathcal{V}$ lead to important special cases. If $\mathcal{V} \subset L^2_\mu(\mathcal{X})$ and $\boldsymbol{\gamma}^{(1)}, \dots, \boldsymbol{\gamma}^{(m)} \in \Gamma$, then we may define a sampling operator, generating (nonadaptive) *linear information*, by

$$\mathcal{L}(F) := \left(\int_{\mathcal{X}} F H_{\boldsymbol{\gamma}^{(i)}, \boldsymbol{\lambda}} d\mu \right)_{i \in [m]} \in \mathbb{R}^m, \quad \forall F \in \mathcal{V}. \tag{5.1}$$

If $\mathcal{V} = C(\mathcal{X})$ and $X_1, \dots, X_m \in \mathcal{V}$, we can define a pointwise sampling operator, generating (nonadaptive) *standard information* [61, Chpt. 4.1.1], by

$$\mathcal{L}(F) := (F(X_i))_{i \in [m]} \in \mathbb{R}^m, \quad \forall F \in C(\mathcal{X}). \tag{5.2}$$

In both cases, the $\boldsymbol{\gamma}^{(i)}$ and the X_i can, in principle, also be chosen adaptively based on previous measurements $\int_{\mathcal{X}} F H_{\boldsymbol{\gamma}^{(j)}} d\mu$ and $F(X_j)$, respectively, for $j \in [i-1]$.

Next, we generalize the definition to the Hilbert-valued case. For notational convenience, for any $Y \in \mathcal{Y}$, $\boldsymbol{v} = (v_i)_{i \in [m]} \in \mathbb{R}^m$, and $F \in L^2_\mu(\mathcal{X})$, we write $Y\boldsymbol{v}$ for the vector $(Yv_i)_{i \in [m]} \in \mathcal{Y}^m$ and YF for the map $X \mapsto YF(X)$.

Definition 5.2 (Adaptive sampling operator; Hilbert-valued case)**.** Let $\mathcal{V} \subset L^2_\mu(\mathcal{X}; \mathcal{Y})$ be a vector subspace with norm $\|\cdot\|_{\mathcal{V}}$ and consider an operator

$$\mathcal{L} : \mathcal{V} \to \mathcal{Y}^m, \;\; \mathcal{L}(F) = \begin{pmatrix} \mathcal{L}_1(F) \\ \mathcal{L}_2(F; \mathcal{L}_1(F)) \\ \vdots \\ \mathcal{L}_m(F; \mathcal{L}_1(F), \dots, \mathcal{L}_{m-1}(F)) \end{pmatrix},$$

where $\mathcal{L}_1 : \mathcal{V} \to \mathcal{Y}$ is a bounded linear operator and $\mathcal{L}_i : \mathcal{V} \times \mathcal{Y}^{i-1} \to \mathcal{Y}$ is bounded and linear in its first component for $i = 2, \dots, m$. Then $\mathcal{L}$ is a *Hilbert-valued adaptive sampling operator* if the following condition holds: There exist $Y, \widetilde{Y} \in \mathcal{Y} \setminus \{0\}$, a normed vector space $\widetilde{\mathcal{V}} \subset L^2_\mu(\mathcal{X})$, and a scalar-valued adaptive sampling operator $\widetilde{\mathcal{L}} : \widetilde{\mathcal{V}} \to \mathbb{R}^m$ such that, if $YF \in \mathcal{V}$ for some $F \in L^2_\mu(\mathcal{X})$, then $F \in \widetilde{\mathcal{V}}$ and $\mathcal{L}(YF) = \widetilde{Y}\widetilde{\mathcal{L}}(F)$.

This definition involves a technical assumption which links the Hilbert-valued case to the scalar-valued case and which we will use to establish a lower bound for the adaptive m-width, see (5.4). However, this condition is not too strong. It holds, for example, in the case of adaptive pointwise sampling. Here, we choose $\mathcal{V} = C(\mathcal{X}, \mathcal{Y})$ and define

$$\mathcal{L}(F) := (F(X_i))_{i \in [m]} \in \mathcal{Y}^m, \quad \forall F \in \mathcal{V},$$

where the ith sample point X_i is potentially chosen based on the previous measurements $F(X_1), \dots, F(X_{i-1})$. We then have

$$\mathcal{L}(YF) = Y\widetilde{\mathcal{L}}(F), \quad \forall F \in \widetilde{\mathcal{Y}} := C(\mathcal{X}), \forall Y \in \mathcal{Y},$$

where $\widetilde{\mathcal{L}} : \widetilde{\mathcal{V}} \to \mathbb{R}^m$ is the adaptive pointwise sampling operator in (5.2). As another example, we will see in the proof of the upper bound of the adaptive m-width that the Hilbert-valued version of linear sampling, as defined in (5.1), is a Hilbert-valued sampling operator in the sense of Definition 5.2 (under a mild condition on $\mathcal{V}$).

5.2 Adaptive m-widths and main result

We now formally define the adaptive m-width and state our main result.

Definition 5.3 (Adaptive m-width)**.** Let $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ be a normed vector subspace of $L^2_\mu(\mathcal{X};\mathcal{Y})$ and let $\mathcal{K} \subset \mathcal{V}$ be a subset. The adaptive m-width of $\mathcal{K}$ in $\mathcal{V}$ is given by

$$\begin{aligned} &\Theta_m(\mathcal{K};\mathcal{V}, L^2_\mu(\mathcal{X};\mathcal{Y})) \\ &\qquad := \inf\left\{ \sup_{F\in\mathcal{K}} \|F - \mathcal{T}(\mathcal{L}(F))\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} : \mathcal{L} : \mathcal{V} \to \mathcal{Y}^m \text{ adaptive}, \mathcal{T} : \mathcal{Y}^m \to L^2_\mu(\mathcal{X};\mathcal{Y}) \right\}. \end{aligned} \tag{5.3}$$

The adaptive m-width describes the smallest worst-case error that can be achieved when we reconstruct all operators in a set $\mathcal{K}$ by a reconstruction map $\mathcal{T}$ from m samples that have been generated by an adaptive Hilbert-valued sampling operator $\mathcal{L}$. It thus quantifies the error that can occur from optimally chosen sampling and reconstruction maps. Note that in (5.3) we allow for any (possibly nonlinear) reconstruction maps. The choice of $\mathcal{V}$, however, determines which sampling operators are allowed. If $\mathcal{V} = C(\mathcal{X}, \mathcal{Y})$, we can use pointwise sampling, whereas, if $\mathcal{V} = L^2_\mu(\mathcal{X};\mathcal{Y})$, we cannot.

We consider two choices for $\mathcal{K}$, namely $\mathcal{K} = B_1(W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y}))$ and $\mathcal{K} = B_1^{\boldsymbol{b}}(\mathrm{Lip}(\mathcal{X},\mathcal{Y}))$, see Definition 3.12. Our lower bound pertains to arbitrary $\mathcal{V}$. For the upper bound, we require the mild additional assumption that $\mathcal{V}$ is continuously embedded in $L^2_\mu(\mathcal{X};\mathcal{Y})$. Our main result in this section is the following characterization of the adaptive m-width of the Sobolev unit (Lipschitz) ball in terms of the weights $u_{\boldsymbol{\gamma}}$:

Theorem 5.4 (Tight characterization of the adaptive m-width)**.** *Let* $\mathcal{K} \in \{B_1(W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})), B_1^{\boldsymbol{b}}(\mathrm{Lip}(\mathcal{X},\mathcal{Y}))\}$ *and let* $m \in \mathbb{N}$*. We have the lower bound*

$$\Theta_m(\mathcal{K};\mathcal{V}, L^2_\mu(\mathcal{X};\mathcal{Y})) \geq u_{\boldsymbol{\pi(m+1)}}, \tag{5.4}$$

where $\pi : \mathbb{N} \to \Gamma$ *is a nonincreasing rearrangement of* $\boldsymbol{u} = (u_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma}\in\Gamma}$*, see* (3.3)*. If, in addition,* $\mathcal{V}$ *is continuously embedded in* $L^2_\mu(\mathcal{X};\mathcal{Y})$*, we have the matching upper bound*

$$\begin{aligned} \Theta_m(\mathcal{K};\mathcal{V}, L^2_\mu(\mathcal{X};\mathcal{Y})) &\leq \inf_{S\subset\Gamma, |S|\leq m} \sup_{F\in\mathcal{K}} \|F - F_S\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \\ &\leq \sup_{F\in\mathcal{K}} \left\|F - F_{\{\boldsymbol{\pi(1)},\dots,\boldsymbol{\pi(m)}\}}\right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \leq u_{\boldsymbol{\pi(m+1)}}. \end{aligned} \tag{5.5}$$

5.3 Proof of Theorem 5.4

The proof of the lower bound is based on the theory of Gelfand and Kolmogorov m-widths. We recall relevant results in Subsection 5.3.1. Further information can be found in [31, Chpt. 10]. Detailed proofs of the lower and upper bound are then given in Subsection 5.3.2 and Subsection 5.3.3, respectively.

5.3.1 Results about widths

Let $\mathcal{K}$ be a subset of a normed vector space $(\mathcal{Z}, \|\cdot\|_{\mathcal{Z}})$ and let $m \in \mathbb{N}$. The *Gelfand m-width* of $\mathcal{K}$ is defined by

$$d^m(\mathcal{K}, \mathcal{Z}) := \inf\left\{ \sup_{x\in\mathcal{K}\cap L^m} \|Z\|_{\mathcal{Z}} : L^m \text{ subspace of } \mathcal{X} \text{ with } \mathrm{codim}(L^m) \leq m \right\}.$$

An equivalent characterization is given by

$$d^m(\mathcal{K},\mathcal{Z}) = \inf\left\{ \sup_{Z\in\mathcal{K}\cap\ker(A)} \|Z\|_{\mathcal{Z}} : A : \mathcal{Z}\to\mathbb{R}^m \text{ linear}\right\}.$$

We also recall the *adaptive compressive m-width* of $\mathcal{K}$,

$$E^m_{\text{ada}}(\mathcal{K},\mathcal{Z}) := \inf\left\{\sup_{Z\in\mathcal{K}} \|Z-\Delta(\Gamma(Z))\|_{\mathcal{Z}} : \Gamma : \mathcal{Z}\to\mathbb{R}^m \text{ adaptive, } \Delta : \mathbb{R}^m\to\mathcal{Z}\right\},$$

and the *Kolmogorov m-width* of $\mathcal{K}$,

$$d_m(\mathcal{K},\mathcal{Z}) := \inf\left\{\sup_{K\in\mathcal{K}}\inf_{Z\in\mathcal{Z}_m} \|Z-K\|_{\mathcal{Z}} : \mathcal{Z}_m \text{ subspace of } \mathcal{Z} \text{ with } \dim(\mathcal{Z}_m)\le m\right\}.$$

Next, we state some standard results which relate the various notions of m-widths to each other.

Theorem 5.5 ([31, Thm. 10.4])**.** *If $\mathcal{K}$ is symmetric with respect to the origin, i.e., $-\mathcal{K}=\mathcal{K}$, then*

$$d^m(\mathcal{K},\mathcal{Z}) \le E^m_{\text{ada}}(\mathcal{K},\mathcal{Z}).$$

In [68], Stesin gave an explicit characterization of the Kolmogorov m-width in (finite) sequence spaces. For this, let us recall from Subsection 2.2 the notation $B^p_{\boldsymbol{w}}(I;\mathcal{Z})$ to denote the unit ball in the sequence space $\ell^p(I;\mathcal{Z})$.

Theorem 5.6 (Stesin)**.** *Let $N\in\mathbb{N}$ with $N>m$, $1\le q<p\le\infty$, and $\boldsymbol{w}\in\mathbb{R}^N$ be a vector of positive weights. Then*

$$d_m(B^p_{\boldsymbol{w}}([N]),\ell^q([N])) = \left(\max_{\substack{i_1,\dots,i_{N-m}\in[N]\\ i_k\neq i_j}}\left(\sum_{j=1}^{N-m} w_{i_j}^{\frac{pq}{p-q}}\right)^{\frac{1}{p}-\frac{1}{q}}\right)^{-1}.$$

In finite sequence spaces it is also possible to relate the Gelfand m-width and the Kolmogorov m-width:

Theorem 5.7 ([8, Thm. B.3])**.** *For $1\le p,q\le\infty$, let $\boldsymbol{w}\in\mathbb{R}^N$ be a vector of positive weights and let $1\le p^*,q^*\le\infty$ be such that $1/p+1/p^*=1$ and $1/q+1/q^*=1$. Then*

$$d_m(B^p([N]),\ell^q_{\boldsymbol{w}}([N])) = d^m(B^{q^*}_{1/\boldsymbol{w}}([N]),\ell^{p^*}([N])).$$

Lemma 5.8 ([8, Lem. B.4])**.** *Let $\boldsymbol{w}\in\mathbb{R}^N$ be a vector of positive weights and $1\le p,q\le\infty$. Then*

$$d_m(B^p([N]),\ell^q_{\boldsymbol{w}}([N])) = d_m(B^p_{1/\boldsymbol{w}}([N]),\ell^q([N])).$$

5.3.2 Lower bound

Since $B^{\boldsymbol{b}}_1(\mathrm{Lip}(\mathcal{X},\mathcal{Y}))$ is a subset of $B_1(W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y}))$, it suffices to prove the lower bound for the adaptive m-width of $\mathcal{K}=B^{\boldsymbol{b}}_1(\mathrm{Lip}(\mathcal{X},\mathcal{Y}))$, that is,

$$\Theta_m(B^{\boldsymbol{b}}_1(\mathrm{Lip}(\mathcal{X},\mathcal{Y}));\mathcal{V},L^2_\mu(\mathcal{X};\mathcal{Y})) \ge u_{\boldsymbol{\pi(m+1)}}, \quad \forall m\in\mathbb{N}. \tag{5.6}$$

This implies the same lower bound in the case $\mathcal{K}=B_1(W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y}))$. The proof consists of two main steps. We first reduce the problem to a discrete one which involves the adaptive compressive m-width of the unit ball in a space of suitably weighted finite sequences, see Lemma 5.9. In the discrete setting, we can then use Theorems 5.5 and 5.7, and Lemma 5.8 to relate the adaptive compressive m-width to the Kolmogorov m-width. In the second step, we apply Theorem 5.6 and combine it with a limiting argument to conclude the claim.

For the next result, let us recall the notation $(c\boldsymbol{u})_I$ with $\boldsymbol{u}=(u_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma}\in\Gamma}$, $c\in\mathbb{R}$, and $I\subset\Gamma$ to denote the scaled subsequence $(cu_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma}\in I}$.

Lemma 5.9 (Reduction to discrete problem)**.** *Let $I \subset \Gamma$ be a finite index set. Then, for every constant $c \in (0,1)$, we have*

$$\Theta_m(B_1^{\boldsymbol{b}}(\mathrm{Lip}(\mathcal{X},\mathcal{Y})); \mathcal{V}, L^2_\mu(\mathcal{X};\mathcal{Y})) \geq d^m(B^2_{(c\boldsymbol{u})_I}(I), \ell^2(I)).$$

The proof of this lemma is based on the construction of a suitable Lipschitz continuous operator. To this end, let $R > 0$ and $n \in \mathbb{N}$, and consider the capped one-dimensional Hermite polynomials

$$\widetilde{H}_{n,R}(x) := \begin{cases} H_n(x) & \text{if } -R \leq x \leq R, \\ H_n(R) & \text{if } x > R, \\ H_n(-R) & \text{if } x < -R. \end{cases} \tag{5.7}$$

For $\boldsymbol{\gamma} \in \Gamma$ and $d \in \mathbb{N}$, we define

$$\widetilde{H}_{\boldsymbol{\gamma},R,d} : \mathbb{R}^d \to \mathbb{R}, \quad \widetilde{H}_{\boldsymbol{\gamma},R,d}(\boldsymbol{x}) := \prod_{i=1}^d \widetilde{H}_{\gamma_i,R}(x_i), \tag{5.8}$$

as well as

$$\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} : \mathcal{X} \to \mathbb{R}, \quad \widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}(X) := \prod_{i=1}^\infty \widetilde{H}_{\gamma_i,R}\left(\frac{\langle X, \phi_i\rangle_{\mathcal{X}}}{\sqrt{\lambda_i}}\right). \tag{5.9}$$

Lemma 5.10 (Lipschitz continuity)**.** *For every $R > 0$ and every $\boldsymbol{\gamma} \in \Gamma$, the functional $\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} : \mathcal{X} \to \mathbb{R}$, defined in* (5.9)*, is Lipschitz continuous.*

Proof. Fix $R > 0$ and $\boldsymbol{\gamma} \in \Gamma$ with $\mathrm{supp}(\boldsymbol{\gamma}) \subset [d]$ for some $d \in \mathbb{N}$. We define the scaling functional

$$S_{\boldsymbol{\lambda},d} : \mathcal{X} \to \mathbb{R}^d, \quad S_{\boldsymbol{\lambda},d}(X) := \left(\frac{\langle X, \phi_i\rangle_{\mathcal{X}}}{\sqrt{\lambda_i}}\right)_{i\in[d]},$$

and note that $\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} = \widetilde{H}_{\boldsymbol{\gamma},R,d} \circ S_{\boldsymbol{\lambda},d}$, with $\widetilde{H}_{\boldsymbol{\gamma},R,d}$ given by (5.8). As $S_{\boldsymbol{\lambda},d}$ is Lipschitz continuous, it suffices to show that $\widetilde{H}_{\boldsymbol{\gamma},R,d}$ is Lipschitz continuous. This, in turn, follows by a simple induction argument over the dimension d. □

By Lemma 5.10 and Theorem 3.9, we have $\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ for every $R > 0$. The next result establishes the connection between $\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}$ and $H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}$ in the limit $R \to \infty$. For its proof we introduce the following notation: For $\boldsymbol{x} \in \mathbb{R}^d$, $d \in \mathbb{N}$, and $1 \leq k \leq d$, we write $\boldsymbol{x}_{[k]} := (x_1, \ldots, x_k) \in \mathbb{R}^k$. We also recall the complementary error function $\mathrm{erfc} : \mathbb{R} \to \mathbb{R}, x \mapsto \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt$, which grows faster than polynomially at infinity, that is,

$$\lim_{t\to\infty} t^m \mathrm{erfc}(t) = 0, \quad \forall m \in \mathbb{N}. \tag{5.10}$$

Lemma 5.11 (Convergence in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X})$)**.** *For every $\boldsymbol{\gamma} \in \Gamma$, we have*

$$\lim_{R\to\infty} \widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} = H_{\boldsymbol{\gamma},\boldsymbol{\lambda}} \quad \text{in } W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X}).$$

Proof. Let $\boldsymbol{\gamma} \in \Gamma$ with $\mathrm{supp}(\boldsymbol{\gamma}) \subset [d]$, $d \in \mathbb{N}$. We first consider convergence in $L^2_\mu(\mathcal{X})$. By Fubini's theorem and a change of variables, one has

$$\left\|\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} - H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}\right\|^2_{L^2_\mu(\mathcal{X})} = \int_{\mathbb{R}^d} \left|\widetilde{H}_{\boldsymbol{\gamma},R,d}(\boldsymbol{x}) - H_{\boldsymbol{\gamma},d}(\boldsymbol{x})\right|^2 d\mu_d(\boldsymbol{x}).$$

It thus suffices to show that

$$\lim_{R\to\infty} \widetilde{H}_{\boldsymbol{\gamma},R,d} = H_{\boldsymbol{\gamma},d} \quad \text{in } L^2_{\mu_d}(\mathbb{R}^d), \ \forall d \in \mathbb{N}. \tag{5.11}$$

For this, we use induction over d and start with $d = 1$. For $n \in \mathbb{N}_0$, we compute

$$\begin{aligned}\int_{\mathbb{R}} \left|\widetilde{H}_{n,R}(x) - H_n(x)\right|^2 d\mu_1(x) &= \int_{-\infty}^{-R} |H_n(-R) - H_n(x)|^2 d\mu_1(x) + \int_{R}^{\infty} |H_n(R) - H_n(x)|^2 d\mu_1(x) \\ &\le \left(H_n(-R)^2 + H_n(R)^2\right) \operatorname{erfc}\left(\frac{R}{\sqrt{2}}\right) + 2\int_{[-R,R]^c} H_n(x)^2 d\mu_1(x) \\ &=: T_1(R) + T_2(R),\end{aligned}$$

where we used the notation $[-R, R]^c := \mathbb{R} \setminus [-R, R]$. By (5.10), we have $\lim_{R\to\infty} T_1(R) = 0$. Since $H_n \in L^2_{\mu_1}(\mathbb{R})$, the second term $T_2(R)$ converges to zero as $R \to \infty$ by the dominated convergence theorem. Now, let $d > 1$ and suppose that (5.11) holds for any $1 \le d' < d$. Without loss of generality, we may assume $R \ge 1$. Then, by Fubini's theorem, we have

$$\begin{aligned}&\int_{\mathbb{R}^d} \left|\widetilde{H}_{\boldsymbol{\gamma},R,d}(\boldsymbol{x}) - H_{\boldsymbol{\gamma},d}(\boldsymbol{x})\right|^2 d\mu_d(\boldsymbol{x}) \\ &\le 2\int_{\mathbb{R}}\int_{\mathbb{R}^{d-1}} \left|\widetilde{H}_{\boldsymbol{\gamma},d-1,R}(\boldsymbol{x}_{[d-1]})\widetilde{H}_{\gamma_d,R}(x_d) - \widetilde{H}_{\boldsymbol{\gamma},d-1,R}(\boldsymbol{x}_{[d-1]})H_{\gamma_d}(x_d)\right|^2 d\mu_{d-1}(\boldsymbol{x}_{[d-1]})d\mu_1(x_d) \\ &\quad + 2\int_{\mathbb{R}}\int_{\mathbb{R}^{d-1}} \left|\widetilde{H}_{\boldsymbol{\gamma},d-1,R}(\boldsymbol{x}_{[d-1]})H_{\gamma_d}(x_d) - H_{\boldsymbol{\gamma}_{[d-1]}}(\boldsymbol{x}_{[d-1]})H_{\gamma_d}(x_d)\right|^2 d\mu_{d-1}(\boldsymbol{x}_{[d-1]})d\mu_1(x_d) \\ &=: t_1(R) + t_2(R).\end{aligned}$$

The term $t_1(R)$ can be bounded from above by

$$t_1(R) \le 2\left(\sup_{R\ge 1}\int_{\mathbb{R}^{d-1}} \left|\widetilde{H}_{\boldsymbol{\gamma},d-1,R}(\boldsymbol{x}_{[d-1]})\right|^2 d\mu_{d-1}(\boldsymbol{x}_{[d-1]})\right)\int_{\mathbb{R}} \left|\widetilde{H}_{\gamma_d,R}(x_d) - H_{\gamma_d}(x_d)\right|^2 d\mu_1(x_d).$$

By induction hypothesis for $d' = d - 1$, the term $\widetilde{H}_{\boldsymbol{\gamma},d-1,R}$ converges in $L^2_{\mu_{d-1}}(\mathbb{R}^{d-1})$ as $R \to \infty$. Hence, the supremum over $R \ge 1$ of the first integral on the right-hand side is finite. Applying the induction hypotheses for $d' = 1$, we conclude that the second integral over $\mathbb{R}$ converges to zero as $R \to \infty$. A similar argument shows that $\lim_{R\to\infty} t_2(R) = 0$. This completes the proof of (5.11).

Next, we consider convergence of $\nabla_{\mathcal{X}_{\boldsymbol{b}}}\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}$ in $L^2_\mu(\mathcal{X}; \mathcal{X}_{\boldsymbol{b}})$. For this, we recall the basis $\{\eta_i\}_{i\in\mathbb{N}}$ of $\mathcal{X}_{\boldsymbol{b}}$, defined in (2.1), and the d-dimensional Hermite polynomials $H_{\boldsymbol{\gamma},d}$, defined in (2.2). Note that the capped Hermite polynomials $\widetilde{H}_{n,R}$, as defined in (5.7), are absolutely continuous along each compact subinterval in $\mathbb{R}$. We can therefore apply Lemma C.13. We write $x_i := \langle X, \phi_i\rangle_{\mathcal{X}}$, $i \in [d]$, and compute

$$\begin{aligned}\frac{\partial}{\partial\eta_i}\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}(X) &= b_i\lambda_i^{-1/2}\partial_i\widetilde{H}_{\boldsymbol{\gamma},R,d}(\lambda_1^{-1/2}x_1,\ldots,\lambda_d^{-1/2}x_d) \\ &= \begin{cases} b_i\lambda_i^{-1/2}\partial_i H_{\boldsymbol{\gamma},d}(\lambda_1^{-1/2}x_1,\ldots,\lambda_d^{-1/2}x_d) & \text{if } x_i \in [-R,R], \\ 0 & \text{if } x_i \in [-R,R]^c.\end{cases}\end{aligned}$$

Moreover, $\frac{\partial}{\partial\eta_i}\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} = \frac{\partial}{\partial\eta_i}H_{\boldsymbol{\gamma},\boldsymbol{\lambda}} = 0$ for $i > d$. Consequently, we have

$$\begin{aligned}\left\|\nabla_{\mathcal{X}_{\boldsymbol{b}}}\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} - \nabla_{\mathcal{X}_{\boldsymbol{b}}}H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}\right\|^2_{L^2_\mu(\mathcal{X};\mathcal{X}_{\boldsymbol{b}})} &= \int_{\mathcal{X}}\sum_{i=1}^d\left|\frac{\partial}{\partial\eta_i}\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}(X) - \frac{\partial}{\partial\eta_i}H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(X)\right|^2 d\mu(X) \\ &= \sum_{i=1}^d\int_{\mathbb{R}^{i-1}\times[-R,R]^c\times\mathbb{R}^{d-i}}\left|b_i\lambda_i^{-1/2}\partial_i H_{\boldsymbol{\gamma},d}(x_1,\ldots,x_d)\right|^2 d\mu_d(\boldsymbol{x}).\end{aligned}$$

Since $\partial_i H_{\boldsymbol{\gamma},d} \in L^2_{\mu_d}(\mathbb{R}^d)$, the right-hand side converges to zero as $R \to \infty$ by the dominated convergence theorem. The proof is now complete. □

Lemma 5.12 (Riesz basis)**.** *Let $I \subset \Gamma$ be finite. Then, for every $\varepsilon > 0$ there exists $\bar{R} > 0$ such that for every $R \geq \bar{R}$ we have*

$$(1-\varepsilon)\|\boldsymbol{x}\|_{\ell^2(I)}^2 \leq \left\| \sum_{\boldsymbol{\gamma}\in I} x_{\boldsymbol{\gamma}} \widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} \right\|_{L^2_\mu(\mathcal{X})}^2 \leq (1+\varepsilon)\|\boldsymbol{x}\|_{\ell^2(I)}^2, \quad \forall \boldsymbol{x} = (x_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma}\in I} \in \mathbb{R}^I. \tag{5.12}$$

In particular, $\{\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}\}_{\boldsymbol{\gamma}\in I}$ is a Riesz basis of $\operatorname{span}\{\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} : \boldsymbol{\gamma} \in I\}$ *for every $R \geq \bar{R}$ with Riesz constants at worst $1 \pm \varepsilon$.*

Proof. We first work in $d = 1$ dimension and compute for $n, m \in \mathbb{N}_0$,

$$\begin{aligned}
&\left\langle \widetilde{H}_{n,R}, \widetilde{H}_{m,R} \right\rangle_{L^2_{\mu_1}(\mathbb{R})} = \int_{\mathbb{R}} \widetilde{H}_{n,R}(x)\widetilde{H}_{m,R}(x) d\mu_1(x) \\
&= \int_{-R}^{R} H_n(x)H_m(x)d\mu_1(x) + \int_{-\infty}^{-R} H_n(-R)H_m(-R)d\mu_1(x) + \int_{R}^{\infty} H_n(R)H_m(R)d\mu_1(x) \\
&= \int_{-R}^{R} H_n(x)H_m(x)d\mu_1(x) + \frac{1}{2}H_n(-R)H_m(-R)\mathrm{erfc}\left(\frac{R}{\sqrt{2}}\right) + \frac{1}{2}H_n(R)H_m(R)\mathrm{erfc}\left(\frac{R}{\sqrt{2}}\right) \\
&=: T_1(R) + T_2(R) + T_3(R).
\end{aligned}$$

Note that H_nH_m is a polynomial of order $n + m$. Hence, by (5.10), we deduce $\lim_{R\to\infty} T_2(R) = 0$ and $\lim_{R\to\infty} T_3(R) = 0$. Moreover, by the dominated convergence theorem and orthonormality of the Hermite polynomials, we have

$$\lim_{R\to\infty} T_1(R) = \int_{-\infty}^{\infty} H_n(x)H_m(x)d\mu_1(x) = \delta_{n,m}.$$

Altogether,

$$\lim_{R\to\infty} \left\langle \widetilde{H}_{n,R}, \widetilde{H}_{m,R} \right\rangle_{L^2_{\mu_1}(\mathbb{R})} = \delta_{n,m}. \tag{5.13}$$

Now, let $\boldsymbol{\gamma}, \boldsymbol{\gamma'} \in \Gamma$ with $\operatorname{supp}(\boldsymbol{\gamma}), \operatorname{supp}(\boldsymbol{\gamma'}) \subset [d]$ for some $d \in \mathbb{N}$. Since

$$\left\langle \widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}, \widetilde{H}_{\boldsymbol{\gamma'},R,\boldsymbol{\lambda}} \right\rangle_{L^2_\mu(\mathcal{X})} = \int_{\mathbb{R}^d} \widetilde{H}_{\boldsymbol{\gamma},R,d}(\boldsymbol{x})\widetilde{H}_{\boldsymbol{\gamma'},d,R}(\boldsymbol{x})d\mu_d(\boldsymbol{x}) = \prod_{i=1}^{d} \int_{\mathbb{R}} \widetilde{H}_{\gamma_i,R}(x_i)\widetilde{H}_{\gamma'_i,R}(x_i)d\mu_1(x_i),$$

we conclude by (5.13) that

$$\lim_{R\to\infty} \left\langle \widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}, \widetilde{H}_{\boldsymbol{\gamma'},R,\boldsymbol{\lambda}} \right\rangle_{L^2_\mu(\mathcal{X})} = \delta_{\boldsymbol{\gamma},\boldsymbol{\gamma'}}.$$

Next, let us fix some arbitrary $\varepsilon > 0$. Then there exists $\bar{R} > 0$ such that for every $R \geq \bar{R}$, we have

$$\left| \left\langle \widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}, \widetilde{H}_{\boldsymbol{\gamma'},R,\boldsymbol{\lambda}} \right\rangle_{L^2_\mu(\mathcal{X})} - \delta_{\boldsymbol{\gamma},\boldsymbol{\gamma'}} \right| \leq \varepsilon.$$

Since $I \subset \Gamma$ is finite, we obtain for any $\boldsymbol{x} = (x_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma}\in I} \in \mathbb{R}^I$,

$$\begin{aligned}
\left\| \sum_{\boldsymbol{\gamma}\in I} x_{\boldsymbol{\gamma}} \widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} \right\|_{L^2_\mu(\mathcal{X})}^2 &= \sum_{\boldsymbol{\gamma}\in I} \sum_{\substack{\boldsymbol{\gamma'}\in I \\ \boldsymbol{\gamma'}\neq\boldsymbol{\gamma}}} x_{\boldsymbol{\gamma}} x_{\boldsymbol{\gamma'}} \underbrace{\left\langle \widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}, \widetilde{H}_{\boldsymbol{\gamma'},R,\boldsymbol{\lambda}} \right\rangle_{L^2_\mu(\mathcal{X})}}_{\leq\varepsilon} + \sum_{\boldsymbol{\gamma}\in\Gamma} x_{\boldsymbol{\gamma}}^2 \underbrace{\left\langle \widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}, \widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} \right\rangle_{L^2_\mu(\mathcal{X})}}_{\leq 1+\varepsilon} \\
&\leq \varepsilon\|\boldsymbol{x}\|_{\ell^1(I)}^2 + (1+\varepsilon)\|\boldsymbol{x}\|_{\ell^2(I)}^2 \leq (\varepsilon\,|I| + 1 + \varepsilon)\|\boldsymbol{x}\|_{\ell^2(I)}^2.
\end{aligned}$$

Similarly, we have

$$\left\|\sum_{\boldsymbol{\gamma}\in I} x_{\boldsymbol{\gamma}} \widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}\right\|_{L^2_\mu(\mathcal{X})}^2 \geq -\varepsilon \|\boldsymbol{x}\|_{\ell^1(I)}^2 + (1-\varepsilon)\|\boldsymbol{x}\|_{\ell^2(I)}^2 \geq (-\varepsilon\,|I| + 1 - \varepsilon)\|\boldsymbol{x}\|_{\ell^2(I)}^2.$$

As $\varepsilon > 0$ was arbitrary, the claim follows. □

We are now ready to reduce the adaptive m-width to the Gelfand m-width in a discrete setting.

Proof of Lemma 5.9. Let $\mathcal{L} : \mathcal{V} \to \mathcal{Y}^m$ be an adaptive sampling operator as in Definition 5.2. Then there exist $Y, \widetilde{Y} \in \mathcal{Y} \setminus \{0\}$ and a normed vector space $\widetilde{\mathcal{V}} \subset L^2_\mu(\mathcal{X})$ such that, if $YF \in \mathcal{V}$ for some $F \in L^2_\mu(\mathcal{X})$, then $F \in \widetilde{\mathcal{V}}$ and $\mathcal{L}(YF) = \widetilde{Y}\widetilde{\mathcal{L}}(F)$, where $\widetilde{\mathcal{L}} : \widetilde{\mathcal{V}} \to \mathbb{R}^m$ is a scalar-valued adaptive sampling operator as in Definition 5.1. Next, let $I \subset \Gamma$ be a finite subset and let us fix $c \in (0,1)$ and $\varepsilon > 0$. By Lemma 5.11 and Lemma 5.12, there exists $R > 0$ sufficiently large such that (5.12) holds and

$$\left\|\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} - H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}\right\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X})} \leq \frac{1-c}{c}\,|I|^{-1/2}, \quad \forall \boldsymbol{\gamma} \in I. \tag{5.14}$$

We fix some arbitrary sequence $\boldsymbol{x} = (x_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma}\in I} \in \mathbb{R}^I$ with $\|\boldsymbol{x}\|_{\ell^2_{\boldsymbol{u}_I}(I)} \leq c$ and define $F_R, F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ by

$$F_R := \frac{Y}{\|Y\|_{\mathcal{Y}}} \sum_{\boldsymbol{\gamma}\in I} x_{\boldsymbol{\gamma}} \widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} \quad \text{and} \quad F := \frac{Y}{\|Y\|_{\mathcal{Y}}} \sum_{\boldsymbol{\gamma}\in I} x_{\boldsymbol{\gamma}} H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}.$$

Note that $\|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} = \|\boldsymbol{x}\|_{\ell^2_{\boldsymbol{u}_I}(I)} \leq c$ by Theorem 3.5, and therefore, by (5.14),

$$\|F_R - F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} \leq \sum_{\boldsymbol{\gamma}\in I} |x_{\boldsymbol{\gamma}}| \left\|\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} - H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}\right\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X})} \leq \frac{1-c}{c}\,|I|^{-1/2} \sum_{\boldsymbol{\gamma}\in I} |x_{\boldsymbol{\gamma}}| \leq 1 - c.$$

Thus, by the triangle inequality, $\|F_R\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} \leq 1$. By Lemma 5.10 and since I is finite, the operator F_R is Lipschitz continuous and we conclude $F_R \in B_1^{\boldsymbol{b}}(\mathrm{Lip}(\mathcal{X},\mathcal{Y}))$.

By Lemma 5.12, the family of functionals $\{\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}\}_{\boldsymbol{\gamma}\in I}$ is a Riesz basis of $\mathcal{F} := \operatorname{span}\{\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} : \boldsymbol{\gamma} \in I\}$. Hence, there exists a unique biorthogonal dual basis $\{\widehat{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}\}_{\boldsymbol{\gamma}\in I}$ such that $\langle \widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}, \widehat{H}_{\boldsymbol{\gamma'},R,\boldsymbol{\lambda}}\rangle = \delta_{\boldsymbol{\gamma},\boldsymbol{\gamma'}}$ for every $\boldsymbol{\gamma}, \boldsymbol{\gamma'} \in I$. The orthogonal projection onto $\mathcal{F}$ is defined by

$$P_{\mathcal{F}} : L^2_\mu(\mathcal{X}) \to \mathcal{F}, \quad P_{\mathcal{F}} g := \sum_{\boldsymbol{\gamma}\in I} \left\langle g, \widehat{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} \right\rangle_{L^2_\mu(\mathcal{X})} \widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}.$$

Let $\{\psi_j\}_{j\in\mathbb{N}}$ be an orthonormal basis of $\mathcal{Y}$. For $G \in L^2_\mu(\mathcal{X};\mathcal{Y})$, we write $g_j := \langle G, \psi_j\rangle_{\mathcal{Y}} \in L^2_\mu(\mathcal{X})$ and find by (5.12) that

$$\begin{aligned}
\|G\|_{L^2_\mu(\mathcal{X};\mathcal{Y})}^2 &= \sum_{j=1}^{\infty} \|g_j\|_{L^2_\mu(\mathcal{X})}^2 \geq \sum_{j=1}^{\infty} \|P_{\mathcal{F}} g_j\|_{L^2_\mu(\mathcal{X})}^2 \geq \sum_{j=1}^{\infty} (1-\varepsilon) \sum_{\boldsymbol{\gamma}\in I} \left| \left\langle g_j, \widehat{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} \right\rangle_{L^2_\mu(\mathcal{X})} \right|^2 \\
&= (1-\varepsilon) \sum_{\boldsymbol{\gamma}\in I} \sum_{j=1}^{\infty} \left| \left\langle g_j, \widehat{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} \right\rangle_{L^2_\mu(\mathcal{X})} \right|^2 = (1-\varepsilon) \sum_{\boldsymbol{\gamma}\in I} \left\| \int_{\mathcal{X}} G \widehat{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} d\mu \right\|_{\mathcal{Y}}^2.
\end{aligned} \tag{5.15}$$

We now define the scalar-valued adaptive sampling operator

$$\Xi_R : \mathbb{R}^I \to \mathbb{R}^m, \quad \Xi_R(\boldsymbol{z}) := \widetilde{\mathcal{L}}\left(\frac{1}{\|Y\|_{\mathcal{Y}}} \sum_{\boldsymbol{\gamma}\in I} z_{\boldsymbol{\gamma}} \widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} \right).$$

We need to show that it is well-defined, that is, $\|Y\|_{\mathcal{Y}}^{-1}\sum_{\boldsymbol{\gamma}\in I} z_{\boldsymbol{\gamma}}\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}} \in \widetilde{\mathcal{V}}$ for every $\boldsymbol{z}\in\mathbb{R}^I$. Since $B_1^{\boldsymbol{b}}(\mathrm{Lip}(\mathcal{X},\mathcal{Y}))\subset\mathcal{V}$, it suffices to observe that $Y\|Y\|_{\mathcal{Y}}^{-1}\sum_{\boldsymbol{\gamma}\in I} z_{\boldsymbol{\gamma}}\widetilde{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}$ is Lipschitz continuous as an operator from $\mathcal{X}$ to $\mathcal{Y}$ and it therefore lies in $\mathcal{V}$. Next, let $\mathcal{T}:\mathcal{Y}^m\to L^2_\mu(\mathcal{X};\mathcal{Y})$ be an arbitrary reconstruction map. We define $\widetilde{\mathcal{T}}:\mathbb{R}^m\to L^2_\mu(\mathcal{X};\mathcal{Y})$ by

$$\widetilde{\mathcal{T}}:\mathbb{R}^m\to L^2_\mu(\mathcal{X};\mathcal{Y}),\quad \widetilde{\mathcal{T}}(\boldsymbol{z}):=\mathcal{T}(\widetilde{Y}\boldsymbol{z}),$$

and observe that

$$\mathcal{T}(\mathcal{L}(F_R))=\mathcal{T}(\widetilde{Y}\Xi_R(\boldsymbol{x}))=\widetilde{\mathcal{T}}(\Xi_R(\boldsymbol{x})).$$

We now set $G:=F_R-\mathcal{T}(\mathcal{L}(F_R))$ in (5.15). We use the estimate $\|Z\|_{\mathcal{Y}}\geq\|Y\|_{\mathcal{Y}}^{-1}|\langle Z,Y\rangle_{\mathcal{Y}}|$, which holds for every $Z\in\mathcal{Y}$ by the Cauchy-Schwarz inequality, and compute

$$\begin{aligned}\|F_R-\mathcal{T}(\mathcal{L}(F_R))\|^2_{L^2_\mu(\mathcal{X};\mathcal{Y})} &\geq (1-\varepsilon)\sum_{\boldsymbol{\gamma}\in I}\left\|\int_{\mathcal{X}}\left(F_R-\mathcal{T}(\mathcal{L}(F_R))\right)\widehat{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}d\mu\right\|^2_{\mathcal{Y}}\\ &=(1-\varepsilon)\sum_{\boldsymbol{\gamma}\in I}\left\|x_{\boldsymbol{\gamma}}\frac{Y}{\|Y\|_{\mathcal{Y}}}-\int_{\mathcal{X}}\widetilde{\mathcal{T}}(\Xi_R(\boldsymbol{x}))\widehat{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}d\mu\right\|^2_{\mathcal{Y}}\\ &\geq(1-\varepsilon)\sum_{\boldsymbol{\gamma}\in I}\|Y\|_{\mathcal{Y}}^{-2}\left|\left\langle x_{\boldsymbol{\gamma}}\frac{Y}{\|Y\|_{\mathcal{Y}}}-\int_{\mathcal{X}}\widetilde{\mathcal{T}}(\Xi_R(\boldsymbol{x}))\widehat{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}d\mu,Y\right\rangle_{\mathcal{Y}}\right|^2\\ &\geq(1-\varepsilon)\sum_{\boldsymbol{\gamma}\in I}\left|x_{\boldsymbol{\gamma}}-\frac{1}{\|Y\|_{\mathcal{Y}}}\int_{\mathcal{X}}\left\langle\widetilde{\mathcal{T}}(\Xi_R(\boldsymbol{x})),Y\right\rangle_{\mathcal{Y}}\widehat{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}d\mu\right|^2.\end{aligned}$$

Finally, we define the (scalar-valued) reconstruction map

$$\Delta_R:\mathbb{R}^m\to\mathbb{R}^I,\quad \Delta_R(\boldsymbol{z}):=\left(\frac{1}{\|Y\|_{\mathcal{Y}}}\int_{\mathcal{X}}\left\langle\widetilde{\mathcal{T}}(\boldsymbol{z}),Y\right\rangle_{\mathcal{Y}}\widehat{H}_{\boldsymbol{\gamma},R,\boldsymbol{\lambda}}d\mu\right)_{\boldsymbol{\gamma}\in I},$$

and conclude

$$\|F_R-\mathcal{T}(\mathcal{L}(F_R))\|^2_{L^2_\mu(\mathcal{X};\mathcal{Y})}\geq(1-\varepsilon)\|\boldsymbol{x}-\Delta_R(\Xi_R(\boldsymbol{x}))\|^2_{\ell^2(I)}.$$

We have thus shown that for any pair $(\mathcal{L},\mathcal{T})$ of a Hilbert-valued adaptive sampling operator and a reconstruction map, the error $\|F_R-\mathcal{T}(\mathcal{L}(F_R))\|_{L^2_\mu(\mathcal{X};\mathcal{Y})}$ can be bounded from below by the error $(1-\varepsilon)\|\boldsymbol{x}-\Delta_R(\Xi_R(\boldsymbol{x}))\|_{\ell^2(I)}$ for some pair (Ξ_R,Δ_R) of a scalar-valued adaptive sampling operator and a (scalar-valued) reconstruction map. Consequently,

$$\begin{aligned}&\Theta_m(B_1^{\boldsymbol{b}}(\mathrm{Lip}(\mathcal{X},\mathcal{Y}));\mathcal{V},L^2_\mu(\mathcal{X};\mathcal{Y}))\\ &=\inf\left\{\sup_{F\in B_1^{\boldsymbol{b}}(\mathrm{Lip}(\mathcal{X},\mathcal{Y}))}\|F-\mathcal{T}(\mathcal{L}(F))\|_{L^2_\mu(\mathcal{X};\mathcal{Y})}:\mathcal{L}:\mathcal{V}\to\mathcal{Y}^m\text{ adaptive, }\mathcal{T}:\mathcal{Y}^m\to L^2_\mu(\mathcal{X};\mathcal{Y})\right\}\\ &\geq(1-\varepsilon)^{1/2}\inf\left\{\sup_{\substack{\boldsymbol{x}\in\mathbb{R}^I\\ \|\boldsymbol{x}\|_{\ell^2_{\boldsymbol{u}_I}(I)}\leq c}}\|\boldsymbol{x}-\Delta(\Xi(\boldsymbol{x}))\|_{\ell^2(I)}:\Xi:\mathbb{R}^I\to\mathbb{R}^m\text{ adaptive, }\Delta:\mathbb{R}^m\to\mathbb{R}^I\right\}\\ &=(1-\varepsilon)^{1/2}E^m_{\mathrm{ada}}(cB^2_{\boldsymbol{u}_I}(I),\ell^2(I))=(1-\varepsilon)^{1/2}E^m_{\mathrm{ada}}(B^2_{(c\boldsymbol{u})_I}(I),\ell^2(I)).\end{aligned}$$

As $\varepsilon>0$ was arbitrary, we can take the limit $\varepsilon\to0^+$. The claim now follows by Theorem 5.5. □

Finally, we can prove the desired lower bound for the adaptive m-width.

Proof of (5.6). Let $N \in \mathbb{N}$ with $N > m$, let $I = \pi([N]) = \{\boldsymbol{\pi(1)}, \ldots, \boldsymbol{\pi(N)}\} \subset \Gamma$ be the index set corresponding to the N largest entries of $\boldsymbol{u}$, and fix $c \in (0,1)$. By Theorem 5.7 and Lemma 5.8, we have

$$d^m(B^2_{(c\boldsymbol{u})_I}(I), \ell^2(I)) = d_m(B^2(I), \ell^2_{1/(c\boldsymbol{u})_I}(I)) = d_m(B^2_{(c\boldsymbol{u})_I}(I), \ell^2(I)) = c \cdot d_m(B^2_{\boldsymbol{u}_I}(I), \ell^2(I)) \tag{5.16}$$

For every $p > 2$ and $r = r(p) := 1/2 - 1/p$, Hölder's inequality implies $N^{-r} B^p_{\boldsymbol{u}_I}(I) \subset B^2_{\boldsymbol{u}_I}(I)$. Consequently,

$$d_m(B^2_{\boldsymbol{u}_I}(I), \ell^2(I)) \geq d_m(N^{-r} B^p_{\boldsymbol{u}_I}(I), \ell^2(I)) = N^{-r} d_m(B^p_{\boldsymbol{u}_I}(I), \ell^2(I)). \tag{5.17}$$

Applying Theorem 5.6 with $q = 2$ yields

$$d_m(B^p_{\boldsymbol{u}_I}(I), \ell^2(I)) = \left(\max_{\substack{i_1, \ldots, i_{N-m} \in I \\ i_k \neq i_j}} \left(\sum_{j=1}^{N-m} u_{i_j}^{\frac{2p}{p-2}} \right)^{\frac{1}{p} - \frac{1}{2}} \right)^{-1}.$$

Since $(u_{\boldsymbol{\pi(i)}})_{i \in \mathbb{N}}$ is nonincreasing, it follows with $q = q(p) := \frac{2p}{p-2} \in (2, \infty)$ that

$$d_m(B^p_{\boldsymbol{u}_I}(I), \ell^2(I)) = \min_{\substack{i_1, \ldots, i_{N-m} \in I \\ i_k \neq i_j}} \left(\sum_{j=1}^{N-m} u_{i_j}^{\frac{2p}{p-2}} \right)^{\frac{1}{2} - \frac{1}{p}} = \left(\sum_{j=m+1}^{N} u^q_{\boldsymbol{\pi(j)}} \right)^{1/q} \geq u_{\boldsymbol{\pi(m+1)}}.$$

We combine this estimate with (5.16), (5.17), and Lemma 5.9, and conclude

$$u_{\boldsymbol{\pi(m+1)}} \leq c^{-1} N^r \Theta_m(B^{\boldsymbol{b}}_1(\mathrm{Lip}(\mathcal{X}, \mathcal{Y})); \mathcal{V}, L^2_\mu(\mathcal{X}; \mathcal{Y})).$$

Taking the limit $p \to 2^+$ yields $r \to 0^+$ and therefore

$$u_{\boldsymbol{\pi(m+1)}} \leq c^{-1} \Theta_m(B^{\boldsymbol{b}}_1(\mathrm{Lip}(\mathcal{X}, \mathcal{Y})); \mathcal{V}, L^2_\mu(\mathcal{X}; \mathcal{Y})).$$

As $c \in (0,1)$ was arbitrary, we can take the limit $c \to 1^-$, and the claim finally follows. □

5.3.3 Upper bound

We now prove the upper bound for the adaptive m-width, that is,

$$\begin{aligned} \Theta_m(\mathcal{K}; \mathcal{V}, L^2_\mu(\mathcal{X}; \mathcal{Y})) &\leq \inf_{S \subset \Gamma, |S| \leq m} \sup_{F \in \mathcal{K}} \|F - F_S\|_{L^2_\mu(\mathcal{X}; \mathcal{Y})} \\ &\leq \sup_{F \in \mathcal{K}} \left\| F - F_{\{\boldsymbol{\pi(1)}, \ldots, \boldsymbol{\pi(m)}\}} \right\|_{L^2_\mu(\mathcal{X}; \mathcal{Y})} \leq u_{\boldsymbol{\pi(m+1)}} \end{aligned} \tag{5.18}$$

for $\mathcal{K} \in \{B_1(W^{1,2}_{\mu, \boldsymbol{b}}(\mathcal{X}; \mathcal{Y})), B^{\boldsymbol{b}}_1(\mathrm{Lip}(\mathcal{X}, \mathcal{Y}))\}$. For this, we assume that $\mathcal{V}$ is continuously embedded in $L^2_\mu(\mathcal{X}; \mathcal{Y})$.

Proof of (5.18). The second and third inequality hold by (4.2), so we only need to prove the first inequality. We fix $m \in \mathbb{N}$ and $S = \{\boldsymbol{\gamma^{(1)}}, \ldots, \boldsymbol{\gamma^{(n)}}\} \subset \Gamma$ with $n \leq m$. We define the adaptive sampling operator

$$\mathcal{L} : \mathcal{V} \to \mathcal{Y}^m, \quad \mathcal{L}_i(F) := \begin{cases} \int_{\mathcal{X}} F H_{\boldsymbol{\gamma^{(i)}}, \boldsymbol{\lambda}} d\mu & \text{if } 1 \leq i \leq n, \\ 0 & \text{if } n+1 \leq i \leq m, \end{cases}$$

and the reconstruction map

$$\mathcal{T} : \mathcal{Y}^m \to L^2_\mu(\mathcal{X}; \mathcal{Y}), \quad \mathcal{T}(\boldsymbol{Y}) := \sum_{i=1}^m Y_i H_{\boldsymbol{\pi(i)}, \boldsymbol{\lambda}}.$$

Since $\mathcal{V}$ is continuously embedded in $L^2_\mu(\mathcal{X}; \mathcal{Y})$, it is easy to see that $\mathcal{L}$ is a well-defined bounded linear operator. We need to show that it satisfies the conditions in Definition 5.2. It suffices to show that there exists $Y \in \mathcal{Y} \setminus \{0\}$, a normed vector space $\widetilde{\mathcal{V}} \subset L^2_\mu(\mathcal{X})$, and a scalar-valued adaptive sampling operator $\widetilde{\mathcal{L}} : \widetilde{\mathcal{V}} \to \mathcal{Y}^m$ such that, if $YF \in \mathcal{V}$ for some $F \in L^2_\mu(\mathcal{X})$, then $F \in \widetilde{\mathcal{V}}$ and $\mathcal{L}(YF) = Y\widetilde{\mathcal{L}}(F)$. To this end, we choose some $Y \in \mathcal{Y}$ with $\|Y\|_{\mathcal{Y}} = 1$ and define the space

$$\widetilde{\mathcal{V}} := \left\{F \in L^2_\mu(\mathcal{X}) : YF \in \mathcal{V}\right\}.$$

It can be readily checked that this defines a normed vector space with norm given by $\|F\|_{\widetilde{\mathcal{V}}} := \|YF\|_{\mathcal{V}}$ for any $F \in \widetilde{\mathcal{V}}$. Moreover, as $\mathcal{V}$ is continuously embedded in $L^2_\mu(\mathcal{X}; \mathcal{Y})$, there exists a constant $C > 0$ such that

$$\|F\|_{L^2_\mu(\mathcal{X})} = \|YF\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \leq C\|YF\|_{\mathcal{V}} = C\|F\|_{\widetilde{\mathcal{V}}}, \quad \forall F \in \widetilde{\mathcal{V}},$$

where in the first step we used the fact that $\|Y\|_{\mathcal{Y}} = 1$. This shows that $\widetilde{\mathcal{V}}$ is continuously embedded in $L^2_\mu(\mathcal{X})$. We now define the operator

$$\widetilde{\mathcal{L}} : \widetilde{\mathcal{V}} \to \mathbb{R}^m, \quad \widetilde{\mathcal{L}}_i(F) := \begin{cases} \int_{\mathcal{X}} F H_{\boldsymbol{\gamma}^{(i)},\boldsymbol{\lambda}} d\mu & \text{if } 1 \leq i \leq n, \\ 0 & \text{if } n+1 \leq i \leq m. \end{cases}$$

Note that $\widetilde{\mathcal{L}}$ is linear and by the continuous embedding of $\widetilde{\mathcal{V}}$ in $L^2_\mu(\mathcal{X})$, it is also bounded. In particular, $\widetilde{\mathcal{L}}$ is a scalar-valued adaptive sampling operator. Moreover, by construction, if $YF \in \mathcal{V}$, then $F \in \widetilde{\mathcal{V}}$ and $\mathcal{L}(YF) = Y\widetilde{\mathcal{L}}(F)$. Hence, $\mathcal{L}$ is indeed an adaptive (Hilbert-valued) sampling operator as in Definition 5.2. Consequently,

$$\Theta_m(\mathcal{K}; \mathcal{V}, L^2_\mu(\mathcal{X}; \mathcal{Y})) \leq \sup_{F\in\mathcal{K}} \|F - \mathcal{T}(\mathcal{L}(F))\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} = \sup_{F\in\mathcal{K}} \|F - F_S\|_{L^2_\mu(\mathcal{X};\mathcal{Y})}.$$

As S was arbitrary, we can now take the infimum over all subsets $S \subset \Gamma$ with $|S| \leq m$ and conclude the claim. □

5.4 Discussion

Note that Theorem 5.4 shows that linear Hermite polynomial approximation based on the index set $S = \{\boldsymbol{\pi}(\mathbf{1}), \ldots, \boldsymbol{\pi}(\boldsymbol{m})\}$ is optimal among all possible recovery strategies which are based on linear (adaptive) information for the uniform approximation of $W^{1,2}_{\mu,\boldsymbol{b}}$- and Lipschitz operators with Sobolev norm at most one.

Moreover, Theorem 5.4 in combination with Theorem 4.6 yields the following **curse of sample complexity**: *No strategy based on finitely many (potentially adaptively chosen) linear samples for the uniform recovery of all operators in the Sobolev unit (Lipschitz) ball can achieve algebraic convergence rates. This holds regardless of the decay rate of the PCA eigenvalues of the covariance operator of the underlying Gaussian measure.*

As already mentioned in the introduction, a related result was previously shown in [40] by means of the so-called sampling nonlinear m-width $s_m(\mathcal{K})_{L^2_\mu(\mathcal{X})}$ of a set $\mathcal{K} \subset L^2_\mu(\mathcal{X})$. The latter is based on standard information. More specifically, compared to the definition of the adaptive m-width in (5.3), the sampling operator $\delta_{\boldsymbol{X}} : \mathcal{K} \to \mathcal{Y}^m$ with $\boldsymbol{X} = (X_1, \ldots, X_m) \in \mathcal{X}^m$ is given by point evaluation at fixed sample points $X_1, \ldots, X_m$, that is, $\delta_{\boldsymbol{X}}(F) = (F(X_1), \ldots, F(X_m)) \in \mathcal{Y}^m$ for every $F \in \mathcal{K}$, and one defines

$$s_m(\mathcal{K})_{L^2_\mu(\mathcal{X})} := \inf\left\{\sup_{F\in\mathcal{K}} \|F - \mathcal{T}(\delta_{\boldsymbol{X}}(F))\|_{L^2_\mu(\mathcal{X})} : \boldsymbol{X} \in \mathcal{X}^m, \mathcal{T} : \mathcal{Y}^m \to L^2_\mu(\mathcal{X})\right\}.$$

Observe that Theorem 2.12 in [40] implies the following result, which was termed the *curse of data complexity*:

Theorem 5.13. *Let μ be a centered Gaussian measure with at most algebraically decreasing (unweighted) PCA eigenvalues $\lambda_i \gtrsim i^{-\alpha}$ of the covariance operator for some $\alpha > 0$. Then there exists a constant $C > 0$ such that*

$$s_m(\mathrm{Lip}(\mathcal{X}))_{L^2_\mu(\mathcal{X})} \geq C \log(m)^{-(\alpha+3)}, \quad \forall m \in \mathbb{N}.$$

Our findings in the present section generalize this result in several directions. First, the adaptive m-width covers recovery based on general linear (adaptive) information, not just standard information. Second, its tight characterization by Theorem 5.4 pertains to *general* centered, nondegenerate Gaussian measures. In addition, we again highlight that Theorem 4.7 provides upper bounds for the adaptive m-width of the Sobolev unit (Lipschitz) ball in terms of the decay of the PCA eigenvalues $\lambda_{\boldsymbol{b},i}$. In particular, the curse of sample complexity described above can be overcome asymptotically in the sense that in the large data limit $m \to \infty$, the adaptive m-width can decay with rates which are arbitrarily close to any algebraic rate if the decay of the $\lambda_{\boldsymbol{b},i}$ is double-exponential.

6 Constructive near-optimal pointwise sampling

Given an operator F in the Sobolev unit (Lipschitz) ball, constructing the optimal (in the sense of Theorem 5.4) polynomial approximant $F_{\pi([s])}$ requires access to the Wiener-Hermite PC coefficients $\int_{\mathcal{X}} FH_{\boldsymbol{\pi(i)},\boldsymbol{\lambda}}d\mu$ for all $i \in [s]$. In practice, however, this data is typically not available. Instead, one often relies on nonintrusive measurements which generate *random standard information*, that is, evaluations of F at points which are (assumed to be) independently and identically distributed with respect to some probability measure. In this section, based on our previous results, we derive algorithms to reconstruct $W^{1,2}_{\mu,\boldsymbol{b}}$- and Lipschitz operators from i.i.d. point samples with near-optimal sample complexity (up to logarithmic and subalgebraic factors) in high probability. The key tool for this is *Christoffel sampling*, that is, we construct a sampling measure based on the Christoffel function of the problem, which we independently draw samples from. Reconstruction of the target operator is then done via a weighted least-squares fit. This strategy is due to [21] and it is closely related to leverage score sampling in data science. See [1] for a review.

In Subsection 6.1, standard results from least-squares approximation are recalled. We prove sample complexity estimates in probability for L^2_μ-, $W^{1,2}_{\mu,\boldsymbol{b}}$-, and Lipschitz operators in Subsection 6.2 and present our algorithms in Subsection 6.3.

6.1 Least-squares: Preliminaries

We recall some important notions from (weighted) least-squares approximation, see also [6, Chpt. 5]. For a set $S \subset \Gamma$ of size $|S| = s$, we first recall from Section 4 the polynomial space

$$\mathcal{P}_{S;\mathcal{Y}} := \left\{ \sum_{\boldsymbol{\gamma}\in S} Y_{\boldsymbol{\gamma}} H_{\boldsymbol{\gamma},\boldsymbol{\lambda}} : Y_{\boldsymbol{\gamma}} \in \mathcal{Y} \right\} \subset L^2_\mu(\mathcal{X};\mathcal{Y})$$

and the corresponding orthogonal L^2_μ-projection

$$(\cdot)_S : L^2_\mu(\mathcal{X};\mathcal{Y}) \to \mathcal{P}_{S;\mathcal{Y}}, \quad F \mapsto F_S := \sum_{\boldsymbol{\gamma}\in S} \left(\int_{\mathcal{X}} FH_{\boldsymbol{\gamma},\boldsymbol{\lambda}} d\mu \right) H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}.$$

We henceforth assume that we are given m distinct sample points $X_1, \ldots, X_m \in \mathcal{X}$ with $m \geq s$. Next, let $w : \mathcal{X} \to (0, \infty)$ be a positive weight function whose reciprocal is a probability density with respect to μ, i.e., $\int_{\mathcal{X}} w(X)^{-1} d\mu(X) = 1$. The corresponding probability measure on $\mathcal{X}$ is given by

$$d\nu(X) := w(X)^{-1} d\mu(X).$$

As we will draw the sample points X_i in $\mathcal{X}$ with respect to ν, we call ν the *sampling measure.* The weight function w is also used to define the *(weighted) discrete semi-norm*

$$\|F\|^2_{\mathrm{disc},w} := \frac{1}{m} \sum_{i=1}^m w(X_i) \|F(X_i)\|^2_{\mathcal{Y}}, \quad \forall F \in L^2_\mu(\mathcal{X};\mathcal{Y}).$$

For fixed $F \in L^2_\mu(\mathcal{X};\mathcal{Y})$, it is well-defined μ- and hence ν-almost surely because the point evaluations $F(X_1),\dots,F(X_m)$ are well-defined μ-almost surely. In the scalar-valued case $\mathcal{Y}=\mathbb{R}$, the corresponding *(weighted) discrete stability constant* of the space

$$\mathcal{P}_S := \mathcal{P}_{S;\mathbb{R}} = \operatorname{span}\{H_{\boldsymbol{\gamma},\boldsymbol{\lambda}} : \boldsymbol{\gamma} \in S\} \subset L^2_\mu(\mathcal{X}) \tag{6.1}$$

is given by

$$\alpha_w = \alpha_w(\mathcal{P}_S) := \inf\left\{ \|p\|_{\mathrm{disc},w} : p \in \mathcal{P}_S,\ \|p\|_{L^2_\mu(\mathcal{X})} = 1 \right\}. \tag{6.2}$$

The *(reciprocal) Christoffel function* of $\mathcal{P}_S$ is defined as

$$K(\mathcal{P}_S) := \sum_{\boldsymbol{\gamma}\in S} |H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}|^2. \tag{6.3}$$

It yields an upper bound on the polynomial approximant F_S:

Lemma 6.1 (Bound on F_S)**.** *Let $S \subset \Gamma$ be finite and $F \in L^2_\mu(\mathcal{X};\mathcal{Y})$. We have*

$$\|F_S(X)\|_{\mathcal{Y}} \le \|F\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \sqrt{K(\mathcal{P}_S)(X)} \quad \text{for } \mu\text{-a.e. } X \in \mathcal{X}.$$

Proof. Suppose that $\|F\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} > 0$. Otherwise there is nothing to show. Let us write $Y_{\boldsymbol{\gamma}} := \int_{\mathcal{X}} F H_{\boldsymbol{\gamma},\boldsymbol{\lambda}} d\mu \in \mathcal{Y}$ for $\boldsymbol{\gamma} \in \Gamma$. By Parseval's identity, we have

$$\sum_{\boldsymbol{\gamma}\in S} \frac{\|Y_{\boldsymbol{\gamma}}\|^2_{\mathcal{Y}}}{\|F\|^2_{L^2_\mu(\mathcal{X};\mathcal{Y})}} + \sum_{\boldsymbol{\gamma}\notin S} \frac{\|Y_{\boldsymbol{\gamma}}\|^2_{\mathcal{Y}}}{\|F\|^2_{L^2_\mu(\mathcal{X};\mathcal{Y})}} = 1.$$

For brevity, we set $a := \sum_{\boldsymbol{\gamma}\notin S} \|Y_{\boldsymbol{\gamma}}\|^2_{\mathcal{Y}} / \|F\|^2_{L^2_\mu(\mathcal{X};\mathcal{Y})}$. We can now apply Jensen's inequality to find

$$\begin{aligned}
\|F_S(X)\|_{\mathcal{Y}} &\le \sum_{\substack{\boldsymbol{\gamma}\in S\\ Y_{\boldsymbol{\gamma}}\neq 0}} \|Y_{\boldsymbol{\gamma}}\|_{\mathcal{Y}} |H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(X)| \le \left[\left(\sum_{\substack{\boldsymbol{\gamma}\in S\\ Y_{\boldsymbol{\gamma}}\neq 0}} \frac{\|Y_{\boldsymbol{\gamma}}\|^2_{\mathcal{Y}}}{\|F\|^2_{L^2_\mu(\mathcal{X};\mathcal{Y})}} \frac{\|F\|^2_{L^2_\mu(\mathcal{X};\mathcal{Y})}}{\|Y_{\boldsymbol{\gamma}}\|_{\mathcal{Y}}} |H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(X)| + a\cdot 0 \right)^2 \right]^{1/2} \\
&\le \left(\sum_{\substack{\boldsymbol{\gamma}\in S\\ Y_{\boldsymbol{\gamma}}\neq 0}} \frac{\|Y_{\boldsymbol{\gamma}}\|^2_{\mathcal{Y}}}{\|F\|^2_{L^2_\mu(\mathcal{X};\mathcal{Y})}} \frac{\|F\|^4_{L^2_\mu(\mathcal{X};\mathcal{Y})}}{\|Y_{\boldsymbol{\gamma}}\|^2_{\mathcal{Y}}} |H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(X)|^2 + a\cdot 0^2 \right)^{1/2} \\
&\le \|F\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \left(\sum_{\boldsymbol{\gamma}\in S} |H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(X)|^2 \right)^{1/2} = \|F\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \sqrt{K(\mathcal{P}_S)(X)}
\end{aligned}$$

for μ-a.e. $X \in \mathcal{X}$. □

Next, let $F \in L^2_\mu(\mathcal{X};\mathcal{Y})$ be a given operator and let $X_1,\dots,X_m \in \mathcal{X}$ be sample points such that the point evaluations $F(X_1),\dots,F(X_m) \in \mathcal{Y}$ are well-defined. We then define an approximant $\widehat{F}$ of F via a weighted least-squares fit,

$$\widehat{F} = \widehat{F}(X_1,\dots,X_m) \in \underset{P\in\mathcal{P}_{S;\mathcal{Y}}}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^m w(X_i) \|P(X_i) - F(X_i)\|^2_{\mathcal{Y}}. \tag{6.4}$$

Note that for every fixed $F \in L^2_\mu(\mathcal{X};\mathcal{Y})$, the loss function in (6.4) and therefore each of its minimizers $\widehat{F}$ (if there are any) are well-defined μ-almost surely and hence also ν-almost surely since pointwise evaluations of F are well-defined μ- and ν-almost surely.

The problem (6.4) can be reformulated as an algebraic least-squares problem. To this end, we introduce the weighted (normalized) measurement matrix and measurement vector

$$\mathbb{A} := \left(\frac{\sqrt{w(X_i)}}{\sqrt{m}} H_{\boldsymbol{\gamma}_j,\boldsymbol{\lambda}}(X_i) \right)_{(i,j)\in[m]\times[s]} \in \mathbb{C}^{m\times s}, \quad \boldsymbol{B} := \left(\frac{\sqrt{w(X_i)}}{\sqrt{m}} F(X_i) \right)_{i\in[m]} \in \mathbb{C}^m,$$

and the associated bounded linear operator

$$T_{\mathbb{A}} : \mathcal{Y}^s \to \mathcal{Y}^m, \quad \boldsymbol{Y} = (Y_j)_{j=1}^s \mapsto \left(\sum_{j=1}^s \mathbb{A}_{ij} Y_j \right)_{i \in [m]}.$$

It is an easy exercise to check that (6.4) is equivalent to

$$\boldsymbol{Y} \in \underset{\boldsymbol{Z} \in \mathcal{Y}^s}{\operatorname{argmin}} \|T_{\mathbb{A}} \boldsymbol{Z} - \boldsymbol{B}\|_{\mathcal{Y}^m}, \tag{6.5}$$

where we use the notation $\|\cdot\|_{\mathcal{Y}^m} := \|\cdot\|_{\ell^2([m];\mathcal{Y})}$. More precisely, any solution of (6.5) yields the polynomial coefficients of a solution $\widehat{F}$ of (6.4) and vice versa. Lemma 6.2 below shows that (6.5), in fact, has a unique solution if α_w is positive.

Lifting to Hilbert spaces

The least-squares problems (6.4), (6.5) are not quite standard, in that they involve operators and vectors, respectively, which take values in a generic separable Hilbert space $\mathcal{Y}$. Fortunately, many of the theoretical tools in the scalar-valued case can be "lifted" to Hilbert spaces, see [3, Sect. 6.2]. The next result is an instance of this lifting concept:

Lemma 6.2. *The Hilbert-valued algebraic least-squares problem* (6.5) *has a unique solution if and only if the discrete stability constant α_w, defined in* (6.2)*, is positive.*

Proof. We commence by fixing an orthonormal basis $\{\psi_k\}_{k \in \mathbb{N}}$ of $\mathcal{Y}$ and reduce the problem to the scalar-valued case. By Parseval's identity, we have

$$\|T_{\mathbb{A}} \boldsymbol{Z} - \boldsymbol{B}\|_{\mathcal{Y}^m}^2 = \sum_{k=1}^{\infty} \left\| \mathbb{A} \boldsymbol{z}^{(\boldsymbol{k})} - \boldsymbol{b}^{(\boldsymbol{k})} \right\|_{\mathbb{R}^m}^2, \quad \forall \boldsymbol{Z} \in \mathcal{Y}^s, \tag{6.6}$$

with $\boldsymbol{z}^{(\boldsymbol{k})} := \langle \boldsymbol{Z}, \psi_k \rangle_{\mathcal{Y}} \in \mathbb{R}^s$ and $\boldsymbol{b}^{(\boldsymbol{k})} := \langle \boldsymbol{B}, \psi_k \rangle_{\mathcal{Y}} \in \mathbb{R}^m$, $k \in \mathbb{N}$. Consequently, $\boldsymbol{Z} \in \mathcal{Y}^s$ is a minimizer of the left-hand side in (6.6) if and only if $\|\mathbb{A} \boldsymbol{z}^{(\boldsymbol{k})} - \boldsymbol{b}^{(\boldsymbol{k})}\|_{\mathbb{R}^m}^2$ is minimal for every k. The scalar-valued least-squares problem

$$\boldsymbol{y}^{(\boldsymbol{k})} \in \underset{\boldsymbol{z} \in \mathbb{R}^s}{\operatorname{argmin}} \left\| \mathbb{A} \boldsymbol{z} - \boldsymbol{b}^{(\boldsymbol{k})} \right\|_{\mathbb{R}^m}$$

always has a solution given by $\boldsymbol{y}^{(\boldsymbol{k})} = (y_1^{(k)}, \dots, y_s^{(k)}) := \mathbb{A}^\dagger \boldsymbol{b}^{(\boldsymbol{k})}$ for every $k \in \mathbb{N}$, where $\mathbb{A}^\dagger$ denotes the pseudoinverse of $\mathbb{A}$. By standard least-squares theory, this solution is unique if and only if $\alpha_w > 0$, see [6, Chpt. 5.2 & Chpt. 5.5.1]. We now define

$$\boldsymbol{Y} = (\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_s) \in \mathcal{Y}^s \quad \text{with} \quad \boldsymbol{Y}_j := \sum_{k=1}^{\infty} y_j^{(k)} \psi_k \in \mathcal{Y}, \quad \forall j \in [s].$$

To show that every $\boldsymbol{Y}_j$ is indeed an element in $\mathcal{Y}$, note that, by definition, we have

$$\sum_{k=1}^{\infty} \left(y_j^{(k)} \right)^2 \leq m \|\mathbb{A}^\dagger\|_F^2 \|\boldsymbol{B}\|_{\mathcal{Y}^m}^2 < \infty,$$

where $\|\mathbb{A}\|_F := \sqrt{\operatorname{trace}(\mathbb{A}^* \mathbb{A})}$ denotes the Frobenius norm. This concludes the proof. □

A second important property which we can lift from the scalar- to the Hilbert-valued case is the following inequality, which holds by definition of α_w:

$$\alpha_w \|p\|_{L^2_\mu(\mathcal{X})}^2 \leq \frac{1}{m} \sum_{i=1}^m w(X_i) \, |p(X_i)|^2, \quad \forall p \in \mathcal{P}_S. \tag{6.7}$$

This is often referred to as a (lower) Marcinkiewicz-Zygmund inequality for $\mathcal{P}_S$, see [70, 39]. The following result can be proven similarly as [5, Lem. 7.5].

Lemma 6.3 (Lower Marcinkiewicz-Zygmund inequality; Hilbert-valued case)**.** *Inequality* (6.7) *is equivalent to a lower Marcinkiewicz-Zygmund inequality for $\mathcal{P}_{S;\mathcal{Y}}$, namely*

$$\alpha_w \|P\|^2_{L^2_\mu(\mathcal{X};\mathcal{Y})} \leq \frac{1}{m}\sum_{i=1}^m w(X_i)\|P(X_i)\|^2_{\mathcal{Y}}, \quad \forall P \in \mathcal{P}_{S;\mathcal{Y}}.$$

6.2 Sample complexities

With the tools from the previous section, we now prove near-optimal sample complexity rates for the approximation of L^2_μ-, $W^{1,2}_{\mu,\boldsymbol{b}}$-, and Lipschitz operators F via the least-squares approximant $\widehat{F}$, as defined in (6.4). For this, we construct a suitable sampling measure ν based on the reciprocal Christoffel function $K(\mathcal{P}_S)$. In the case of L^2_μ-operators, our results hold for any finite index set $S \subset \Gamma$ of size s for which $K(\mathcal{P}_S) > 0$. We remark that the positivity constraint can be avoided by defining the weight function w in a slightly different way, see [1, Eq. (6.2)]. However, for ease of presentation, we stick with the mild constraint $K(\mathcal{P}_S) > 0$. For $W^{1,2}_{\mu,\boldsymbol{b}}$- and Lipschitz operators, we make a specific choice for S so that $\mathbf{0} \in S$ and therefore

$$K(\mathcal{P}_S)(X) \geq |H_{\mathbf{0},\boldsymbol{\lambda}}(X)|^2 = 1, \quad \forall X \in \mathcal{X}. \tag{6.8}$$

As it will repeatedly appear in our estimates below, we introduce the universal constant

$$c := 2\left(\log(1/2)+1\right)^{-1} \approx 6.518. \tag{6.9}$$

6.2.1 L^2_μ-operators

We fix a finite index set $S \subset \Gamma$ of size s such that $K(\mathcal{P}_S) > 0$ and define

$$w(X) := \left(\frac{1}{s}K(\mathcal{P}_S)(X)\right)^{-1} = \left(\frac{1}{s}\sum_{\boldsymbol{\gamma}\in S}|H_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(X)|^2\right)^{-1}, \quad \forall X \in \mathcal{X}. \tag{6.10}$$

This is indeed a probability density with respect to the measure μ because the Hermite polynomials are orthonormal in $L^2_\mu(\mathcal{X})$.

Theorem 6.4 (Near-optimal sampling for L^2_μ-operators in probability)**.** *Let $0 < \epsilon < 1$ denote the failure probability and suppose that $F \in L^2_\mu(\mathcal{X};\mathcal{Y})$. Let $X_1, \ldots, X_m \in \mathcal{X}$ be drawn independently from the sampling measure ν, where $d\nu = w^{-1}d\mu$ with w as in* (6.10)*. Suppose that m satisfies*

$$m \geq cs\log(s/\epsilon) \tag{6.11}$$

with c given by (6.9)*. Then, with ν-probability at least $1-\epsilon$, the weighted least-squares approximant $\widehat{F}$ in* (6.4) *is unique and well-defined and satisfies*

$$\left\|F - \widehat{F}\right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \leq \left(1 + \frac{2\sqrt{2}}{\sqrt{\epsilon}}\right)\|F - F_S\|_{L^2_\mu(\mathcal{X};\mathcal{Y})}. \tag{6.12}$$

Proof. Let $F \in L^2_\mu(\mathcal{X};\mathcal{Y})$ and suppose that $\alpha_w > 0$. Then, by Lemma 6.2, $\widehat{F}$ is unique and well-defined μ- and ν-almost surely. Using Lemma 6.3, we can generalize the standard arguments in the proof of [6, Thm. 5.3] to the Hilbert-valued case to obtain

$$\left\|F - \widehat{F}\right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \leq \|F - F_S\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} + \alpha_w^{-1}\|F - F_S\|_{\mathrm{disc},w} \quad \nu\text{-almost surely.}$$

It thus suffices to bound α_w from below away from zero with high probability and to estimate $\|F - F_S\|_{\mathrm{disc},w}$ accordingly. To this end, let $0 < \epsilon < 1$. Standard arguments, based on the matrix Chernoff bound, see, e.g., [6, Thm. 5.8], yield

$$\mathbb{P}_\nu\left[\alpha_w \leq 1/2\right] \leq \epsilon/2 \quad \text{if} \quad m \geq cs\log(s/\epsilon) \tag{6.13}$$

with c given by (6.9). We refer to the proof of [6, Thm. 5.19]) (with $\delta = 1/2$) for further details. Next, observe that

$$\mathbb{E}_\nu\left[\|F - F_S\|^2_{\mathrm{disc},w}\right] = \|F - F_S\|^2_{L^2_\mu(\mathcal{X};\mathcal{Y})}.$$

Hence, by Markov's inequality, we get

$$\mathbb{P}_\nu\left[\|F - F_S\|_{\mathrm{disc},w} > \frac{\|F - F_S\|_{L^2_\mu(\mathcal{X};\mathcal{Y})}}{\sqrt{\epsilon/2}}\right] \leq \frac{\mathbb{E}_\nu\left[\|F - F_S\|^2_{\mathrm{disc},w}\right]}{\|F - F_S\|^2_{L^2_\mu(\mathcal{X};\mathcal{Y})}/(\epsilon/2)} = \frac{\epsilon}{2}. \tag{6.14}$$

We now combine (6.13) and (6.14) and apply the union bound to conclude the claim. □

6.2.2 Sobolev and Lipschitz operators

We now make a specific choice for the index set S, namely

$$S = \pi([s]) = \{\boldsymbol{\pi}(\mathbf{1}), \ldots, \boldsymbol{\pi}(\boldsymbol{s})\}.$$

In particular, $\boldsymbol{\pi}(\mathbf{1}) = \mathbf{0} \in \pi([s])$ and therefore (6.8) is satisfied. The following result addresses the sample complexity for $W^{1,2}_{\mu,\boldsymbol{b}}$-operators. It is an immediate consequence of Theorem 6.4 and (4.1), (4.2).

Corollary 6.5 (Near-optimal sampling for $W^{1,2}_{\mu,\boldsymbol{b}}$-operators in probability)**.** *Let $0 < \epsilon < 1$ denote the failure probability and suppose that $F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$. Let $X_1, \ldots, X_m \in \mathcal{X}$ be drawn independently from the sampling measure ν, where $d\nu = w^{-1}d\mu$ with w as in* (6.10)*. Suppose that m satisfies*

$$m \geq cs\log(s/\epsilon) \tag{6.15}$$

with c given by (6.9)*. Then, with ν-probability at least $1-\epsilon$, the weighted least-squares approximant $\widehat{F}$ in* (6.4) *is unique and well-defined and satisfies*

$$\left\|F - \widehat{F}\right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \leq \left(1 + \frac{2\sqrt{2}}{\sqrt{\epsilon}}\right) u_{\boldsymbol{\pi}(\boldsymbol{s}+\mathbf{1})} \|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})}. \tag{6.16}$$

Note that the bound on the approximation error in (6.16) has poor scaling in the reciprocal of the failure probability ϵ. This can be removed if we restrict to Lipschitz operators under a mild alteration of the sampling measure ν. However, this leads a to a slightly worse sample complexity which is linear in s up to a log-factor and an additional subalgebraic term. Note that for Lipschitz operators F, each least-squares approximant $\widehat{F}$ (if there are any) is always well-defined, not only almost surely.

We proceed by defining a new weight function

$$w(X) := \left(\frac{1}{s+1}\left(\|X\|^2_{\mathcal{X}} + K(\mathcal{P}_{\pi([s])})(X)\right)\right)^{-1} = \left(\frac{1}{s+1}\left(\|X\|^2_{\mathcal{X}} + \sum_{i=1}^{s}\left|H_{\boldsymbol{\pi}(\boldsymbol{i}),\boldsymbol{\lambda}}(X)\right|^2\right)\right)^{-1}, \quad \forall X \in \mathcal{X}. \tag{6.17}$$

Observe that it differs from the weight function in (6.10) only by the additional additive term $\|X\|^2_{\mathcal{X}}$ (and the corresponding normalizing factor). It is again a well-defined probability measure by our assumption $\sum_{i\in\mathbb{N}} \lambda_i = 1$ which implies $\int_{\mathcal{X}} \|X\|^2_{\mathcal{X}} d\mu(X) = 1$.

Theorem 6.6 (Near-optimal sampling for Lipschitz operators in probability)**.** *Let $0 < \epsilon < 1$ denote the failure probability and suppose that $F \in \mathrm{Lip}(\mathcal{X}, \mathcal{Y})$ with Lipschitz constant $L > 0$. Let $X_1, \dots, X_m \in \mathcal{X}$ be drawn independently from ν, where $d\nu = w^{-1} d\mu$ with w as in* (6.17)*. Suppose that m satisfies*

$$m \geq C s u_{\boldsymbol{\pi(s+1)}}^{-2} \log(4s/\epsilon), \qquad C := \max\left\{8\left(\|F(0)\|_{\mathcal{Y}} + L\right)^2, c\right\}, \tag{6.18}$$

where c is given by (6.9)*. Then, with ν-probability at least $1-\epsilon$, the weighted least-squares approximant $\widehat{F}$ in* (6.4) *is unique and satisfies*

$$\left\|F - \widehat{F}\right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \leq \sqrt{2} u_{\boldsymbol{\pi(s+1)}} \left(\|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} + 1\right).$$

The proof is based on the following result which follows from Bernstein's inequality for bounded random variables, see, e.g., [6, Lemma 7.18]. It can be proven the same way as [6, Lemma 7.11(ii)], we only remark that $\|\sqrt{w}F\|_{L^2_\nu(\mathcal{X};\mathcal{Y})} = \|F\|_{L^2_\mu(\mathcal{X};\mathcal{Y})}$ and $\|\sqrt{w}F\|_{L^\infty_\nu(\mathcal{X};\mathcal{Y})} = \|\sqrt{w}F\|_{L^\infty_\mu(\mathcal{X};\mathcal{Y})}$ for every $F \in L^2_\mu(\mathcal{X};\mathcal{Y})$.

Lemma 6.7 (Bound on the weighted discrete approximation error)**.** *Let $F \in L^2_\mu(\mathcal{X};\mathcal{Y})$ and $S \subset \Gamma$ be finite. Suppose that $X_1, \dots, X_m \in \mathcal{X}$ are drawn independently from ν, where $d\nu = w^{-1} d\mu$ with w^{-1} as in* (6.17)*. Then, for any $0 < \epsilon < 1$ and any $k \in (0, \infty)$, we have*

$$\|F - F_S\|_{\mathrm{disc},w} \leq \sqrt{2}\left(\|F - F_S\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} + \frac{\|\sqrt{w}\,(F - F_S)\|_{L^\infty_\mu(\mathcal{X};\mathcal{Y})}}{\sqrt{k}}\right)$$

with ν-probability at least $1-\epsilon$, provided that $m \geq 2k \log(2/\epsilon)$.

Proof of Theorem 6.6. We fix $0 < \epsilon < 1$ and a Lipschitz operator $F : \mathcal{X} \to \mathcal{Y}$ with Lipschitz constant $L > 0$. We proceed as in the proof of Theorem 6.4 to derive the bound

$$\left\|F - \widehat{F}\right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \leq \left\|F - F_{\pi([s])}\right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} + \alpha_w^{-1}\left\|F_{\pi([s])} - F\right\|_{\mathrm{disc},w},$$

and to show that

$$\mathbb{P}_\nu\left[\alpha_w > 1/2\right] \leq \epsilon/2 \quad \text{if} \quad m \geq cs\log(2s/\epsilon), \tag{6.19}$$

with c given by (6.9). Next, we use Lemma 6.7 to bound $\|F_{\pi([s])} - F\|_{\mathrm{disc},w}$. First note that by Lipschitz continuity we have

$$\|F(X)\|_{\mathcal{Y}} \leq \|F(0)\|_{\mathcal{Y}} + L\|X\|_{\mathcal{X}}, \quad \forall X \in \mathcal{X}. \tag{6.20}$$

Together with (6.8), we conclude for any $X \in \mathcal{X}$,

$$\begin{aligned} \left\|\sqrt{w(X)}F(X)\right\|_{\mathcal{Y}} &= \frac{\sqrt{s+1}}{\sqrt{\|X\|_{\mathcal{X}}^2 + K(\mathcal{P}_{\pi([s])})(X)}}\|F(X)\|_{\mathcal{Y}} \\ &\leq \frac{\sqrt{s+1}}{\sqrt{\|X\|_{\mathcal{X}}^2 + 1}}\left(\|F(0)\|_{\mathcal{Y}} + L\|X\|_{\mathcal{X}}\right) \\ &\leq \sqrt{s+1}\left(\|F(0)\|_{\mathcal{Y}} + L\right). \end{aligned} \tag{6.21}$$

Moreover, it follows from (6.20) and $\int_{\mathcal{X}} \|X\|_{\mathcal{X}}^2 d\mu(X) = 1$ that $\|F\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \leq \|F(0)\|_{\mathcal{Y}} + L$. Hence, by Lemma 6.1, we obtain

$$\begin{aligned} \left\|\sqrt{w(X)}F_{\pi([s])}(X)\right\|_{\mathcal{Y}} &\leq \frac{\sqrt{s+1}}{\sqrt{\|X\|_{\mathcal{X}}^2 + K(\mathcal{P}_{\pi([s])})(X)}}\|F\|_{L^2_\mu(\mathcal{X};\mathcal{Y})}\sqrt{K(\mathcal{P}_{\pi([s])})(X)} \\ &\leq \sqrt{s+1}\left(\|F(0)\|_{\mathcal{Y}} + L\right). \end{aligned} \tag{6.22}$$

We now set

$$C := \max\left\{8\left(\|F(0)\|_{\mathcal{Y}} + L\right)^2, c\right\} \tag{6.23}$$

and combine (6.21) and (6.22) to conclude

$$\left\|\sqrt{w}\left(F - F_{\pi([s])}\right)\right\|_{L^\infty_\mu(\mathcal{X};\mathcal{Y})} \le \sqrt{Cs}.$$

Next, we apply Lemma 6.7 with $k := Csu^{-2}_{\boldsymbol{\pi(s+1)}}$ to find that, with probability at least $1 - \epsilon/2$, we have

$$\left\|F - F_{\pi([s])}\right\|_{\mathrm{disc},w} \le \sqrt{2}\left(\left\|F - F_{\pi([s])}\right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} + u_{\boldsymbol{\pi(s+1)}}\right) \le \sqrt{2}u_{\boldsymbol{\pi(s+1)}}\left(\|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} + 1\right), \tag{6.24}$$

provided that

$$m \ge 2Csu^{-2}_{\boldsymbol{\pi(s+1)}}\log(4/\epsilon). \tag{6.25}$$

The last inequality in (6.24) follows from (4.1) and (4.2). Finally, we compare (6.25) with the lower bound on m in (6.19). By definition of C, we find

$$2Csu^{-2}_{\boldsymbol{\pi(s+1)}}\log(4s/\epsilon) \ge cs\log(2s/\epsilon),$$

where we used that $u_{\boldsymbol{\pi(s+1)}} \le 1$. Consequently, we can combine the bound on α_w in (6.19) and the bound on $\|F - F_{\pi([s])}\|_{\mathrm{disc},w}$ in (6.24) via the union bound. The proof is now complete. □

6.3 Algorithms

Recall from Subsection 4.1 the relation of the index set $\pi([s])$ to an anisotropic total degree index set via (4.5)–(4.7). Combined with the results from the Subsection 6.2, it yields a constructive way to approximate $W^{1,2}_{\mu,\boldsymbol{b}}$- and Lipschitz operators via Christoffel sampling and a weighted least-squares fit with near-optimal sample complexity. In the following, we present (high-level) algorithms for the approximation of general continuous $W^{1,2}_{\mu,\boldsymbol{b}}$-operators (Algorithm 1) and of Lipschitz operators (Algorithm 2), based on Corollary 6.5 and Theorem 6.6, respectively.

We introduce the following notation for the largest possible dimension of the approximation space $\mathcal{P}_S$, defined in (6.1): Let m be a given number of samples and let $0 < \epsilon < 1$ denote the failure probability. We define

$$s_1 = s_1(m,\epsilon) := \max\{s \in \mathbb{N} : cs\log(s/\epsilon) \le m\}, \tag{6.26}$$

where $c = 2\left(\log(1/2) + 1\right)^{-1}$. For a Lipschitz continuous operator $F : \mathcal{X} \to \mathcal{Y}$, suppose that we know the Lipschitz constant L and the value $\|F(0)\|_{\mathcal{Y}}$. In this case, we set

$$s_2 = s_2(m,\epsilon,L,\|F(0)\|_{\mathcal{Y}}) := \max\{s \in \mathbb{N} : 2Csu^{-2}_{\boldsymbol{\pi(s+1)}}\log(4/\epsilon) \le m\}, \tag{6.27}$$

where $C = C(L, \|F(0)\|_{\mathcal{Y}})$ is given by (6.23). We leave s_1 and s_2 undefined in cases where the maximum is taken over the empty set.

We make the following assumptions:

(i) The sequence of weighted PCA eigenvalues $(\lambda_{\boldsymbol{b},i})_{i\in\mathbb{N}}$ is nonincreasing (see Assumption 3.6) and we additionally have $\lim_{i\to\infty}\lambda_{\boldsymbol{b},i} = 0$ so that the effective dimension of the problem is finite, see Remark 4.4 and Remark 4.5.

(ii) We can exactly compute any finite number of the unweighted PCA eigenvalues λ_i, see steps 5 and 8 in Algorithm 1 and steps 2 and 5 in Algorithm 2.

(iii) We have pointwise access to the target operator F. In particular, we require continuity of the $W^{1,2}_{\mu,\boldsymbol{b}}$-operators in Algorithm 1 to ensure that pointwise evaluations are well-defined.

Algorithm 1 Least-squares approximation of an operator $F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y}) \cap C(\mathcal{X},\mathcal{Y})$

```
Input:
    $\mathbf{0} < \boldsymbol{b} \leq \mathbf{1}$                                  ▷ Weight sequence
    $d\nu = w^{-1} d\mu$                      ▷ Sampling measure with $w$ given by (6.10)
    $m \in \mathbb{N}$                                          ▷ Number of samples
    $\epsilon \in (0,1)$                                       ▷ Failure probability
    $0 < h < 1$                                          ▷ (Small) step size
Output: Least-squares approximant $\widehat{F}$ of $F$ which satisfies
    $$\left\| F - \widehat{F} \right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \leq \left(1 + \frac{2\sqrt{2}}{\sqrt{\epsilon}}\right) u_{\boldsymbol{\pi(s_1+1)}} \|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})}$$
    with $\nu$-probability at least $1-\epsilon$, where $s_1$ is given by (6.26).
$c \leftarrow 2\left(\log(1/2)+1\right)^{-1}$                          ▷ Sample complexity constant
if $c\log(1/\epsilon) > m$ then
    abort                                  ▷ Abort if there are too few samples
end if
Compute $\lambda_1$.
$\tau \leftarrow \lambda_{\boldsymbol{b},1}$                                        ▷ Eigenvalue tolerance
while TRUE do
    Compute all PCA eigenvalues $\lambda_i$ for which $\lambda_{\boldsymbol{b},i} \geq \tau$.
    $d \leftarrow \min\{l \in \mathbb{N} : \lambda_{\boldsymbol{b},l+1} < \tau\}$                     ▷ Effective dimension
    Construct the index set
        $$\widetilde{\Gamma} = \left\{ \boldsymbol{\gamma} \in \mathbb{N}_0^d : \sum_{i=1}^{d} \frac{\gamma_i}{\lambda_{\boldsymbol{b},i}} \leq \tau^{-1} \right\}$$
    and compute the corresponding weights
        $$u_{\boldsymbol{\gamma}} = \left(1 + \sum_{i=1}^{d} \frac{\gamma_i}{\lambda_{\boldsymbol{b},i}}\right)^{-1/2}, \quad \boldsymbol{\gamma} \in \widetilde{\Gamma}.$$
    $s \leftarrow |\widetilde{\Gamma}|$
    Construct a nonincreasing rearrangement $\pi : [s] \to \widetilde{\Gamma}$ of $(u_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma} \in \widetilde{\Gamma}}$.
    $s' \leftarrow \max\{r \in [s] : cr\log(r/\epsilon) \leq m\}$
    if $s' \leq s-1$ then
        $S \leftarrow \{\boldsymbol{\pi(1)}, \ldots, \boldsymbol{\pi(s')}\}$                          ▷ $s' = s_1$
        break
    else
        $\tau \leftarrow \tau - h$
    end if
end while
Draw $m$ samples $X_1, \ldots, X_m \sim_{i.i.d.} \nu$ and compute $F(X_1), \ldots, F(X_m)$.
Compute the least-squares approximant $\widehat{F}$ based on $\pi([s'])$ via (6.4).
```

Algorithm 2 Least-squares approximation of a Lipschitz operator $F : \mathcal{X} \to \mathcal{Y}$

```
Input:
    $\mathbf{0} < \boldsymbol{b} \leq \mathbf{1}$ with $\boldsymbol{b} \in \ell^2(\mathbb{N})$ if $\dim(\mathcal{Y}) = \infty$      ▷ Weight sequence
    $d\nu = w^{-1} d\mu$      ▷ Sampling measure with $w$ given by (6.17)
    $m \in \mathbb{N}$      ▷ Number of samples
    $\epsilon \in (0,1)$      ▷ Failure probability
    $\|F(0)\|_{\mathcal{Y}}$      ▷ Norm of operator value at origin
    $L > 0$      ▷ Lipschitz constant
    $0 < h < 1$      ▷ (Small) step size
Output: Least-squares approximant $\widehat{F}$ of $F$ which satisfies
    $\left\| F - \widehat{F} \right\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \leq \sqrt{2} u_{\boldsymbol{\pi(s_2+1)}} \left( \|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} + 1 \right)$
    with $\nu$-probability at least $1 - \epsilon$, where $s_2$ is given by (6.27).
$C \leftarrow \max \left\{ 8 \left( \|F(0)\|_{\mathcal{Y}} + L \right)^2, 2 \left( \log(1/2) + 1 \right)^{-1} \right\}$      ▷ Sample complexity constant
Compute $\lambda_1$.
$\tau \leftarrow \lambda_{\boldsymbol{b},1}$      ▷ Eigenvalue tolerance
while TRUE do
    Compute all PCA eigenvalues $\lambda_i$ for which $\lambda_{\boldsymbol{b},i} \geq \tau$.
    $d \leftarrow \min\{ l \in \mathbb{N} : \lambda_{\boldsymbol{b},l+1} < \tau \}$      ▷ Effective dimension
    Construct the index set
        $\widetilde{\Gamma} = \left\{ \boldsymbol{\gamma} \in \mathbb{N}_0^d : \sum_{i=1}^{d} \frac{\gamma_i}{\lambda_{\boldsymbol{b},i}} \leq \tau^{-1} \right\}$
    and corresponding weights
        $u_{\boldsymbol{\gamma}} = \left( 1 + \sum_{i=1}^{d} \frac{\gamma_i}{\lambda_{\boldsymbol{b},i}} \right)^{-1/2}, \quad \boldsymbol{\gamma} \in \widetilde{\Gamma}.$
    $s \leftarrow |\widetilde{\Gamma}|$
    Construct a nonincreasing rearrangement $\pi : [s] \to \widetilde{\Gamma}$ of $(u_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma} \in \widetilde{\Gamma}}$.
    if $2Cu^{-2}_{\boldsymbol{\pi(2)}} \log(4/\epsilon) > m$ then
        abort      ▷ Abort if there are too few samples
    else
        $s' \leftarrow \max\{ r \in [s] : 2Cru^{-2}_{\boldsymbol{\pi(r+1)}} \log(4/\epsilon) \leq m \}$
        if $s' \leq s - 1$ then
            $S \leftarrow \{ \boldsymbol{\pi(1)}, \dots, \boldsymbol{\pi(s')} \}$      ▷ $s' = s_2$, see (6.27)
            break
        else
            $\tau \leftarrow \tau - h$
        end if
    end if
end while
Draw $m$ samples $X_1, \dots, X_m \sim_{i.i.d.} \nu$ and compute $F(X_1), \dots, F(X_m)$.
Compute the least-squares approximant $\widehat{F}$ based on $\pi([s_2])$ via (6.4).
```

6.4 Discussion

The results in this section show that weighted least-squares approximation via Christoffel sampling is an efficient way to reconstruct L^2_μ-, $W^{1,2}_{\mu,\boldsymbol{b}}$-, and Lipschitz operators from finitely many point samples. In light of Theorem 5.4, the resulting approximation errors in Corollary 6.5 and Theorem 6.6 are *quasi-optimal*, that is, they are, up to constants, best possible uniformly for all operators in the Sobolev unit (Lipschitz) ball. The sample complexity in (6.15) is *near-optimal* in the sense that it is linear in s up to a log-factor. The sample complexity in (6.25) is linear in s up to a log-term and the factor $u^{-2}_{\boldsymbol{\pi}(s+1)}$. The latter grows only subalgebraically in s by Theorem 4.6. In this (broader) sense, the sample complexity in (6.25) is again near-optimal.

The presented algorithms implement our theoretical findings as constructive approximation schemes. Note that the data points $(X_i, F(X_i)) \in \mathcal{X} \times \mathcal{Y}$ are infinite-dimensional, so additional discretization steps are necessary to make the method applicable in practice. As already mentioned in the introduction, actual implementations of any algorithm for learning Lipschitz operators on hardware always require in-memory costs, i.e., number of bits, that are exponential in the reciprocal of the approximation error, see [48]. Nevertheless, our algorithms show that only finitely many PCA eigenvalues λ_i are necessary to construct near-optimal approximants for $W^{1,2}_{\mu,\boldsymbol{b}}$- and Lipschitz operators. In particular, infinite eigenvalue searches can be avoided.

7 Conclusion and outlook

In this article, we analyzed the approximation of Hilbert-valued Lipschitz operators from finite data. We first extended results from infinite-dimensional analysis and showed that all Lipschitz operators lie in a Gaussian Sobolev space $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$. We then studied Hermite polynomial s-term approximations and proved that they cannot achieve algebraic convergence rates. This curse of parametric complexity is independent of the decay of the (weighted) PCA eigenvalues $\lambda_{\boldsymbol{b},i}$ of the covariance operator of the Gaussian measure μ. We illustrated how the decay of the $\lambda_{\boldsymbol{b},i}$ influences the approximation rate and proved that convergence rates arbitrarily close to any algebraic rate can be attained at least asymptotically for $s \to \infty$ if the eigenvalues decay double-exponentially.

We studied the smallest worst-case error for reconstructing $W^{1,2}_{\mu,\boldsymbol{b}}$- and Lipschitz operators from m (potentially adaptively chosen) samples in terms of the adaptive m-width and tightly quantified the dependence of the latter on the $\lambda_{\boldsymbol{b},i}$. We showed that no recovery strategy based on finite (adaptive) linear information can achieve algebraic convergence rates for all $W^{1,2}_{\mu,\boldsymbol{b}}$-operators. This curse of sample complexity (which implies the curse of parametric complexity) holds for a general (centered, non-degenerate) Gaussian measure independently of its spectral properties. The same is true for Lipschitz operators. In particular, restricting the set of all $W^{1,2}_{\mu,\boldsymbol{b}}$-operators to only those which are Lipschitz continuous does not provide enough additional regularity to overcome the curse of sample complexity. It is an active area of research to identify classes of operators for which efficient learning in the sense of algebraic convergence rates is possible. As discussed in Subsection 1.1, examples include holomorphic operators and solution operators of certain PDEs. On the positive side, we proved that $W^{1,2}_{\mu,\boldsymbol{b}}$-regularity, and Lipschitz regularity in particular, suffices to achieve approximation rates which are arbitrarily close to any algebraic rate in the large data limit $m \to \infty$, provided that the PCA eigenvalues $\lambda_{\boldsymbol{b},i}$ decay double-exponentially.

Finally, we passed from general (adaptive) linear information to standard information and studied the approximation of $W^{1,2}_{\mu,\boldsymbol{b}}$- and Lipschitz operators based on finitely many point samples. We showed that by means of Christoffel sampling and weighted least-squares approximation it is possible to achieve near-optimal sample complexities and we presented corresponding constructive algorithms.

We conclude with discussing several open problems: The estimates in Corollary 6.5 and Theorem 6.6 are *nonuniform* in the given class of operators, that is, they only hold for a fixed operator of that class and therefore do not provide upper bounds for the adaptive m-width. It is an open problem to prove corresponding uniform bounds based on Christoffel sampling and weighted least-squares approximation. We remark that it is possible to derive such a uniform bound in the context of Theorem 6.4 with the L^∞-norm and without ϵ on the right-hand side in (6.12), see [6, Cor. 5.9]. However, this is not helpful in our setting as we do not expect L^∞-convergence of the approximation error.

We also remark in passing that there has been a series of recent works that refine Christoffel sampling to remove the logarithmic dependence on s in the sample complexity estimates, as it appears in Theorem 6.4 and Theorem 6.6, see, e.g., [26, 27, 46, 47, 54, 59, 69]. However, they involve more elaborate constructions, so for simplicity we have considered standard Christoffel sampling. Furthermore, these results only consider scalar-valued functions, not operators, so they are not immediately applicable in our setting. They also do not always provide upper bounds in the L^2_μ-norm, which is essential for our analysis. We remark that recent works also consider other L^p-norms [45, 44]. We leave it to the future to study such refinements and generalizations.

It remains unclear whether Christoffel sampling provides any benefits over Monte Carlo sampling for learning Lipschitz operators in the Gaussian setting, that is, whether (uniform) near-optimal sample complexities can be achieved with samples drawn directly from the underlying Gaussian measure. We refer to [9], where the same question was studied in the context of holomorphic operators and Jacobi measures. In that case, Monte Carlo sampling is as good as Christoffel sampling.

In the present paper, we did not consider encoding and decoding errors, errors due to noisy observations, or the error of computing empirical PCA bases, as in PCA-Net [49]. In subsequent work we shall take these errors into account and study the convergence of PCA-Net-like approaches for learning Lipschitz operators as well as analyze deep neural network approximations, e.g., in terms of practical existence theorems [7, 2, 3, 4, 32], or techniques from statistical learning theory [64, 55]. The present article provides the theoretical foundations for these future research directions and serves as an important first step towards practical implementations of (near-)optimal approximation algorithms for Lipschitz operators.

Acknowledgments

Ben Adcock was supported by the the *Natural Sciences and Engineering Research Council of Canada of Canada* (NSERC) through grant RGPIN-2021-611675. Michael Griebel and Gregor Maier were supported by the *Hausdorff Center for Mathematics* (HCM) in Bonn, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2047/1 – 390685813 and by the CRC 1060 *The Mathematics of Emergent Effects* – 211504053 of the Deutsche Forschungsgemeinschaft. A part of this work was done when Gregor Maier was visiting the Department of Mathematics at Simon Fraser University (SFU). Gregor Maier thanks the Department of Mathematics at SFU for providing a fruitful research environment and working conditions. He also thanks Ben Adcock for his invitation and hospitality as well as the Bonn International Graduate School of Mathematics (BIGS) of the HCM for financial support during the visit.

References

[1] B. Adcock. Optimal sampling for least-squares approximation (2024). arXiv: 2409.02342.

[2] B. Adcock, S. Brugiapaglia, N. Dexter, and S. Moraga. Deep neural networks are effective at learning high-dimensional Hilbert-valued functions from limited data. *Proceedings of the 2nd mathematical and scientific machine learning conference*. Mathematical and Scientific Machine Learning. PMLR, 2022, pp. 1–36.

[3] B. Adcock, S. Brugiapaglia, N. Dexter, and S. Moraga. Learning smooth functions in high dimensions. *Handbook of Numerical Analysis*. Vol. 25. Elsevier, 2024, pp. 1–52.

[4] B. Adcock, S. Brugiapaglia, N. Dexter, and S. Moraga. Near-optimal learning of Banach-valued, high-dimensional functions via deep neural networks (2024). arXiv: 2211.12633.

[5] B. Adcock, S. Brugiapaglia, N. Dexter, and S. Moraga. *On Efficient Algorithms for Computing Near-Best Polynomial Approximations to High-Dimensional, Hilbert-Valued Functions from Limited Samples*. 1st ed. Vol. 13. Memoirs of the European Mathematical Society. EMS Press, 2024.

[6] B. Adcock, S. Brugiapaglia, and C. Webster. *Sparse Polynomial Approximation of High-Dimensional Functions*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2022.

[7] B. Adcock and N. Dexter. The gap between theory and practice in function approximation with deep neural networks. *SIAM Journal on Mathematics of Data Science* 3.2 (2021), pp. 624–655.

[8] B. Adcock, N. Dexter, and S. Moraga. Optimal approximation of infinite-dimensional holomorphic functions. *Calcolo* 61.1 (2024), p. 12.

[9] B. Adcock, N. Dexter, and S. Moraga. Optimal approximation of infinite-dimensional holomorphic functions II: recovery from i.i.d. pointwise samples (2023). arXiv: 2310.16940.

[10] B. Adcock, N. Dexter, and S. Moraga. Optimal deep learning of holomorphic operators between Banach spaces (2024). arXiv: 2406.13928.

[11] N. Aronszajn. Differentiability of Lipschitzian mappings between Banach spaces. *Studia Mathematica* 57.2 (1976), pp. 147–190.

[12] F. Bartel and D. Dũng. Sampling recovery in Bochner spaces and applications to parametric PDEs with random inputs (2024). arXiv: 2409.05050.

[13] A. Beged-dov. Lower and upper bounds for the number of lattice points in a simplex. *SIAM Journal on Applied Mathematics* 22.1 (1972), pp. 106–108.

[14] J. Berezanskij, Z. Sheftel, and G. Us. *Functional Analysis*. Vol. 2. Operator theory 86. Basel: Birkhäuser, 1996.

[15] K. Bhattacharya, B. Hosseini, N. Kovachki, and A. Stuart. Model reduction and neural networks for parametric PDEs. *The SMAI Journal of computational mathematics* 7 (2021), pp. 121–157.

[16] V. Bogachev. *Gaussian Measures*. Mathematical surveys and monographs Volume 62. Providence, RI: American Mathematical Society, 1998.

[17] N. Boullé and A. Townsend. A mathematical guide to operator learning. *Handbook of Numerical Analysis*. Vol. 25. Elsevier, 2024, pp. 83–125.

[18] T. Chen and H. Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks* 6.4 (1995), pp. 911–917.

[19] A. Chojnowska-Michalik and B. Goldys. Generalized Ornstein–Uhlenbeck semigroups: Littlewood-Paley–Stein inequalities and the P. A. Meyer equivalence of norms. *Journal of Functional Analysis* 182.2 (2001), pp. 243–279.

[20] A. Cohen and R. DeVore. Approximation of high-dimensional parametric PDEs. *Acta Numerica* 24 (2015), pp. 1–159.

[21] A. Cohen and G. Migliorati. Optimal weighted least-squares methods. *The SMAI Journal of computational mathematics* 3 (2017), pp. 181–203.

[22] G. Da Prato and J. Zabczyk. Regular densities of invariant measures in Hilbert spaces. *Journal of Functional Analysis* 130.2 (1995), pp. 427–449.

[23] G. Da Prato. *An Introduction to Infinite-Dimensional Analysis.* 1st ed. Universitext. Springer Berlin, Heidelberg, 2006.

[24] G. Da Prato. *Introduction to Stochastic Analysis and Malliavin Calculus.* Pisa: Scuola Normale Superiore, 2014.

[25] T. De Ryck and S. Mishra. Generic bounds on the approximation error for physics-informed (and) operator learning. *Advances in neural information processing systems.* Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. 2022.

[26] M. Dolbeault and A. Cohen. Optimal pointwise sampling for L^2 approximation. *Journal of Complexity* 68 (2022), p. 101602.

[27] M. Dolbeault, D. Krieg, and M. Ullrich. A sharp upper bound for sampling numbers in L_2. *Applied and Computational Harmonic Analysis* 63 (2023), pp. 113–134.

[28] D. Dũng and M. Griebel. Hyperbolic cross approximation in infinite dimensions. *Journal of Complexity* 33 (2016), pp. 55–88.

[29] D. Dũng, V. Nguyen, and D. Pham. Deep ReLU neural network approximation in Bochner spaces and applications to parametric PDEs (2022). arXiv: 2111.05854.

[30] L. Evans and R. Gariepy. *Measure Theory and Fine Properties of Functions.* Revised edition. Textbooks in mathematics. Boca Raton, FL: CRC press Taylor & Francis group, 2015.

[31] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing.* Applied and Numerical Harmonic Analysis. New York, NY: Springer, 2013.

[32] N. Franco and S. Brugiapaglia. A practical existence theorem for reduced order models based on convolutional autoencoders (2024). arXiv: 2402.00435.

[33] M. Griebel and J. Oettershagen. On tensor product approximation of analytic functions. *Journal of Approximation Theory* 207 (2016), pp. 348–379.

[34] M. Griebel and P. Oswald. Stable splittings of Hilbert spaces of functions of infinitely many variables. *Journal of Complexity* 41 (2017), pp. 126–151.

[35] J. Gwinner, B. Jadamba, A. Khan, and F. Raciti. *Uncertainty Quantification in Variational Inequalities: Theory, Numerics, and Applications.* New York, NY: Chapman and Hall/CRC, 2021.

[36] A.-L. Haji-Ali, H. Harbrecht, M. Peters, and M. Siebenmorgen. Novel results for the anisotropic sparse grid quadrature. *Journal of Complexity* 47 (2018), pp. 62–85.

[37] L. Herrmann, C. Schwab, and J. Zech. Neural and spectral operator surrogates: unified construction and expression rate bounds. *Advances in Computational Mathematics* 50.4 (2024), p. 72.

[38] T. Hytönen, J. van Neerven, M. Veraar, and L. Weis. *Analysis in Banach Spaces: Volume I: Martingales and Littlewood-Paley Theory.* 1st ed. 2016. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics 63. Cham: Springer International Publishing, 2016.

[39] B. Kashin, E. Kosov, I. Limonova, and V. Temlyakov. Sampling discretization and related problems. *Journal of Complexity.* Approximation and Geometry in High Dimensions 71 (2022), p. 101653.

[40] N. Kovachki, S. Lanthaler, and H. Mhaskar. Data complexity estimates for operator learning (2024). arXiv: 2405.15992.

[41] N. Kovachki, S. Lanthaler, and S. Mishra. On universal approximation and error bounds for Fourier Neural Operators. *Journal of Machine Learning Research* 22.290 (2021), pp. 1–76.

[42] N. Kovachki, S. Lanthaler, and A. Stuart. Operator learning: Algorithms and analysis. *Handbook of Numerical Analysis*. Vol. 25. Elsevier, 2024, pp. 419–467.

[43] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research* 24.89 (2023), pp. 1–97.

[44] D. Krieg, K. Pozharska, M. Ullrich, and T. Ullrich. Sampling projections in the uniform norm (2024). arXiv: `2401.02220`.

[45] D. Krieg, K. Pozharska, M. Ullrich, and T. Ullrich. Sampling recovery in L_2 and other norms (2023). arXiv: `2305.07539`.

[46] D. Krieg and M. Ullrich. Function values are enough for L_2-approximation. *Foundations of Computational Mathematics* 21.4 (2021), pp. 1141–1151.

[47] D. Krieg and M. Ullrich. Function values are enough for L_2-approximation: Part II. *Journal of Complexity* 66 (2021), p. 101569.

[48] S. Lanthaler. Operator learning of Lipschitz operators: An information-theoretic perspective (2024). arXiv: `2406.18794`.

[49] S. Lanthaler. Operator learning with PCA-Net: upper and lower complexity bounds. *Journal of Machine Learning Research* 24.318 (2023), pp. 1–67.

[50] S. Lanthaler, Z. Li, and A. Stuart. Nonlocality and nonlinearity implies universality in operator learning (2024). arXiv: `2304.13221`.

[51] S. Lanthaler, S. Mishra, and G. Karniadakis. Error estimates for DeepONets: a deep learning framework in infinite dimensions. *Transactions of Mathematics and Its Applications* 6.1 (2022), tnac001.

[52] S. Lanthaler and A. Stuart. The parametric complexity of operator learning (2024). arXiv: `2306.15924`.

[53] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. *International conference on learning representations*. 2021.

[54] I. Limonova and V. Temlyakov. On sampling discretization in L_2. *Journal of Mathematical Analysis and Applications* 515.2 (2022), p. 126457.

[55] H. Liu, H. Yang, M. Chen, T. Zhao, and W. Liao. Deep nonparametric estimation of operators between infinite dimensional spaces. *Journal of Machine Learning Research* 25.24 (2024), pp. 1–67.

[56] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence* 3.3 (2021), pp. 218–229.

[57] A. Lunardi, M. Miranda, and D. Pallara. Infinite dimensional analysis. 2016. URL: `https://www.mathematik.tu-darmstadt.de/media/analysis/lehrmaterial_anapde/hallerd/Lectures.pdf` (accessed on Oct. 25, 2024).

[58] H. Luo, X. Zhang, and S. Zhao. Sobolev embeddings in infinite dimensions. *Science China Mathematics* 66.10 (2023), pp. 2157–2178.

[59] N. Nagel, M. Schäfer, and T. Ullrich. A new upper bound for sampling numbers. *Foundations of Computational Mathematics* 22.2 (2022), pp. 445–468.

[60] E. Novak. On the power of adaption. *Journal of Complexity* 12.3 (1996), pp. 199–237.

[61] E. Novak and H. Woźniakowski. *Tractability of Multivariate Problems. Volume I, Linear information.* Zürich: European Mathematical Society, 2008.

[62] D. Nualart and E. Nualart. *Introduction to Malliavin Calculus.* 1st ed. Cambridge University Press, 2018.

[63] J. Opschoor, C. Schwab, and J. Zech. Exponential ReLU DNN expression of holomorphic maps in high dimension. *Constructive Approximation* 55.1 (2022), pp. 537–582.

[64] J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics* 48.4 (2020), pp. 1875–1897.

[65] C. Schwab, A. Stein, and J. Zech. Deep operator network approximation rates for Lipschitz operators (2023). arXiv: `2307.09835`.

[66] C. Schwab and J. Zech. Deep learning in high dimension: Neural network expression rates for analytic functions in $L^2(\mathbb{R}^d, \gamma_d)$. *SIAM/ASA Journal on Uncertainty Quantification* 11.1 (2023), pp. 199–234.

[67] I. Shigekawa. Sobolev spaces over the Wiener space based on an Ornstein-Uhlenbeck operator. *Kyoto Journal of Mathematics* 32.4 (1992), pp. 731–748.

[68] M. Stesin. Aleksandrov diameters of finite-dimensional sets and classes of smooth functions. *Doklady Akademii Aauk SSSR* 220.6 (1975), pp. 1278–1281.

[69] V. Temlyakov. On optimal recovery in L_2. *Journal of Complexity* 65 (2021), p. 101545.

[70] V. Temlyakov. The Marcinkiewicz-type discretization theorems. *Constructive Approximation* 48.2 (2018), pp. 337–369.

Appendices

A Notions of differentiability

We recall several notions of differentiability, loosely following [16, Chpt. 5.1]. The constructions in this section hold for general Banach spaces $\mathcal{X}$ and $\mathcal{Y}$. As usual, we denote by $L(\mathcal{X}, \mathcal{Y})$ the space of all bounded linear operators F from $\mathcal{X}$ to $\mathcal{Y}$ with finite operator norm $\|F\|_{L(\mathcal{X},\mathcal{Y})} := \sup_{X\in\mathcal{X}, X\neq 0} \|F(X)\|_{\mathcal{Y}}/\|X\|_{\mathcal{X}}$. We set $L(\mathcal{X}) := L(\mathcal{X}, \mathbb{R})$.

Definition A.1 (Differentiability)**.** Let $\mathcal{M}$ be a collection of non-empty subsets of $\mathcal{X}$, let $X \in \mathcal{X}$, and let Ω be an open neighborhood of X. A mapping $F : \Omega \to \mathcal{Y}$ is said to be differentiable with respect to $\mathcal{M}$ at the point X if there exists a continuous linear mapping $\ell \in L(\mathcal{X}, \mathcal{Y})$ such that for every fixed set $M \in \mathcal{M}$, we have

$$\lim_{t\to 0} \sup_{Z\in M} \left\| \frac{F(X + tZ) - F(X)}{t} - \ell(Z) \right\|_{\mathcal{Y}} = 0.$$

In that case, ℓ is unique and we write $D^{\mathcal{M}}F(X) := \ell$ for the derivative of F at X.

If $\mathcal{M}$ is the class of all *finite*, *compact*, or *bounded* subsets of $\mathcal{X}$, then we say that F is *Gâteaux*, *Hadamard*, or *Fréchet differentiable at X*, respectively. If $\mathcal{X}$ is finite-dimensional, then Hadamard and Fréchet differentiability at X are equivalent and the corresponding derivatives of F at X coincide. We usually drop the superscript $\mathcal{M}$ in the notation of the derivative and write $DF(X)$ instead of $D^{\mathcal{M}}F(X)$. It will be clear from context which notion of derivative we refer to. We call F *differentiable* (in the corresponding sense) if it is differentiable (in the corresponding sense) at every point $X \in \mathcal{X}$. The resulting derivative DF is a mapping from $\mathcal{X}$ to $L(\mathcal{X}, \mathcal{Y})$.

If $\mathcal{E}$ is a linear subspace of $\mathcal{X}$ (possibly with a stronger norm), then we say that F is *differentiable along $\mathcal{E}$ at the point X* (in the corresponding sense) if the mapping $Z \mapsto F(X + Z)$ is differentiable from $\mathcal{E}$ to $\mathcal{Y}$ at $Z = 0$ (in the corresponding sense). If F is Fréchet differentiable along $\mathcal{E}$ at X, then it is Hadamard differentiable along $\mathcal{E}$ at X. If F is Hadamard differentiable along $\mathcal{E}$ at X, then it is Gâteaux differentiable along $\mathcal{E}$ at X. In both cases, the corresponding derivatives at X coincide and we denote them by $D_{\mathcal{E}}F(X)$. We call F *differentiable along $\mathcal{E}$* (in the corresponding sense) if it is differentiable along $\mathcal{E}$ at every point $X \in \mathcal{X}$ (in the corresponding sense). The resulting derivative $D_{\mathcal{E}}F$ is a mapping from $\mathcal{X}$ to $L(\mathcal{E}, \mathcal{Y})$. Moreover, the differential operator $D_{\mathcal{E}}$, mapping F to its derivative $D_{\mathcal{E}}F$, is linear in F.

Example A.2. If $\mathcal{E} = \mathcal{X}$, then $D_{\mathcal{E}}F = DF$. If F is Fréchet differentiable and we choose $\mathcal{E}$ to be the Cameron-Martin space $\mathcal{H}$ of a Gaussian measure on $\mathcal{X}$ (see Appendix C.1), then $D_{\mathcal{E}}F$ is the $\mathcal{H}$-derivative of F which is commonly used in infinite-dimensional analysis, see [57, Sect. 9].

If $\mathcal{E} = \operatorname{span}\{Z\}$, $Z \in \mathcal{X} \setminus \{0\}$, is one-dimensional, we obtain the usual directional derivative

$$\frac{\partial}{\partial Z}F(X) := D_{\operatorname{span}\{Z\}}F(X)(Z) = \lim_{t\to 0} \frac{F(X + tZ) - F(X)}{t} \in \mathcal{Y}.$$

In this case, the Gâteaux, Hadamard, and Fréchet derivatives along $\mathcal{E}$ at X coincide. For any subspaces $\mathcal{E}' \subset \mathcal{E} \subset \mathcal{X}$, it can be readily seen that, if $F : \mathcal{X} \to \mathcal{Y}$ is differentiable along $\mathcal{E}$ at a point $X \in \mathcal{X}$ (in the corresponding sense), then F is also differentiable along $\mathcal{E}'$ at X (in the corresponding sense) and

$$D_{\mathcal{E}}F(X)|_{\mathcal{E}'} = D_{\mathcal{E}'}F(X).$$

In particular, if F is differentiable along $\mathcal{E}$ at X (in the corresponding sense), then, for any $Z \in \mathcal{E} \setminus \{0\}$, the directional derivative $\frac{\partial}{\partial Z}F(X)$ at X exists and

$$\frac{\partial}{\partial Z}F(X) = D_{\mathcal{E}}F(X)(Z).$$

If $\mathcal{X} = \mathbb{R}^n$ and $Z = \boldsymbol{e_i}$ is the ith standard unit vector, we use the standard notation $\partial_i F(\boldsymbol{x}) = \partial_{x_i} F(\boldsymbol{x}) := \frac{\partial}{\partial \boldsymbol{e_i}} F(\boldsymbol{x})$ for $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$.

If $\mathcal{E}$ is a Hilbert subspace of $\mathcal{X}$ and if $F : \mathcal{X} \to \mathbb{R}$ is Fréchet differentiable along $\mathcal{E}$ at X, then, by the Riesz representation theorem, there exists a unique $Z \in \mathcal{E}$ such that $D_{\mathcal{E}} F(X)(X') = \langle Z, X' \rangle_{\mathcal{E}}$ for every $X' \in \mathcal{E}$. In this case, we call Z the *$\mathcal{E}$-gradient of F at X* and write

$$\nabla_{\mathcal{E}} F(X) := Z.$$

Definition A.3 (The space $C_b^1(\mathcal{X})$)**.** We denote by $C_b^1(\mathcal{X})$ the set of all *boundedly Fréchet differentiable functionals on $\mathcal{X}$*, that is, the set of all Fréchet differentiable mappings $F : \mathcal{X} \to \mathbb{R}$ which are bounded on $\mathcal{X}$ and whose derivative DF is bounded in $L(\mathcal{X})$. The corresponding norm is given by

$$\|F\|_{C_b^1(\mathcal{X})} := \sup_{X \in \mathcal{X}} |F(X)| + \|DF\|_{L(\mathcal{X})}.$$

B Results from operator theory

B.1 Closability and closure of operators

We define closability and the closure of operators between Hilbert spaces and state standard properties. For further details we refer to [14, Chpt. 12].

Let $\mathcal{H}_1, \mathcal{H}_2$ be two Hilbert spaces. A linear $\mathcal{H}_2$-valued operator (not necessarily bounded) acting on $\mathcal{H}_1$ is a linear mapping $A : \mathrm{dom}(A) \to \mathcal{H}_2$ from a linear subspace $\mathrm{dom}(A) \subset \mathcal{H}_1$ to $\mathcal{H}_2$. The set $\mathrm{dom}(A)$ is called the *domain* of A. The *graph* of A is defined as the set

$$\Gamma_A := \{(H, AH) \in \mathcal{H}_1 \oplus \mathcal{H}_2 : H \in \mathrm{dom}(A)\}.$$

Considered as a subspace of the direct sum $\mathcal{H}_1 \oplus \mathcal{H}_2$ and equipped with the *graph inner product*,

$$\langle H, K \rangle_{\Gamma_A} := \langle H, K \rangle_{\mathcal{H}_1} + \langle AH, AK \rangle_{\mathcal{H}_2}, \quad H, K \in \mathrm{dom}(A),$$

this becomes a Hilbert space with *graph norm*

$$\|(H, AH)\|_{\Gamma_A} := \sqrt{\langle H, H \rangle_{\Gamma_A}} = \left(\|H\|_{\mathcal{H}_1}^2 + \|AH\|_{\mathcal{H}_2}^2 \right)^{1/2}.$$

Definition B.1 (Closability and closure)**.** A linear operator $A : \mathrm{dom}(A) \to \mathcal{H}_2$ is called *closable* (in $\mathcal{H}_1$) if the closure of its graph $\overline{\Gamma}_A$ in $\mathcal{H}_1 \oplus \mathcal{H}_2$ is the graph of some (necessarily unique) linear operator, that is, there exists a linear operator $\overline{A} : \mathrm{dom}(\overline{A}) \to \mathcal{H}_2$ such that $\overline{\Gamma}_A = \Gamma_{\overline{A}}$. In this case, we call $\overline{A}$ the *closure* of A.

If A is closable, then the domain of its closure is given by

$$\mathrm{dom}(\overline{A}) = \left\{ H \in \mathcal{H}_1 : \exists (H_n)_{n\in\mathbb{N}} \subset \mathrm{dom}(A) : \lim_{n\to\infty} H_n = H, \ (AH_n)_{n\in\mathbb{N}} \text{ converges in } \mathcal{H}_2 \right\}.$$

For $H \in \mathrm{dom}(\overline{A})$, we have

$$\overline{A}H = \lim_{n\to\infty} AH_n \quad \text{in } \mathcal{H}_2$$

for every sequence $(H_n)_{n\in\mathbb{N}} \subset \mathrm{dom}(\overline{A})$ such that $H_n \to H$ in $\mathcal{H}_1$, and the limit $\lim_{n\to\infty} AH_n$ is independent of the sequence $(H_n)_{n\in\mathbb{N}}$ (cf. [14, Thm. 2.1]). If A is closable, we equip $\mathrm{dom}(\overline{A})$ with the graph inner product, which turns $(\mathrm{dom}(\overline{A}), \langle \cdot, \cdot \rangle_{\Gamma_{\overline{A}}})$ into a Hilbert space.

B.2 Hilbert-Schmidt operators

We recall the notion of Hilbert-Schmidt operators. Further details can be found, e.g., in [14, Chpt. 8.7]. Let $\mathcal{H}_1, \mathcal{H}_2$ be two separable Hilbert spaces.

Definition B.2 (Hilbert-Schmidt operator)**.** A bounded linear operator $A \in L(\mathcal{H}_1, \mathcal{H}_2)$ is called a *Hilbert-Schmidt operator* if there exists an orthonormal basis $\{\zeta_i\}_{i\in\mathbb{N}}$ of $\mathcal{H}_1$ such that

$$\sum_{i=1}^{\infty} \|A\zeta_i\|_{\mathcal{H}_2}^2 < \infty. \tag{B.1}$$

The convergence of the series (B.1) and its value are independent of the basis of $\mathcal{H}_1$. We denote the space of all Hilbert-Schmidt operators from $\mathcal{H}_1$ to $\mathcal{H}_2$ by $HS(\mathcal{H}_1, \mathcal{H}_2)$ and set

$$\|A\|_{HS(\mathcal{H}_1,\mathcal{H}_2)} := \left(\sum_{i=1}^{\infty} \|A\zeta_i\|_{\mathcal{H}_2}^2\right)^{1/2}$$

for any orthonormal basis $\{\zeta_i\}_{i\in\mathbb{N}}$ of $\mathcal{H}_1$. This norm is induced by the inner product

$$\langle A, B\rangle_{HS(\mathcal{H}_1,\mathcal{H}_2)} := \sum_{i=1}^{\infty} \langle A\zeta_i, B\zeta_i\rangle_{\mathcal{H}_2},$$

where for any pair of Hilbert-Schmidt operators A, B, the series on the right-hand side converges for every orthonormal basis $\{\zeta_i\}_{i\in\mathbb{N}}$ of $\mathcal{H}_1$ and its value is independent of the basis. The space $HS(\mathcal{H}_1, \mathcal{H}_2)$ with this inner product is a separable Hilbert space.

C Results from infinite-dimensional analysis

We recall some well-known results from infinite-dimensional analysis which are used to define the Gaussian Sobolev space $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ (see Definition 3.1) and to prove Theorem 3.9 in Appendix D. We first define the Cameron-Martin space $\mathcal{H}$ of μ in $\mathcal{X}$ in Appendix C.1 and then discuss the construction of $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ as well as some of its important properties in Appendix C.2. We mainly follow [57] and [23], which consider the cases $\boldsymbol{b} = \sqrt{\boldsymbol{\lambda}}$ and $\boldsymbol{b} = \boldsymbol{1}$, respectively, and generalize the proofs therein to the case $\boldsymbol{0} < \boldsymbol{b} \leq \boldsymbol{1}$. More information can also be found in [16]. Throughout this section, we use notation as introduced in Sections 2 and 3.

C.1 The Cameron-Martin space

We commence with the celebrated *Fernique theorem*:

Theorem C.1 (Fernique, [57, Thm. 2.3.1])**.** *There exists $\alpha > 0$ such that*

$$\int_{\mathcal{X}} \exp(\alpha\|X\|_{\mathcal{X}}^2) d\mu(X) < \infty.$$

Fernique's theorem implies that any mapping $\mathcal{X} \to \mathcal{Y}$ which grows at most polynomially at infinity belongs to $L^2_\mu(\mathcal{X};\mathcal{Y})$. In particular, the mapping

$$j : \mathcal{X}^* \to L^2_\mu(\mathcal{X}), \quad F \mapsto j(F) = F,$$

is well-defined. We use it to define the Cameron-Martin space:

Definition C.2 (Cameron-Martin space)**.** The *Cameron-Martin space of μ (in $\mathcal{X}$)* is the set of all $X \in \mathcal{X}$ whose $\mathcal{H}$-norm is finite, where

$$\|X\|_{\mathcal{H}} := \sup \left\{ F(X) : F \in \mathcal{X}^*, \|j(F)\|_{L^2_\mu(\mathcal{X})} \leq 1 \right\}.$$

To further describe the structure of the space $\mathcal{H}$, we introduce the *reproducing kernel Hilbert space $\mathcal{X}^*_\mu$* of μ as the closure of $j(\mathcal{X}^*)$ under the $L^2_\mu(\mathcal{X})$-norm, that is,

$$\mathcal{X}^*_\mu := \overline{j(\mathcal{X}^*)}^{\|\cdot\|_{L^2_\mu(\mathcal{X})}},$$

together with the mapping

$$R_\mu : \mathcal{X}^*_\mu \to \mathcal{X}, \ \ F \mapsto \int_{\mathcal{X}} X F(X) d\mu(X),$$

where the integral is to be understood in the sense of Bochner. Note that R_μ is well-defined by Theorem C.1.

Proposition C.3 (Relation between $\mathcal{H}$ and $\mathcal{X}^*_\mu$, [57, Prop. 3.1.2])**.** *An element $H \in \mathcal{X}$ belongs to $\mathcal{H}$ if and only if there exists $\hat{H} \in \mathcal{X}^*_\mu$ such that $H = R_\mu(\hat{H})$. In that case, we have*

$$\|H\|_{\mathcal{H}} = \|\hat{H}\|_{L^2_\mu(\mathcal{X})}.$$

Hence, $R_\mu : \mathcal{X}^*_\mu \to \mathcal{H}$ is an isometric isomorphism that turns $\mathcal{H}$ into a Hilbert space with inner product

$$\langle H, K \rangle_{\mathcal{H}} := \left\langle \hat{H}, \hat{K} \right\rangle_{L^2_\mu(\mathcal{X})}$$

whenever $H = R_\mu \hat{H}$ and $K = R_\mu \hat{K}$.

In our case, where $\mathcal{X}$ is a separable Hilbert space, the Cameron-Martin space has a particularly simple structure in terms of the covariance operator Q of μ and the corresponding orthonormal PCA eigenbasis $\{\phi_i\}_{i \in \mathbb{N}}$ of $\mathcal{X}$ and PCA eigenvalues λ_i:

Theorem C.4 (Cameron-Martin space in a separable Hilbert space, cf. [57, Thm. 4.2.7])**.**

(i) The Cameron-Martin space of μ is given by

$$\mathcal{H} = Q^{1/2}(\mathcal{X}).$$

(ii) For $H = Q^{1/2}(Z) \in \mathcal{H}$ with $Z \in \mathcal{X}$, we have

$$\hat{H}(X) = \sum_{i=1}^{\infty} \lambda_i^{-1/2} \langle X, \phi_i \rangle_{\mathcal{X}} \langle Z, \phi_i \rangle_{\mathcal{X}} \quad \text{for } \mu\text{-a.e. } X \in \mathcal{X}.$$

(iii) The inner product in $\mathcal{H}$ satisies

$$\langle H, K \rangle_{\mathcal{H}} = \left\langle Q^{-1/2} H, Q^{-1/2} K \right\rangle_{\mathcal{X}}, \quad \forall H, K \in \mathcal{H}.$$

In particular, the family of vectors $\{\xi_i\}_{i \in \mathbb{N}}$ with

$$\xi_i := \sqrt{\lambda_i} \phi_i, \quad \forall i \in \mathbb{N}, \tag{C.1}$$

is an orthonormal basis of $\mathcal{H}$ and $\hat{\xi}_i \in \mathcal{X}^*_\mu$ is given by

$$\hat{\xi}_i(\cdot) = \lambda_i^{-1/2} \langle \cdot, \phi_i \rangle_{\mathcal{X}}, \quad \forall i \in \mathbb{N}. \tag{C.2}$$

The right-hand side in (C.2) *defines an element in $\mathcal{X}^*$ and we identify each $\hat{\xi}_i$ with its version in $\mathcal{X}^*$.*

C.2 The Gaussian Sobolev space $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$

Recall from Subsection 2.3 the weighted space $\mathcal{X}_{\boldsymbol{b}}$ with weight sequence $\boldsymbol{b} = (b_i)_{i\in\mathbb{N}}$, $\mathbf{0} < \boldsymbol{b} \leq \mathbf{1}$, and with orthonormal basis $\{\eta_i\}_{i\in\mathbb{N}}$, defined in (2.1).

C.2.1 Construction

The construction of Gaussian Sobolev spaces is based on so-called *cylindrical functionals* which have an explicit simple structure.

Definition C.5 (Cylindrical functionals and operators)**.** A functional $\varphi : \mathcal{X} \to \mathbb{R}$ is called a *cylindrical functional* if there exist $n \in \mathbb{N}$, $\ell_1, \ldots, \ell_n \in \mathcal{X}^*$, and a function $\omega : \mathbb{R}^n \to \mathbb{R}$ such that

$$\varphi(X) = \omega(\ell_1(X), \ldots, \ell_n(X)), \quad \forall X \in \mathcal{X}.$$

We call φ a *cylindrical boundedly Fréchet differentiable functional* if, with the above notation, $\omega \in C^1_b(\mathbb{R}^n)$. The space of all such functionals is denoted by $\mathcal{F}C^1_b(\mathcal{X})$. Moreover, we define the set of all *cylindrical boundedly Fréchet differentiable $\mathcal{Y}$-valued operators* by

$$\mathcal{F}C^1_b(\mathcal{X},\mathcal{Y}) := \operatorname{span}\left\{\mathcal{X} \ni X \mapsto \varphi(X)Y \in \mathcal{Y} : \varphi \in \mathcal{F}C^1_b(\mathcal{X}), Y \in \mathcal{Y}\right\}.$$

Lemma C.6. *For every $F \in \mathcal{F}C^1_b(\mathcal{X},\mathcal{Y})$ the derivative $D_{\mathcal{X}_{\boldsymbol{b}}}F$ lies in $L^2_\mu(\mathcal{X}; HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y}))$.*

Proof. By linearity of the differential operator $D_{\mathcal{X}_{\boldsymbol{b}}}$, it suffices to consider cylindrical operators $F \in \mathcal{F}C^1_b(\mathcal{X},\mathcal{Y})$ of the form $F(\cdot) = \varphi(\cdot)Y$ for some $\varphi \in \mathcal{F}C^1_b(\mathcal{X})$ and $Y \in \mathcal{Y}$. We fix any such F and note that it is Fréchet differentiable along $\mathcal{X}_{\boldsymbol{b}}$ at every point $X \in \mathcal{X}$ with derivative

$$D_{\mathcal{X}_{\boldsymbol{b}}}F(X)(Z) = \langle \nabla_{\mathcal{X}_{\boldsymbol{b}}}\varphi(X), Z\rangle_{\mathcal{X}_{\boldsymbol{b}}} Y, \quad \forall Z \in \mathcal{X}_{\boldsymbol{b}}.$$

Next, define the map

$$J_Y : \mathcal{X}_{\boldsymbol{b}} \to HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y}), \quad J_Y(X)(Z) := \langle X, Z\rangle_{\mathcal{X}_{\boldsymbol{b}}} Y,$$

It is well-defined as well as linear and bounded. Indeed, using the orthonormal basis $\{\eta_i\}_{i\in\mathbb{N}}$ of $\mathcal{X}_{\boldsymbol{b}}$, we have, by Parseval's identity,

$$\|J_Y(X)\|^2_{HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y})} = \sum_{i=1}^{\infty} \|J_Y(X)(\eta_i)\|^2_{\mathcal{Y}} = \sum_{i=1}^{\infty} \left|\langle X, \eta_i\rangle_{\mathcal{X}_{\boldsymbol{b}}}\right|^2 \|Y\|^2_{\mathcal{Y}} = \|X\|^2_{\mathcal{X}_{\boldsymbol{b}}} \|Y\|^2_{\mathcal{Y}}.$$

Moreover, by definition, the map $X \mapsto \nabla_{\mathcal{X}_{\boldsymbol{b}}}\varphi(X)$ is continuous and bounded from $\mathcal{X}$ to $\mathcal{X}_{\boldsymbol{b}}$. Since $D_{\mathcal{X}_{\boldsymbol{b}}}F = J_Y \circ \nabla_{\mathcal{X}_{\boldsymbol{b}}}\varphi$, we conclude that $F \mapsto D_{\mathcal{X}_{\boldsymbol{b}}}F$ is continuous and bounded as a map from $\mathcal{X}$ to $HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y})$. In particular, by Theorem C.1, it belongs to $L^2_\mu(\mathcal{X}; HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y}))$. □

Proposition C.7 (Closability of $D_{\mathcal{X}_{\boldsymbol{b}}}$)**.** *The (Fréchet) differential operator along $\mathcal{X}_{\boldsymbol{b}}$, $D_{\mathcal{X}_{\boldsymbol{b}}} : \mathcal{F}C^1_b(\mathcal{X};\mathcal{Y}) \to L^2_\mu(\mathcal{X}; HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y}))$ is closable in $L^2_\mu(\mathcal{X};\mathcal{Y})$.*

Proof. First note that the mapping $D_{\mathcal{X}_{\boldsymbol{b}}} : \mathcal{F}C^1_b(\mathcal{X};\mathcal{Y}) \to L^2_\mu(\mathcal{X}; HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y}))$ is well-defined by Lemma C.6. The proof of closability is a straight-forward modification of the proof of [57, Lem. 10.2.4], replacing the derivative along the Cameron-Martin space by the derivative along the space $\mathcal{X}_{\boldsymbol{b}}$. □

The previous result justifies the definition of the space $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ as described in Definition 3.1.

C.2.2 Properties

The following lemma provides an important criterion for an operator to belong to $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$:

Lemma C.8. *If $F_n \to F$ in $L^2_\mu(\mathcal{X};\mathcal{Y})$ and $\sup_{n\in\mathbb{N}} \|F_n\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} < \infty$, then $F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$.*

Proof. Since $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ is a Hilbert space, it is reflexive. As $(F_n)_{n\in\mathbb{N}}$ is bounded in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ by assumption, there exists a subsequence $(F_{n_k})_{k\in\mathbb{N}}$ which converges weakly in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ to some G as $k\to\infty$. Since $F_{n_k} \to F$ in $L^2_\mu(\mathcal{X};\mathcal{Y})$ by assumption, we conclude that $F = G$, and the claim follows. □

Next, we consider the ℓ^2-characterization of $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$, see Theorem 3.5. For $\boldsymbol{\gamma} \in \Gamma$ and $i \in \mathbb{N}$, we define $\boldsymbol{\gamma^{(i)}} = (\gamma_k^{(i)})_{k\in\mathbb{N}} \in \Gamma$ as follows: If $\gamma_i = 0$, we set $\boldsymbol{\gamma^{(i)}} := \boldsymbol{0}$, and if $\gamma_i > 0$, we set

$$\gamma_k^{(i)} := \begin{cases} \gamma_k - 1 & \text{if } k = i, \\ \gamma_k & \text{if } k \neq i. \end{cases}$$

Proposition C.9. *Let $F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$. Then, we have*

$$\frac{\partial}{\partial \eta_i} F = \sum_{\boldsymbol{\gamma}\in\Gamma} b_i \sqrt{\frac{\gamma_i}{\lambda_i}} \left(\int_{\mathcal{X}} F H_{\boldsymbol{\gamma},\boldsymbol{\lambda}} d\mu \right) H_{\boldsymbol{\gamma^{(i)}},\boldsymbol{\lambda}}, \quad \forall i \in \mathbb{N}, \tag{C.3}$$

and

$$\|F\|^2_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} = \sum_{\boldsymbol{\gamma}\in\Gamma} \left(1 + \sum_{i=1}^{\infty} b_i^2 \frac{\gamma_i}{\lambda_i} \right) \left\| \int_{\mathcal{X}} F H_{\boldsymbol{\gamma},\boldsymbol{\lambda}} d\mu \right\|^2_{\mathcal{Y}}. \tag{C.4}$$

Conversely, if for a family of vectors $(Y_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma}\in\Gamma} \subset \mathcal{Y}$, one has

$$\sum_{\boldsymbol{\gamma}\in\Gamma} \left(1 + \sum_{i=1}^{\infty} b_i^2 \frac{\gamma_i}{\lambda_i} \right) \|Y_{\boldsymbol{\gamma}}\|^2_{\mathcal{Y}} < \infty, \tag{C.5}$$

then

$$F := \sum_{\boldsymbol{\gamma}\in\Gamma} Y_{\boldsymbol{\gamma}} H_{\boldsymbol{\gamma},\boldsymbol{\lambda}} \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y}).$$

Proof. First, let us fix a cylindrical functional $\varphi \in \mathcal{F}C^1_b(\mathcal{X})$. By [23, Lemma 10.14], the partial derivatives of φ satisfy

$$\frac{\partial}{\partial \phi_i} \varphi = \sum_{\boldsymbol{\gamma}\in\Gamma} \sqrt{\frac{\gamma_i}{\lambda_i}} \left(\int_{\mathcal{X}} \varphi H_{\boldsymbol{\gamma},\boldsymbol{\lambda}} d\mu \right) H_{\boldsymbol{\gamma^{(i)}},\boldsymbol{\lambda}}, \quad \forall i \in \mathbb{N}. \tag{C.6}$$

Since $\frac{\partial}{\partial \eta_i}\varphi = b_i \frac{\partial}{\partial \phi_i}\varphi$, it follows follows from (C.6) together with orthonormality of the Hermite polynomials that

$$\int_{\mathcal{X}} \left(\frac{\partial}{\partial \eta_i} \varphi \right) H_{\boldsymbol{\gamma^{(i)}},\boldsymbol{\lambda}} d\mu = b_i \sqrt{\frac{\gamma_i}{\lambda_i}} \int_{\mathcal{X}} \varphi H_{\boldsymbol{\gamma},\boldsymbol{\lambda}} d\mu, \quad \forall i \in \mathbb{N}. \tag{C.7}$$

By linearity, this holds, in fact, for every cylindrical operator $\varphi \in \mathcal{F}C^1_b(\mathcal{X},\mathcal{Y})$.

Now suppose that $F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$. It suffices to prove (C.3) as (C.4) then follows by Parseval's identity (2.5). For this, it is enough to show that

$$\int_{\mathcal{X}} \left(\frac{\partial}{\partial \eta_i} F \right) H_{\boldsymbol{\gamma^{(i)}},\boldsymbol{\lambda}} d\mu = b_i \sqrt{\frac{\gamma_i}{\lambda_i}} \int_{\mathcal{X}} F H_{\boldsymbol{\gamma},\boldsymbol{\lambda}} d\mu, \quad \forall i \in \mathbb{N}. \tag{C.8}$$

By definition of $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$, there exists a sequence $(\varphi_n)_{n\in\mathbb{N}} \subset \mathcal{F}C^1_b(\mathcal{X},\mathcal{Y})$ such that $\lim_{n\to\infty} \varphi_n = F$ and $\lim_{n\to\infty} \frac{\partial}{\partial \eta_i}\varphi_n = \frac{\partial}{\partial \eta_i} F$ in $L^2_\mu(\mathcal{X};\mathcal{Y})$. We set $\varphi = \varphi_n$ in (C.7) and take the limit $n\to\infty$ to obtain (C.8).

Conversely, suppose that (C.5) holds for a sequence $(Y_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma}\in\Gamma} \subset \mathcal{Y}$. We fix an enumeration $\tau : \mathbb{N} \to \Gamma$ of Γ and define

$$F_n := \sum_{j=1}^{n} Y_{\boldsymbol{\tau(j)}} H_{\boldsymbol{\tau(j)},\boldsymbol{\lambda}}, \quad \forall n \in \mathbb{N}.$$

By Parseval's identity, we can bound the $L^2_\mu(\mathcal{X};\mathcal{Y})$-norm of the F_n by

$$\|F_n\|^2_{L^2_\mu(\mathcal{X};\mathcal{Y})} = \sum_{j=1}^{n} \|Y_{\boldsymbol{\gamma}}\|^2_{\mathcal{Y}} \leq \sum_{\boldsymbol{\gamma}\in\Gamma} \|Y_{\boldsymbol{\gamma}}\|^2_{\mathcal{Y}} \tag{C.9}$$

and the right-hand side is finite by (C.5). This implies, in particular, that $(F_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $L^2_\mu(\mathcal{X};\mathcal{Y})$. Hence, there exists some $F \in L^2_\mu(\mathcal{X};\mathcal{Y})$ such that

$$F_n \to F \quad \text{in } L^2_\mu(\mathcal{X};\mathcal{Y}) \text{ as } n \to \infty. \tag{C.10}$$

Since $F_n \in \mathcal{F}C^1_b(\mathcal{X},\mathcal{Y})$, we can set $\varphi = F_n$ in (C.7), and obtain

$$\frac{\partial}{\partial \eta_i} F_n = \sum_{j=1}^{n} b_i \sqrt{\frac{\tau(j)_i}{\lambda_i}} Y_{\boldsymbol{\tau(j)}} H_{\boldsymbol{\tau(j)}^{(i)}}, \quad \forall i \in \mathbb{N}.$$

Consequently, we can bound the $L^2_\mu(\mathcal{X}; HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y}))$-norm of the derivatives $D_{\mathcal{X}_{\boldsymbol{b}}} F_n$ by

$$\begin{aligned}
\int_{\mathcal{X}} \|D_{\mathcal{X}_{\boldsymbol{b}}} F_n(X)\|^2_{HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y})} d\mu(X) &= \int_{\mathcal{X}} \sum_{i=1}^{\infty} \left\| \frac{\partial}{\partial \eta_i} F_n(X) \right\|^2_{\mathcal{Y}} d\mu(X) \\
&= \sum_{i=1}^{\infty} \int_{\mathcal{X}} \left\| \sum_{j=1}^{n} b_i \sqrt{\frac{\tau(j)_i}{\lambda_i}} Y_{\tau(j)} H_{\tau(j)^{(i)}}(X) \right\|^2_{\mathcal{Y}} d\mu(X) \\
&= \sum_{i=1}^{\infty} \sum_{j=1}^{n} b_i^2 \frac{\tau(j)_i}{\lambda_i} \|Y_{\boldsymbol{\tau(j)}}\|^2_{\mathcal{Y}} \\
&\leq \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} b_i^2 \frac{\tau(j)_i}{\lambda_i} \|Y_{\boldsymbol{\tau(j)}}\|^2_{\mathcal{Y}},
\end{aligned} \tag{C.11}$$

and the right-hand side is again finite by (C.5). Combining (C.9) and (C.11), we conclude that the F_n are uniformly bounded in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$. By Lemma C.8, it then follows that $F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$. □

We now consider operators with a certain structure. They will become important in the proof of Theorem 3.9 in Appendix D. To this end, recall from Theorem C.4(iii) the orthonormal basis $\{\xi_i\}_{i\in\mathbb{N}}$ of $\mathcal{H}$ with $\xi_i = \sqrt{\lambda_i}\phi_i$, $R_\mu \hat{\xi}_i = \xi_i$, and $\hat{\xi}_i \in \mathcal{X}^*$. For every $F \in L^2_\mu(\mathcal{X};\mathcal{Y})$ and $n \in \mathbb{N}$, we define

$$\mathbb{E}_n F : \mathcal{X} \to \mathcal{Y}, \quad \mathbb{E}_n F := \mathbb{E}[F \mid \hat{\xi}_1, \dots, \hat{\xi}_n], \tag{C.12}$$

to be the conditional expectation of F with respect to the σ-algebra generated by the random variables $\hat{\xi}_1, \dots, \hat{\xi}_n$. Furthermore, for $n \in \mathbb{N}$, we define the mapping

$$P_n : \mathcal{X} \to \operatorname{span}\{\hat{\xi}_i : i \in [n]\}, \quad P_n X := \sum_{i=1}^{n} \hat{\xi}_i(X)\xi_i. \tag{C.13}$$

Proposition C.10 (Properties of $\mathbb{E}_n F$ in L^2_μ)**.** *For $F \in L^2_\mu(\mathcal{X};\mathcal{Y})$, let $\mathbb{E}_n F$ and P_n be given as in* (C.12) *and* (C.13)*, respectively. Then the following holds:*

(i) For every $F \in L^2_\mu(\mathcal{X};\mathcal{Y})$ and $n \in \mathbb{N}$,

$$\mathbb{E}_n F(X) = \int_{\mathcal{X}} F(P_n X + (I - P_n)Z) d\mu(Z) \quad \text{for } \mu\text{-a.e. } X \in \mathcal{X}.$$

In particular, $\mathbb{E}_n F$ can be identified with an operator on $P_n(\mathcal{X})$ by setting $F_n(Z) := \mathbb{E}_n F(X)$ for $Z = P_n(X)$.

(ii) For every $F \in L^2_\mu(\mathcal{X};\mathcal{Y})$, the sequence $(\mathbb{E}_n F)_{n\in\mathbb{N}}$ converges to F in $L^2_\mu(\mathcal{X};\mathcal{Y})$.

Proof. We refer to the proofs of Proposition 7.4.1 and Proposition 7.4.4, respectively, in [57], which can be adopted almost verbatim, only changing the Lebesgue integrals to Bochner integrals. □

Proposition C.10 can be extended to operators in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$:

Proposition C.11 (Properties of $\mathbb{E}_n F$ in $W^{1,2}_{\mu,\boldsymbol{b}}$)**.** *Let $F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ and let $\mathbb{E}_n F$ be defined as in* (C.12)*. We have $\mathbb{E}_n F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ for every $n \in \mathbb{N}$ and*

$$\lim_{n\to\infty} \mathbb{E}_n F = F \quad \text{in } W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y}).$$

The case of $\mathcal{H}$-differentiable functionals, that is, $\boldsymbol{b} = \sqrt{\boldsymbol{\lambda}}$ (see Remark 2.1) and $\mathcal{Y} = \mathbb{R}$, is covered by [57, Prop. 10.1.2] which reads as follows:

Proposition C.12. *Let $F \in W^{1,2}_{\mu,\sqrt{\boldsymbol{\lambda}}}(\mathcal{X})$. Then, for every $n \in \mathbb{N}$, we have $\mathbb{E}_n F \in W^{1,2}_{\mu,\sqrt{\boldsymbol{\lambda}}}(\mathcal{X})$ and the following properties hold:*

(i) For every $i \in \mathbb{N}$, the ith partial derivative of $\mathbb{E}_n F$ is given by

$$\frac{\partial}{\partial \xi_i} \mathbb{E}_n F = \begin{cases} \mathbb{E}_n(\frac{\partial}{\partial \xi_i} F) & \text{if } j \le n, \\ 0 & \text{if } j > n. \end{cases}$$

(ii) We have $\lim_{n\to\infty} \mathbb{E}_n F = F$ in $W^{1,2}_{\mu,\sqrt{\boldsymbol{\lambda}}}(\mathcal{X})$.

Proof of Proposition C.11. We first prove the claim for $F \in \mathcal{F}C^1_b(\mathcal{X},\mathcal{Y})$ and then for general operators in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ by density and a diagonal argument. Let us fix $F \in \mathcal{F}C^1_b(\mathcal{X},\mathcal{Y})$. By linearity, it suffices to consider cylindrical operators of the form $F(\cdot) = \varphi(\cdot)Y$ with $\varphi \in \mathcal{F}C^1_b(\mathcal{X})$ and $Y \in \mathcal{Y}$. Since $\eta_i = b_i \lambda_i^{-1/2} \xi_i$, we have

$$\frac{\partial}{\partial \eta_i} \mathbb{E}_n F(X) = b_i \lambda_i^{-1/2} \frac{\partial}{\partial \xi_i} \mathbb{E}_n F(X) = b_i \lambda_i^{-1/2} Y \frac{\partial}{\partial \xi_i} \mathbb{E}_n \varphi(X), \quad \forall X \in \mathcal{X}, \ \forall i \in \mathbb{N}.$$

Let $\{\psi_j\}_{j\in\mathbb{N}}$ be an orthonormal basis of $\mathcal{Y}$. Using Proposition C.12(i), Parseval's identity, and the contraction property of the conditional expectation, we can bound the norm of the derivative $D_{\mathcal{X}_{\boldsymbol{b}}}\mathbb{E}_n F$ by

$$\begin{aligned}\|D_{\mathcal{X}_{\boldsymbol{b}}}\mathbb{E}_n F\|^2_{L^2_\mu(\mathcal{X};HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y}))} &= \int_{\mathcal{X}} \sum_{i=1}^{\infty} \left\| \frac{\partial}{\partial \eta_i}\mathbb{E}_n F(X)\right\|^2_{\mathcal{Y}} d\mu(X) \\ &= \int_{\mathcal{X}} \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} b_i^2\lambda_i^{-1}\|Y\|^2_{\mathcal{Y}} \left| \left\langle \frac{\partial}{\partial \xi_i}\mathbb{E}_n\varphi(X), \psi_j\right\rangle_{\mathcal{Y}}\right|^2 d\mu(X) \\ &= \sum_{i=1}^{n} b_i^2\lambda_i^{-1}\|Y\|^2_{\mathcal{Y}} \sum_{j=1}^{\infty}\int_{\mathcal{X}} \left| \left\langle \mathbb{E}_n\left(\frac{\partial}{\partial \xi_i}\varphi\right)(X), \psi_j\right\rangle_{\mathcal{Y}}\right|^2 d\mu(X) \\ &= \sum_{i=1}^{n} b_i^2\lambda_i^{-1}\|Y\|^2_{\mathcal{Y}} \left\|\mathbb{E}_n\left(\frac{\partial}{\partial \xi_i}\varphi\right)\right\|^2_{L^2_\mu(\mathcal{X})} \\ &\le \sum_{i=1}^{n} b_i^2\lambda_i^{-1}\|Y\|^2_{\mathcal{Y}} \left\|\frac{\partial}{\partial \xi_i}\varphi\right\|^2_{L^2_\mu(\mathcal{X})} \\ &\le \int_{\mathcal{X}} \sum_{i=1}^{\infty} \left\| \frac{\partial}{\partial \eta_i} F(X)\right\|^2_{\mathcal{Y}} d\mu(X) = \|D_{\mathcal{X}_{\boldsymbol{b}}} F\|^2_{L^2_\mu(\mathcal{X};HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y}))}.\end{aligned} \tag{C.14}$$

Since in addition $\|\mathbb{E}_n F\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \le \|F\|_{L^2_\mu(\mathcal{X};\mathcal{Y})}$ by the contraction property of the conditional expectation, we conclude

$$\|\mathbb{E}_n F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} \le \|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})}, \quad \forall F \in \mathcal{F}C^1_b(\mathcal{X},\mathcal{Y}). \tag{C.15}$$

In particular, this shows that $\mathbb{E}_n F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ for every $n \in \mathbb{N}$. To prove convergence in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$, we first note that by Proposition C.10(ii), $\mathbb{E}_n F$ converges to F in $L^2_\mu(\mathcal{X};\mathcal{Y})$ as $n \to \infty$. Since φ is cylindrical, it can be written as $\varphi(X) = \omega(\hat{\xi}_1(X), \dots, \hat{\xi}_k(X))$ for some $k \in \mathbb{N}$. This implies

$$\|D_{\mathcal{X}_{\boldsymbol{b}}}\mathbb{E}_n F - D_{\mathcal{X}_{\boldsymbol{b}}} F\|_{L^2_\mu(\mathcal{X};HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y}))} \le \left(\max_{i\in[k]} b_i^2\lambda_i^{-1}\right)\|Y\|_{\mathcal{Y}} \left\|\nabla_{\mathcal{X}_{\sqrt{\boldsymbol{\lambda}}}}\mathbb{E}_n\varphi - \nabla_{\mathcal{X}_{\sqrt{\boldsymbol{\lambda}}}}\varphi\right\|_{L^2_\mu(\mathcal{X};\mathcal{X}_{\sqrt{\boldsymbol{\lambda}}})},$$

and the right-hand side converges to 0 as $n \to \infty$ by Proposition C.12(ii). Altogether, we conclude

$$\lim_{n\to\infty} \mathbb{E}_n F = F \quad \text{in } W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y}) \text{ for all } F \in \mathcal{F}C^1_b(\mathcal{X},\mathcal{Y}). \tag{C.16}$$

For the general case, let $F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$, and let $(F_k)_{k\in\mathbb{N}} \subset \mathcal{F}C^1_b(\mathcal{X},\mathcal{Y})$ be a sequence converging to F in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$. Again by the contraction property of the conditional expectation, we have

$$\|\mathbb{E}_n F_k - \mathbb{E}_n F\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \le \|F_k - F\|_{L^2_\mu(\mathcal{X};\mathcal{Y})}, \quad \forall n, k \in \mathbb{N},$$

which implies $\lim_{k\to\infty} \mathbb{E}_n F_k = \mathbb{E}_n F$ in $L^2_\mu(\mathcal{X};\mathcal{Y})$. In addition, by (C.15), $(\mathbb{E}_n F_k)_{k\in\mathbb{N}}$ is a Cauchy sequence in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$, which yields

$$\lim_{k\to\infty} D_{\mathcal{X}_{\boldsymbol{b}}}\mathbb{E}_n F_k = D_{\mathcal{X}_{\boldsymbol{b}}}\mathbb{E}_n F \quad \text{in } L^2_\mu(\mathcal{X}; HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y})).$$

Consequently, again by (C.15),

$$\begin{aligned}\|D_{\mathcal{X}_{\boldsymbol{b}}}\mathbb{E}_n F\|_{L^2_\mu(\mathcal{X};HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y}))} &= \lim_{k\to\infty}\|D_{\mathcal{X}_{\boldsymbol{b}}}\mathbb{E}_n F_k\|_{L^2_\mu(\mathcal{X};HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y}))} \le \lim_{k\to\infty}\|D_{\mathcal{X}_{\boldsymbol{b}}} F_k\|_{L^2_\mu(\mathcal{X};HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y}))} \\ &= \|D_{\mathcal{X}_{\boldsymbol{b}}} F\|_{L^2_\mu(\mathcal{X};HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y}))}.\end{aligned}$$

Together with $\|\mathbb{E}_n F\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \le \|F\|_{L^2_\mu(\mathcal{X};\mathcal{Y})}$, we conclude

$$\|\mathbb{E}_n F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} \le \|F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})}, \quad \forall F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y}), \tag{C.17}$$

and therefore $\mathbb{E}_n F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ for every $n \in \mathbb{N}$. To prove convergence, first note that, for any $k \in \mathbb{N}$,

$$\begin{aligned}\|\mathbb{E}_n F - F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} &\leq \|\mathbb{E}_n F - \mathbb{E}_n F_k\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} + \|\mathbb{E}_n F_k - F_k\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} + \|F_k - F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} \\ &\leq \|\mathbb{E}_n F_k - F_k\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} + 2\|F_k - F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})},\end{aligned}$$

where we used (C.17) in the second step. Next, let $\epsilon > 0$ be arbitrary and choose k large enough such that $\|F_k - F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} \leq \epsilon$. Then,

$$\|\mathbb{E}_n F - F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} \leq \|\mathbb{E}_n F_k - F_k\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} + 2\epsilon,$$

and taking the limsup $n \to \infty$ yields, by (C.16),

$$\limsup_{n\to\infty} \|\mathbb{E}_n F - F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} \leq 2\epsilon.$$

As this holds for any $\epsilon > 0$, the claim follows. □

We provide one more technical lemma which asserts that certain functionals lie in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X})$. For $X \in \mathcal{X}$ and $i \in \mathbb{N}$, we use use the notation $x_i := \langle X, \phi_i\rangle_{\mathcal{X}}$, where $\{\phi_i\}_{i\in\mathbb{N}}$ is the PCA basis of $\mathcal{X}$.

Lemma C.13. *Let $\varphi : \mathcal{X} \to \mathbb{R}$ be a functional of the form*

$$\varphi(X) = \omega(\lambda_1^{-1/2} x_1, \ldots, \lambda_n^{-1/2} x_n), \quad X \in \mathcal{X},$$

with $n \in \mathbb{N}$ and $\omega \in L^2(\mathbb{R}^n)$. Suppose that ω (possibly modified on an $\mathcal{L}^n$-null set) is absolutely continuous along each compact subinterval of almost every line parallel to one of the coordinate axes with (weak) partial derivatives in $L^2(\mathbb{R}^n)$. That is, there exists $g : \mathbb{R}^n \to \mathbb{R}$ such that $\psi = g$ $\mathcal{L}^n$-a.e., and for each $k \in [n]$, the functions

$$g_k(x,t) := g(x_1, \ldots, x_{k-1}, t, x_{k+1}, \ldots, x_n)$$

are absolutely continuous in t on compact subsets of $\mathbb{R}$ for $\mathcal{L}^{n-1}$-a.e. point $x = (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)$ in $\mathbb{R}^{n-1}$, and $\partial_t g_k \in L^2(\mathbb{R}^n)$. Then, $\varphi \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X})$ and

$$\frac{\partial}{\partial \eta_i}\varphi(X) = b_i \lambda_i^{-1/2} \partial_i \omega(\lambda_1^{-1/2} x_1, \ldots, \lambda_n^{-1/2} x_n)$$

for μ-a.e. $X \in \mathcal{X}$ and every $i \in [n]$.

Proof. First note that for $i \in [n]$, $X \in \mathcal{X}$, we have $\lambda_i^{-1/2} x_i = \hat{\xi}_i(X)$ with $\hat{\xi}_i \in \mathcal{X}^*$ given by (C.2). For brevity, we write $\hat{\boldsymbol{\xi}}(X) := (\hat{\xi}_1(X), \ldots, \hat{\xi}_n(X))$. The assumptions on ω imply that it lies in the space $W^{1,2}_{\mathrm{loc}}(\mathbb{R}^n)$ of weakly differentiable functions which, up to their first derivatives, are locally L^2-integrable, see, e.g., [30, Thm. 4.21]. Let $(\omega_j)_{j\in\mathbb{N}} \subset C_c^\infty(\mathbb{R}^n)$ be a sequence of smooth, compactly supported functions such that $\lim_{j\to\infty} \omega_j = \omega$ and $\lim_{j\to\infty} \partial_i \omega_j = \partial_i \omega$ in $L^2(\mathbb{R}^n)$ for every $i \in [n]$. We define

$$\varphi_j := \omega_j \circ \hat{\boldsymbol{\xi}}, \quad \forall j \in \mathbb{N},$$

and note that, by construction, $\varphi_j \in \mathcal{F}C_b^1(\mathcal{X})$. It is then easy to see that, as $j \to \infty$,

$$\varphi_j \to \varphi \quad \text{in } L^2_\mu(\mathcal{X})$$

as well as

$$\frac{\partial}{\partial \eta_i}\varphi_j(\cdot) \to b_i \lambda_i^{-1/2} \partial_i \omega(\hat{\boldsymbol{\xi}}(\cdot)) \quad \text{in } L^2_\mu(\mathcal{X}) \text{ for every } i \in [n].$$

This shows the claim by definition of $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X})$. □

D Proof of Theorem 3.9

We modify the argument in the proof of [57, Prop. 10.1.4]. Let $F : \mathcal{X} \to \mathcal{Y}$ be Lipschitz continuous with $L := [F]_{\mathrm{Lip}(\mathcal{X},\mathcal{Y})}$. The idea is to use Lemma C.8 to show that $F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$. By Lipschitz continuity, $F(X) \le F(0) + L\|X\|_{\mathcal{X}}$ for every $X \in \mathcal{X}$ and therefore, by Theorem C.1, $F \in L^2_\mu(\mathcal{X};\mathcal{Y})$. As approximating sequence to F we take $F_n := \mathbb{E}_n F$, as defined in (C.12). By Proposition C.10, we have $\lim_{n\to\infty} \mathbb{E}_n F = F$ in $L^2_\mu(\mathcal{X};\mathcal{Y})$, and we can write

$$\mathbb{E}_n F(X) = V_n(T_n(X))$$

for some $V_n : \mathbb{R}^n \to \mathcal{Y}$ and $T_n : \mathcal{X} \to \mathbb{R}^n$, $T_n(X) := (\hat{\xi}_1(X), \ldots, \hat{\xi}_n(X))$ with $\xi_i, \hat{\xi}_i$ defined in (C.1), (C.2). Note that V_n inherits Lipschitz continuity of F. Indeed, using Proposition C.10(i) and the fact that $\hat{\xi}_i(\xi_j) = \delta_{i,j}$, we find for $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^n$ that

$$\begin{aligned}
\|V_n(\boldsymbol{x}+\boldsymbol{z}) - V_n(\boldsymbol{x})\|_{\mathcal{Y}} &= \left\| \mathbb{E}_n F\left(\sum_{i=1}^n x_i\xi_i + \sum_{i=1}^n z_i\xi_i\right) - \mathbb{E}_n F\left(\sum_{i=1}^n x_i\xi_i\right)\right\|_{\mathcal{Y}} \\
&\le \int_{\mathcal{X}} \left\| F\left(\sum_{i=1}^n x_i\xi_i + \sum_{i=1}^n z_i\xi_i + (I-P_n)Y\right) - F\left(\sum_{i=1}^n x_i\xi_i + (I-P_n)Y\right)\right\|_{\mathcal{Y}} d\mu(Y) \\
&\le L\left\|\sum_{i=1}^n z_i\xi_i\right\|_{\mathcal{X}} = L\left\|\sum_{i=1}^n \sqrt{\lambda_i} z_i\phi_i\right\|_{\mathcal{X}} = L\left(\sum_{i=1}^n \lambda_i z_i^2\right)^{1/2} \qquad \text{(D.1)} \\
&\le L\left(\max_{i\in[n]} \sqrt{\lambda_i}\right)\|\boldsymbol{z}\|_{\mathbb{R}^n}. \qquad \text{(D.2)}
\end{aligned}$$

Proof of (i). Suppose that $\dim(\mathcal{Y}) = m \in \mathbb{N}$ and let $\{\psi_j\}_{j\in[m]}$ be an orthonormal basis of $\mathcal{Y}$. For $j \in [m]$, we define the function(al)s

$$\begin{aligned}
V_n^{(j)} &: \mathbb{R}^n \to \mathbb{R}, \quad x \mapsto \langle V_n(x), \psi_j\rangle_{\mathcal{Y}} \\
\mathbb{E}_n^{(j)} F &: \mathcal{X} \to \mathbb{R}, \quad X \mapsto \langle \mathbb{E}_n F(X), \psi_j\rangle_{\mathcal{Y}} = V_n^{(j)}(T_n(X)).
\end{aligned}$$

From (D.2) it follows that $V_n^{(j)}$ is Lipschitz continuous. By Rademacher's theorem it is therefore Fréchet differentiable $\mathcal{L}^n$-almost everywhere. Suppose that $X \in \mathcal{X}$ is a point such that $V_n^{(j)}$ is differentiable at $T_n(X)$. We can then apply the chain rule and since $\hat{\xi}_j(\eta_i) = b_i\lambda_i^{-1/2}\delta_{i,j}$, we get

$$D_{\mathcal{X}_{\boldsymbol{b}}}(\mathbb{E}_n^{(j)} F)(X)(\eta_i) = \begin{cases} \left\langle \nabla V_n^{(j)}(T_n(X)), b_i\lambda_i^{-1/2}\boldsymbol{e_i}\right\rangle_{\mathbb{R}^n} = b_i\lambda_i^{-1/2}\partial_i V_n^{(j)}(T_n(X)) & \text{if } 1 \le i \le n, \\ 0 & \text{otherwise.}\end{cases}$$

Recall that $\boldsymbol{e_i}$ denotes the ith standard unit vector in $\mathbb{R}^n$. To derive an upper bound for $|\partial_i V_n^{(j)}(\boldsymbol{x})|$, $\boldsymbol{x} \in \mathbb{R}^n$, we equip $\mathbb{R}^n$ with a rescaled Euclidean inner product,

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle_{\mathbb{R}^n_{\boldsymbol{\lambda}}} := \sum_{i=1}^n \sqrt{\lambda_i} x_i y_i$$

with induced norm $\|\boldsymbol{x}\|_{\mathbb{R}^n_{\boldsymbol{\lambda}}} := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x}\rangle_{\mathbb{R}^n_{\boldsymbol{\lambda}}}}$. We denote the hereby defined space as $\mathbb{R}^n_{\boldsymbol{\lambda}}$ and observe that $\{\lambda_i^{-1/2}\boldsymbol{e_i}\}_{i\in\mathbb{N}}$ is an orthonormal basis. Now note that (D.1) implies that $V_n^{(j)}$ is L-Lipschitz as a function from $\mathbb{R}^n_{\boldsymbol{\lambda}}$ to $\mathbb{R}$. Hence, for $\mathcal{L}^n$-almost every $\boldsymbol{x} \in \mathbb{R}^n$ and $1 \le i \le n$, it follows

$$\begin{aligned}
\sum_{i=1}^n \lambda_i^{-1}\left|\partial_i V_n^{(j)}(\boldsymbol{x})\right|^2 &= \sum_{i=1}^n \left|DV_n^{(j)}(\boldsymbol{x})(\lambda_i^{-1/2}\boldsymbol{e_i})\right|^2 \\
&= \left\|DV_n^{(j)}(\boldsymbol{x})\right\|^2_{HS(\mathbb{R}^n_{\boldsymbol{\lambda}},\mathbb{R})} = \left\|DV_n^{(j)}(\boldsymbol{x})\right\|^2_{L(\mathbb{R}^n_{\boldsymbol{\lambda}},\mathbb{R})} \le L^2.
\end{aligned} \qquad \text{(D.3)}$$

Consequently, whenever $V_n^{(j)}$ is differentiable at $T_n(X)$, we have for every $j \in [m]$,

$$\begin{aligned}
\|D_{\mathcal{X}_{\boldsymbol{b}}}(\mathbb{E}_n F)(X)\|^2_{HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y})} &= \left\| \sum_{j=1}^m D_{\mathcal{X}_{\boldsymbol{b}}}(\mathbb{E}_n^{(j)} F)(X)\psi_j \right\|^2_{HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y})} = \sum_{i=1}^\infty \left\| \sum_{j=1}^m D_{\mathcal{X}_{\boldsymbol{b}}}(\mathbb{E}_n^{(j)} F)(X)(\eta_i)\psi_j \right\|^2_{\mathcal{Y}} \\
&= \sum_{i=1}^\infty \sum_{j=1}^m \left| D_{\mathcal{X}_{\boldsymbol{b}}}(\mathbb{E}_n^{(j)} F)(X)(\eta_i) \right|^2 = \sum_{i=1}^n \sum_{j=1}^m b_i^2 \lambda_i^{-1} \left| \partial_i V_n^{(j)}(T_n(X)) \right|^2 \\
&\leq mL^2,
\end{aligned} \tag{D.4}$$

where we used in the last step that $b_i \leq 1$ for every $i \in \mathbb{N}$.

Next, we show that for μ-a.e. $X \in \mathcal{X}$ and every $j \in [m]$, $V_n^{(j)}$ is Fréchet differentiable at $T_n(X)$. To this end, let $A \subset \mathbb{R}^n$ be the set such that for every $j \in [m]$, $V_n^{(j)}$ is differentiable at every point $\boldsymbol{x} \in \mathbb{R}^n \setminus A$. Then each $V_n^{(j)}$ is Fréchet differentiable at any point $T_n(X)$ with $X \in \mathcal{X} \setminus T_n^{-1}(A)$. It thus suffices to show that $\mu(T_n^{-1}(A)) = 0$. We know that $\mathcal{L}^n(A) = 0$, and since μ_n is absolutely continuous with respect to $\mathcal{L}^n$, it follows that $\mu_n(A) = 0$. Moreover, it is easy to see that μ_n is equal to the push-forward measure $(T_n)_\sharp\mu$, and thus $\mu(T_n^{-1}(A)) = 0$. Hence, (D.4) holds for μ-a.e. $X \in \mathcal{X}$, which implies

$$\int_{\mathcal{X}} \|D_{\mathcal{X}_{\boldsymbol{b}}}(\mathbb{E}_n F)(X)\|^2_{HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y})} d\mu(X) \leq mL^2, \quad \forall n \in \mathbb{N}.$$

As $\mathbb{E}_n F \to F$ in $L^2_\mu(\mathcal{X};\mathcal{Y})$, we can now use Lemma C.8 to conclude that $F \in W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$. This shows that $\operatorname{Lip}(\mathcal{X},\mathcal{Y}) \subset W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$.

Next, we show that $C^{0,1}(\mathcal{X},\mathcal{Y})$ is continuously embedded in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$. By (D.4), we know that $\|D_{\mathcal{X}_{\boldsymbol{b}}}(\mathbb{E}_n F)(X)\|_{L^2_\mu(\mathcal{X};HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y})} \leq \sqrt{m}L$. Moreover, from Proposition C.10(i) it follows that $\|\mathbb{E}_n F\|_{L^2_\mu(\mathcal{X};\mathcal{Y})} \leq \sup_{X\in\mathcal{X}} \|F(X)\|_{\mathcal{Y}}$. In total, we have

$$\|\mathbb{E}_n F\|_{W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})} \leq \sup_{X\in\mathcal{X}} \|F(X)\|_{\mathcal{Y}} + \sqrt{m}L \leq \sqrt{m}\|F\|_{C^{0,1}(\mathcal{X},\mathcal{Y})}.$$

As $\lim_{n\to\infty} \mathbb{E}_n F = F$ in $W^{1,2}_{\mu,\boldsymbol{b}}(\mathcal{X};\mathcal{Y})$ by Proposition C.11, the claim follows.

Proof of (ii). Suppose that $\dim(\mathcal{Y}) = \infty$. Notice that in this case we cannot use the same argument as in the proof of (i) for two reasons. First, (D.4) does not give a meaningful bound for $m = \infty$. Second, and more subtly, a similar estimate as in (D.3) does not hold because equality of the operator norm and the Hilbert-Schmidt norm is only true for rank-one-operators. We thus have to argue differently.

To this end, assume that $\boldsymbol{b} \in \ell^2(\mathbb{N})$. We derive bounds for the partial derivatives $\partial_i V_n$ and use square-summability of $\boldsymbol{b}$ to ensure finiteness even if $\mathcal{Y}$ has infinite dimension. First, as $V_n : \mathbb{R}^n \to \mathcal{Y}$ is Lipschitz continuous, we can use the generalized Rademacher theorem to conclude that V_n is Hadamard differentiable (and, in fact, Fréchet differentiable since $\mathbb{R}^n$ has finite dimension) $\mathcal{L}^n$-almost everywhere in $\mathbb{R}^n$, see [11, Thm. 1 in Chpt. 2 & Rmk. 2 in Chpt. 1]. Suppose that $X \in \mathcal{X}$ is a point such that V_n is differentiable at $T_n(X)$. As in the proof of (i), we can then apply the chain rule to get

$$D_{\mathcal{X}_{\boldsymbol{b}}}(\mathbb{E}_n F)(X)(\eta_i) = \begin{cases} DV_n(T_n(X))(b_i\lambda_i^{-1/2}\boldsymbol{e_i}) = b_i\lambda_i^{-1/2}\partial_i V_n(T_n(X)) & \text{if } 1 \leq i \leq n, \\ 0 & \text{otherwise.} \end{cases}$$

Note that setting $\boldsymbol{z} = \boldsymbol{e_i}$ in (D.2) implies that $\|\partial_i V_n(\boldsymbol{x})\|_{\mathcal{Y}} \leq L\sqrt{\lambda_i}$ for $\mathcal{L}^n$-a.e. $\boldsymbol{x} \in \mathbb{R}^n$ and every $1 \leq i \leq n$. Consequently, whenever V_n is differentiable at $T_n(X)$, it follows that

$$\begin{aligned}
\|D_{\mathcal{X}_{\boldsymbol{b}}}(\mathbb{E}_n F)(X)\|^2_{HS(\mathcal{X}_{\boldsymbol{b}},\mathcal{Y})} &= \sum_{i=1}^\infty \|D_{\mathcal{X}_{\boldsymbol{b}}}(\mathbb{E}_n F)(X)(\eta_i)\|^2_{\mathcal{Y}} = \sum_{i=1}^n b_i^2\lambda_i^{-1}\|\partial_i V_n(X)\|^2_{\mathcal{Y}} \\
&\leq L^2 \sum_{i=1}^n b_i^2 \leq L^2\|\boldsymbol{b}\|^2_{\ell^2(\mathbb{N})}.
\end{aligned}$$

We can now proceed as in the proof of (i) to conclude the claim. □