Towards optimal sampling for learning sparse approximations in high dimensions

Ben Adcock, Juan M. Cardenas, Nick Dexter and Sebastian Moraga

Abstract In this chapter, we discuss recent work on learning sparse approximations to high-dimensional functions on data, where the target functions may be scalar-, vector- or even Hilbert space-valued. Our main objective is to study how the sampling strategy affects the sample complexity – that is, the number of samples that suffice for accurate and stable recovery – and to use this insight to obtain optimal or near-optimal sampling procedures. We consider two settings. First, when a target sparse representation is known, in which case we present a near-complete answer based on drawing independent random samples from carefully-designed probability measures. Second, we consider the more challenging scenario when such representation is unknown. In this case, while not giving a full answer, we describe a general construction of sampling measures that improves over standard Monte Carlo sampling. We present examples using algebraic and trigonometric polynomials, and for the former, we also introduce a new procedure for function approximation on irregular (i.e., nontensorial) domains. The effectiveness of this procedure is shown through numerical examples. Finally, we discuss a number of structured sparsity models, and how they may lead to better approximations.

1 Introduction

Learning an accurate approximation to an unknown function from data is a fundamental problem at the heart of many key tasks in applied mathematics and computer science. This problem is rendered challenging by the famous curse of dimensionality. In many relevant applications, the domain of the function is a high-dimensional space, thus standard algorithms (those well suited in lower dimensions) often suffer.
from an exponential blow-up in sample complexity (the number of samples required to obtain an accurate approximation). This is particularly problematic in many practical settings, since the amount of data available is often highly limited.

Fortunately, it is well known that functions arising in practice often possess low-dimensional structure. Specifically, they admit approximately sparse representations, meaning that they can be efficiently approximated using a relatively small number $s$ of functions from a particular dictionary. With this in mind, the aim of this chapter is to address the following fundamental question: supposing a function has an approximately sparse representation, how many samples (of a given type) suffice to learn such an approximation from data, and how can it be computed?

### 1.1 Main problem

Let $(\mathcal{D}, \mathcal{F}, \rho)$ be a probability space. Here $D$ is typically a subset of $\mathbb{R}^d$, where $d \gg 1$, is the dimension of the problem. We consider approximating functions defined over $D$. In many applications, such a function takes scalar values. However, other applications call for the approximation of functions that are vector- or function-space valued. To this end, in this work we let $V$ be a separable Hilbert space over the field $\mathbb{C}$ and consider a function of the form

$$f : D \rightarrow V.$$  

Note that $V$ may be taken as $(\mathbb{C}, |\cdot|)$ in the case of scalar-valued function approximation or $(\mathbb{C}^k, \|\cdot\|_2)$ in the case of vector-valued function approximation. Alternatively, it may be an infinite-dimensional Hilbert space of functions. We discuss several motivations for studying this case later. Note also that we consider vector spaces over complex fields. Doing so presents a number of additional challenges over considering the real case only.

Let $L^2_\rho(D)$ be the Lebesgue space of complex scalar-valued, square-integrable functions on $D$. We now consider a known dictionary of functions

$$\Phi = \{\phi_i : i \in \mathcal{I}\} \subset L^2_\rho(D),$$

which may be finite, countable or uncountable, and we assume that $f$ has an approximate $s$-sparse representation in $\Phi$. That is to say, there exists a set $S \subset \mathcal{F}$ of size $|S| \leq s$ for which

$$f \approx f_S := \sum_{i \in S} c_i \phi_i,$$

where the coefficients $c_i \in V$ are elements of the Hilbert space $V$.

Motivated by the high-dimensionality of the domain $D$, our primary focus in this work is on random sampling schemes. To this end, we assume that there are probability measures $\mu_1, \ldots, \mu_m$ on $D$, we draw $m$ independent samples $y_1, \ldots, y_m$ with $y_i \sim \mu_i$, $i = 1, \ldots, m$, and we assume that the data takes the form
Towards optimal sampling for learning sparse approximations in high dimensions

\[(y_i, f(y_i) + n_i), \quad i = 1, \ldots, m,\]  

(2)

where \(n_i \in \mathbb{V}\) is measurement noise. With this in hand, we may reformulate the main question stated previously as follows: how should one choose the number of samples \(m\), the sampling measures \(\mu_1, \ldots, \mu_m\), and the learning procedure so that an approximation to \(f\) yielding an error close to that of the sparse representation \(f_S\) can be computed from the data \(2\)? Furthermore, is this approximation stable to measurement noise?

Lacking any further insight, the standard random sampling strategy involves drawing samples in a Monte Carlo fashion from the underlying measure \(\rho\); in other words, we let \(\mu_1 = \ldots = \mu_m = \rho\). We consider this strategy the starting point for the discussion. To this end, we also consider the related question: what is the sample complexity of Monte Carlo sampling, and to what extent can this be improved by changing the sampling measures \(\mu_i\)?

Note that the focus of this work is on approximations that can be computed (potentially up to some tolerance) in finite time. When \(f\) takes values in an infinite-dimensional Hilbert space \(\mathbb{V}\), this presents an issue. To address it, we assume the \(\mathbb{V}\) can be discretized via a finite-dimensional space \(\mathbb{V}_h\) (here \(h > 0\) is a discretization parameter), and then proceed to perform computations in \(\mathbb{V}_h\), as opposed to \(\mathbb{V}\). Thus, another important question we discuss in this chapter is: what is the effect of this discretization on the ensuing approximation to \(f\)?

1.2 Overview

The purpose of this chapter is to survey a recent body of work that has sought to answer these questions. See \(\S1.4\) for a detailed summary of relevant literature. We divide our discussion into two main cases:

First, we consider the case where the target set \(S\) in the sparse representation \(1\) is known. This is by far the simpler situation, yet it can indeed occur in certain problems arising in practice. For example, \(S\) may be obtained by \(a\ priori\) regularity estimates on \(f\). Moreover, even though it may not be applicable in general, examining this case helps provide insight into what can possibly be achieved in the second setting, where \(S\) is unknown.

We provide an almost complete set of answers to the above questions in this first setting. The approximation is learned through a simple (weighted) least-squares fit, which is readily shown to provide accurate and stable approximations. We also obtain a general condition on the sampling measures \(\mu_i\), as well as explicit examples satisfying such a condition, for which only

\[m \gtrsim s \cdot \log(2s/\epsilon)\]  

(3)

such samples suffice for recovery, with probability at least \(1 - \epsilon\) for some \(\epsilon > 0\). This condition is optimal up to the constant implied by the \(\gtrsim\) symbol and the log
factor. As we also discuss, the near-optimal sample complexity bound typically does not hold in the case of Monte Carlo sampling. We discuss examples where the corresponding bound for Monte Carlo sampling can be arbitrarily large.

§3 is devoted to weighted least-squares approximation.

Unfortunately, the first case is rather rare in practice. It is more common to encounter the situation where \( S \) is unknown a priori. To overcome this, one may seek to estimate \( S \) by employing adaptive sampling while building \( S \) in an iterative manner, typically via a greedy scheme. While such procedures can sometimes work well in practice – especially when \( s \) is relatively small – they often lack theoretical guarantees. Instead, we pursue a different approach using tools from sparse regularization, in which we seek to promote the sparsity of \( f \) in the dictionary \( \Phi \) via \( \ell_1 \)-minimization-type techniques. Analysis of this case can then be performed using tools from compressed sensing theory.

In this case, we assume that \( \Phi \) is a finite set of linearly-independent elements. Let \( n = |\Phi| \). Our main result on sample complexity in this case demonstrates that there exist choices of sampling measures \( \mu_1, \ldots, \mu_m \) for which

\[
m \gtrsim \left( \frac{b}{a} \right) \cdot \left( \frac{\theta^2}{a} \right) \cdot s \cdot \left( \log(en) \cdot \log\left( e\left( \frac{b}{a} \right) \left( \frac{\theta^2}{a} \right) s \right) + \log\left( \frac{2}{\varepsilon} \right) \right)
\]

sample suffice for recovery, where \( a, b > 0 \) are the Riesz basis constants of \( \Phi \) (see (45)). Here \( \theta \) is an explicit constant, given by

\[
\theta^2 = \int_D \max_{\psi \in \mathcal{F}} |\psi(y)|^2 \, d\rho(y).
\]

We also present a sample complexity bound for Monte Carlo sampling, which takes the form

\[
m \gtrsim \left( \frac{b}{a} \right) \cdot \left( \Theta^2 / a \right) \cdot s \cdot \left( \log(en) \cdot \log\left( e\left( \frac{b}{a} \right) \left( \frac{\theta^2}{a} \right) s \right) + \log\left( \frac{2}{\varepsilon} \right) \right),
\]

where

\[
\Theta^2 = \max_{i \in \mathcal{F}} \|\phi_i\|^2_{L^2(D)}.
\]

Notice that \( \theta \leq \Theta \). Hence the former strategy is always at least as good as Monte Carlo sampling. We present examples where \( \theta = \Theta = 1 \) (in which case, Monte Carlo sampling is sufficient) and where \( \theta \ll \Theta \) (in which case, the former strategy is strictly better).

### 1.3 Additional contributions

In tandem with the various sample complexity bounds, we also present error bounds for the learned approximations. These show that such approximations are accurate – i.e. the error is bounded by the best approximation error \( f - f_S \), measured in some norm – and stable to noise, i.e. the error scales linearly with the noise values.
Towards optimal sampling for learning sparse approximations in high dimensions

In the Hilbert-valued setting, we also determine stability to discretization error, in the sense that the error involves an additional term that is proportional to the orthogonal projection onto $\mathbb{V}_k$.

Several of our examples consider function approximation on tensor-product domains, such as the symmetric hypercube $D = [-1,1]^d$ in $d$ dimensions. However, certain practical applications result in approximation problems on irregular domains. Another contribution of this chapter is to introduce a new approach for function approximation on irregular domains via sparse regularization. We demonstrate the efficacy of this new approach both through theoretical guarantees and numerical examples.

Finally, we also discuss settings where $f$ admits a structured sparse approximation in the dictionary $\Phi$. Such representations arise frequently in practice, and can lead to tangible benefits in accuracy. We consider two such models, weighted sparsity and lower set sparsity, and briefly describe the extension of the main results to these settings. Focusing on the irregular domain case, we also showcase the benefits of such structured sparsity models via numerical examples.

1.4 Related literature

This work is motivated in great part by applications arising in parametric models. Here, one seeks to understand how the parameters in a physical model – a weather or climate model, a chemical or biological process, a fluid flow model such as groundwater flow, a nuclear reactor, an aircraft engine, etc – affect its output. Parametric models are ubiquitous in engineering and the physical sciences. Approximating the input-output map of a parametric model is a problem that lies at the heart of many key tasks in parametric modelling, such as performing uncertainty quantification, parameter optimization or solving parametric inverse problems. See [42, 61, 83, 85] for detailed introductions to this topic.

Parametric models are often formulated as (systems of) DEs. In such problems, the function $f$ is the solution $u$ of a PDE system of the form

$$\mathcal{L}_x(u,y) = 0,$$

defined over a physical domain $\Omega$ and subject to suitable boundary conditions, where $\mathcal{L}_x(\cdot,y)$ denotes a differential operator in the physical variable $x$ which depends on $y$. Therefore, the solution $u$ is also a function defined over $\Omega \times D$, and for each fixed $y \in D$ the solution $u(\cdot,y)$ is an element of a function space $\mathbb{V}$. Note here that $\mathbb{V}$ may be a Hilbert or a Banach space, depending on the particular form of $\mathcal{L}_x$. The typical goal in such settings is then to compute a quantity of interest (QoI) $Q : \mathbb{V} \rightarrow \mathbb{R}$ depending on $u$, e.g., the expectation or variance of $u$ with respect to $y$ at certain points $(x_i)$ or the integral of $u$ with respect to the physical variable $x$ as a function of $y$. Depending on the task at hand, a number of quantities of interest
may be required, and in such scenarios computation of a fast surrogate of the full parameter-to-solution map $y \mapsto u(\cdot, y) \in \mathcal{V}$ is desirable.

Generally speaking, evaluating $u(\cdot, y)$ (or some QoI of $u$) at a fixed value of $y$ is expensive. This either involves a costly physical experiment, or a computationally-intensive numerical simulation to (approximately) solve (4). Hence, the objective approximate $u$, or some QoI, from as few sample values

$$u(\cdot, y_1), \ldots, u(\cdot, y_m),$$

as possible. There are many different approaches to effect such an approximation, many of which seek to exploit low-dimensional structure of the solution $u$, typically in the form of sparsity with respect to a dictionary. Amongst the most popular methods are those which use a basis of algebraic polynomials (termed polynomial chaos expansions in uncertainty quantification), which are motivated by the fact that solutions of many parametric DEs (4) are smooth functions of their parameters. But there are also techniques based on multiscale or hierarchical bases, radial basis functions, trigonometric polynomials, and various others. Furthermore, there are adaptive or learned bases methods, such as, most recently, techniques involving deep neural networks.

The systematic study of least-squares approximation in general finite-dimensional subspaces from Monte Carlo samples began with the work of [28], with a focus on spaces of algebraic polynomials. Other early works on algebraic polynomials include [22, 66, 70]. It was observed that Monte Carlo sampling can lead to large sample complexities or poor approximations, which in turn led to a series of investigations into the design of improved sampling strategies. See [9, 38, 41, 48, 49, 72, 82, 88, 103, 105] and references therein. The matter of optimal sampling was theoretically resolved in [49] for specific polynomial subspaces, and later [32] for general spaces. However, drawing samples from the resulting measures may not always be straightforward in practice. The measures are also nonadaptive. This led to various further extensions, including adaptive strategies [16, 67], more practical approaches based on discrete measures [10, 38, 68] and recent work on boosting [38, 47]. For other reviews of this topic, see [9, 33, 45, 48].

The application of $\ell_1$-minimization for computing sparse polynomial approximations of functions was first considered in [18, 39, 65, 79, 97]. This led to substantial amounts of subsequent research, including [3, 4, 13, 14, 19, 26, 26, 43, 44, 52, 55, 57, 63, 72, 76, 78, 81, 86, 90, 91, 94, 95, 98, 102]. Specific extensions to weighted and lower sparsity models were developed in [1, 3, 5, 9, 25, 75, 80, 99]. The generalization to Hilbert-valued functions was considered in [36]. As in the case of least squares, Monte Carlo sampling can lead to poor sample complexity bounds. Thus, a series of works considered improved sampling strategies [15, 37, 46, 50, 59, 62, 87, 96]. Weighted and lower set sparsity were developed in series of works [2, 8, 25, 80]. For additional reviews of this topic, see [8, 9, 51, 60, 63, 64, 73].
1.5 Outline

The remainder of this chapter surveys the topic of constructing sparse approximations to scalar- or Hilbert-valued functions from sample values via least squares or \( \ell_1 \)-minimization. Our focus is on the question of sampling, and, in particular, whether or not optimal sampling can be achieved. We combine ideas from many of the aforementioned works, which are generally specific to polynomial approximations, and describe them in the setting of general dictionaries of functions.

The outline of the remainder of this chapter is as follows. First, in §2 we introduce various preliminary concepts and notation. We then formalize the main problem and three main questions, and introduce the main examples considered later to highlight the main results. Next, in §3 we consider least-squares approximation. We provide definitive answers to all three main questions, and present several numerical examples. In §4 we consider \( \ell_1 \)-minimization. We present a series of theoretical results and then describe the extent to which they resolve the three main questions. In §6 we consider the extension to weighted and lower set sparsity models. Finally, we end in §7 with some conclusions and open problems.

2 Preliminaries

In this section, we first provide some key notation, then we describe the setup and main problems in further detail.

2.1 Notation

As noted, throughout \((D, \mathcal{D}, \rho)\) is a probability space and \( V \) is separable Hilbert space over the field \( \mathbb{C} \) with inner product \( \langle \cdot, \cdot \rangle_V \) and corresponding norm \( \| v \|_V = \sqrt{\langle v, v \rangle_V} \) for \( v \in V \). We write \( L^2_\rho(D; V) \) for the Lebesgue–Bochner space of functions \( f : D \to V \) for which the norm

\[
\| f \|_{L^2_\rho(D; V)} := \left( \int_D \| f(y) \|^2_V \, d\rho(y) \right)^{1/2} < \infty.
\]

We also write \( L^\infty(D; V) \) for the Lebesgue–Bochner space of functions \( f : D \to V \) for which the norm

\[
\| f \|_{L^\infty_\rho(D; V)} := \text{ess sup}_{y \in D} \| f(y) \|_V < \infty.
\]

Note that we also denote the classical Lebesgue spaces of complex-valued functions \( f : D \to \mathbb{C} \) as \( L^2(D) \) and \( L^\infty(D) \). We write \( \| \cdot \|_{L^2(D)} \) and \( \| \cdot \|_{L^\infty(D)} \) for their norms, respectively. These coincide with the Lebesgue–Bochner spaces whenever \( V \) is taken as \( \mathbb{C} \) with the obvious inner product.
Given an index set $\mathcal{J}$ that is at most countable, we write $\ell^p(\mathcal{J}; \mathbb{V})$ for the $\ell^p$ space of $\mathbb{V}$-valued sequences $(v_\iota)_{\iota \in \mathcal{J}}$ with finite $\ell^p$-norm, defined by

$$\|v\|_{\ell^p(\mathcal{J}; \mathbb{V})} = \left( \sum_{\iota \in \mathcal{J}} \|v_\iota\|_V^p \right)^{1/p}, \quad 1 \leq p < \infty,$$

$$\|v\|_{\ell^p(\mathcal{J}; \mathbb{V})} = \sup_{\iota \in \mathcal{J}} \|v_\iota\|_V, \quad p = \infty.$$ 

When $p = 2$, we also write $\langle \cdot, \cdot \rangle_{\ell^2(\mathcal{J}; \mathbb{V})}$ for its inner product. Note that when $\mathbb{V} = \mathbb{C}$, we write $\ell^p(\mathcal{J})$ and $\|\|_{\ell^p(\mathcal{J})}$, or simply $\|\|$ when the choice of $\mathcal{J}$ is clear.

Likewise, for $p = 2$ and $\mathbb{V} = \mathbb{C}$, we write $\langle \cdot, \cdot \rangle$ for the $\ell^2$-inner product on $\ell^2(\mathcal{J})$.

As discussed above, the space $\mathbb{V}$ may be infinite dimensional. Hence, performing computations in $\mathbb{V}$ directly is often not possible. To this end, we introduce a finite-dimensional discretization of $\mathbb{V}$, denoted by $\mathbb{V}_h$, where $h > 0$ is a discretization parameter. We assume that $\mathbb{V}_h \subseteq \mathbb{V}$ is a subspace of $\mathbb{V}$ and write $\mathcal{P}_h : \mathbb{V} \rightarrow \mathbb{V}_h$ for the orthogonal projection onto this subspace. Further, given $f \in L^2_\rho(D; \mathbb{V})$, we write $\mathcal{P}_h f \in L^2_\rho(D; \mathbb{V}_h)$ for the almost everywhere defined function given by

$$(\mathcal{P}_h f)(y) = \mathcal{P}_h(f(y)), \quad y \in D.$$ 

When necessary, we also employ a (not necessarily orthonormal basis) of $\mathbb{V}_h$. We write $\{ \psi_i \}_{i=1}^k$ for such a basis, where $k = \dim(\mathbb{V}_h)$.

Finally, we require a few additional pieces of notation. For convenience, we write $[n] := \{1, \ldots, n\}$ for $n \in \mathbb{N}$. We also use the notation $A \lesssim B$ to mean that there exists a numerical constant $c > 0$ such that $A \leq cB$, and likewise for $A \gtrsim B$. Further, we write $A \lesssim_x B$ if $A \leq c_x B$ for some constant $c_x > 0$ depending on a variable $x$, and likewise for $A \gtrsim_x B$.

### 2.2 Problem and key questions

As above, we let $\Phi = \{ \phi_t : t \in \mathcal{J} \} \subset L^2_\rho(D)$ be a dictionary and $f \in L^2_\rho(D; \mathbb{V})$ be the function we seek to learn. We consider $s$-sparse representations of $f$ of the form

$$f_S := \sum_{t \in S} c_t \phi_t, \quad c_t \in \mathbb{V},$$ 

where $S \subseteq \mathcal{J}$, $|S| \leq s$ is a subset of $s$ indices. We now formalize the two main settings we consider in this work:

**Problem 1 (Sparsity in a known subset).** The function $f$ has an approximate $s$-sparse representation of the form (5) for some known set $S$. 
Problem 2 (Sparsity in an unknown subset). The function $f$ has an approximate $s$-sparse representation of the form (5) for some unknown set $S$.

As discussed above, we consider sample points drawn randomly according to probability measures $\mu_1, \ldots, \mu_m$. We term these the sampling measures. We make the following assumption:

Assumption 1 (Absolute continuity and positivity) The additive mixture

$$\mu := \frac{1}{m} \sum_{i=1}^{m} \mu_i,$$

is absolutely continuous with respect to $\rho$ and moreover its Radon–Nikodym derivative is strictly positive almost everywhere on $\text{supp}(\mu)$.

This means that we can write

$$\frac{1}{m} \sum_{i=1}^{m} d\mu_i(y) = \frac{1}{w(y)} d\rho(y), \quad (6)$$

where $w: D \to \mathbb{R}$ is finite almost everywhere on $\text{supp}(\mu)$. We refer to $w$ as the weight function. Note that it satisfies

$$\int_D \frac{1}{w(y)} d\rho(y) = 1. \quad (7)$$

Given such sampling measures, we now draw samples $y_i \sim \mu_i$, $i = 1, \ldots, m$, independently from these measures and consider noisy data of the form

$$(y_i, f(y_i) + n_i) \in D \times \mathbb{V}_h, \quad i = 1, \ldots, m. \quad (8)$$

Here, the $n_i \in \mathbb{V}$ are terms that capture the measurement error. We focus on the case where these terms are small in norm, but we do not assume they follow a specific distribution (e.g. Gaussian noise in the scalar or vector-valued case). Note that we assume the measurements $f(y_i) + n_i$ are elements of the finite-dimensional space $\mathbb{V}_h$. Our motivation for doing so is the following. Since $f$ is $\mathbb{V}$-valued, the noiseless sample $f(y_i)$ is an element of the (potentially) infinite-dimensional Hilbert space $\mathbb{V}$. In general, this quantity cannot be stored, let alone used as the input to an algorithm for learning an approximation to $f$. Hence, we assume that the measurements are elements of the finite-dimensional subspace $\mathbb{V}_h$, which means they can be both stored – for example, by storing their coefficients with respect to the basis $\{\psi_i\}_{i=1}^k$ for $\mathbb{V}_h$ – and used as input to a learning algorithm. Note that the quantity $n_i \in \mathbb{V}$ accounts for both the discretization error in approximating the true sample $f(y_i)$ by an element of $\mathbb{V}_h$, as well as any other errors that arise in the measurement process (e.g. noise, numerical error, and so forth). Further, we do not specify how measurements are processed to give elements of $\mathbb{V}_h$; we simply assume that this process yields an error that can be captured by the generic noise term $n_i$. In other words, the model (8) is very general, and therefore sufficient for many applications.
We now formalize the three main questions considered in this work:

**Question 1.** Suppose \( f \) satisfies either Problem 1 or 2. How does one learn an approximation to \( f \) from the data \((8)\) that is accurate – i.e. the approximation error is bounded by the errors \( f - f_S \) and, in the case of Hilbert-valued functions, \( f - P_h(f) \), measured in suitable norms – and stable – i.e. the error depends moderately on the noise values \( n_i \), measured in a suitable norm?

**Question 2.** In the setting of either Problem 1 or 2 how many samples \( m \) are sufficient to obtain such an approximation, and how does the choice of sampling measures affect this bound?

**Question 3.** In the setting of either Problem 1 or 2, does Monte Carlo sampling lead to near-optimal sample complexity – i.e. linear in \( s \), up to constants and log factors? If not, is there a near-optimal choice of sampling measures?

### 2.3 Examples

We now introduce the main examples considered in this work.

**Example 1 (Trigonometric polynomial approximation on the \( d \)-torus).** Let \( d \geq 1 \), \( D = \mathbb{T}^d \) is the unit torus in \( d \) dimensions and \( dp(y) = dy \) be the uniform measure. In this example, we set \( I = \mathbb{Z}^d \) and consider the set of functions

\[
\phi_t(y) = \exp(2\pi i t \cdot y), \quad t = (t_1, \ldots, t_d) \in I, \quad y = (y_1, \ldots, y_d) \in D.
\]

Observe that the dictionary \( \Phi = \{ \phi_t : t \in I \} \) forms an orthonormal basis of \( L^2_{p_\rho}(D) \).

Trigonometric polynomial approximation of smooth and periodic functions in high dimensions is a classical topic \[35, 77, 89\]. The nature of the Fourier basis makes it a relatively straightforward case to study, and as we see later, this also yields clear answers to Questions 1–3.

Unfortunately, many problems – in particular parametric model problems – do not involve periodic functions. Since such functions are often smooth, however, this motivates the study of algebraic polynomial approximations:

**Example 2 (Algebraic polynomial approximation in the symmetric hypercube).** Let \( D = [-1, 1]^d \) be the symmetric hypercube in \( d \geq 1 \) dimensions of side length 2, and \( dp(y) = 2^{-d} dy \) be the uniform measure. We set \( I = \mathbb{N}^d_0 \) and consider

\[
\phi_t(y) = p_{t_1}(y_1) \cdots p_{t_d}(y_d), \quad t = (t_1, \ldots, t_d) \in I, \quad y = (y_1, \ldots, y_d) \in D,
\]

where, on the right-hand side, \( p_t \) denotes the one-dimensional Legendre polynomial of degree, \( t \in \mathbb{N}_0 \), normalized with respect to the one-dimensional uniform measure. Note that \( p_1(y) = \sqrt{2t+1}P_1(y) \), where \( P_1(y) \) is the classical Legendre polynomial with normalization \( P_1(1) = 1 \).
As observed, algebraic polynomial approximation is used widely in parametric model problems. This is motivated by the observation that many classes of parametric differential equations are holomorphic (analytic) functions of their parameters (see [9, 23, 24, 29, 30, 53, 56, 92] and references therein). This means the polynomial coefficients decay rapidly, yielding approximately sparse representations in a given polynomial basis. Note that Problem 1 is naturally motivated by such problems. For certain classes of parametric differential equations, one can use \textit{a priori} analysis to determine coefficient estimates, and using these obtain a candidate set $S$. However, this is not feasible for more complicated parametric differential equations, or problems where $f$ is given as a black box. In this setting we resort to Problem 2.

Note that it is also common to consider other systems of polynomials. Common examples include Chebyshev polynomials on $[-1, 1]^d$, Laguerre or Hermite polynomials on $[0, \infty)^d$ and $\mathbb{R}^d$, respectively, or nonorthogonal polynomials such as Taylor polynomials. For succinctness we consider Legendre polynomials only, although our analysis readily extends to more general settings.

Many polynomial approximation problems are naturally formulated on compact hyperrectangles. Using a change of variables, these can all be reduced to the setting of Example 2. In parametric models, this is inspired by the notion that the parameters are independent, with each one varying between a finite upper and lower value. However, this assumption can fail in practice. In parametric models, for example, there may often be dependencies between the parameters [40, 58, 61, 84]. This leads to polynomial approximation problems where the domain $D$, while still compact, is no longer a hypercube but an irregular-shaped domain. This poses a number of challenges, which we shall review later in this chapter. It motivates our third and final example:

Example 3 (Polynomial approximation on general compact domains). Let $D \subseteq \mathbb{R}^d$ be a measurable set with nonzero measure. We assume without loss of generality that $D \subseteq [-1, 1]^d$ is contained in the symmetric, $d$-dimensional hypercube with side length 2. Following a well-known approach, studied in detail in [12], we then construct a polynomial dictionary by restricting the orthonormal basis of Example 2 to $D$. For consistency of notation, we continue to denote this dictionary as $\Phi \rightarrow \{\phi_i : i \in \mathcal{I}\}$. We also let $\rho$ denote the uniform measure on $D$, i.e. $d\rho(y) = |D|^{-1} dy$. An important observation in this case is that, in contrast to the previous two examples, this dictionary is not a basis of $L^2_\rho(D)$. Rather, it forms a frame [27]. On the other hand, every finite set of elements from $\Phi$ is linearly independent – indeed, no finite linear combination of polynomials can vanish on a set of nonzero measure – and therefore a Riesz basis for its span. These two observations will be particularly important later when we consider $\ell^1$-minimization techniques in the setting of Problem 2.
2.4 Multi-index sets

Notice that the dictionary $\Phi$ in Example 1, 2 or 3 is indexed over a multi-index $\iota \in \mathcal{I} = \mathbb{Z}^d$ or $\iota \in \mathcal{I} = \mathbb{N}_0^d$. It is useful to define a number of standard choices for finite subsets of $\mathcal{I}$. In the case of Problem 1 such an index set could be used as a potential choice for $S$. Whereas in Problem 2 we see later that it is important to truncate the infinite set of multi-indices $\mathbb{N}_0^d$ or $\mathbb{Z}^d$ to some finite, but large subset in which we expect the indices of the sparse representation to belong.

Several standard subsets are the tensor product index set

$$\mathcal{S}^{TP}_t = \left\{ \iota = (\iota_k)_{k=1}^d : \iota_k \leq t, \ k = 1, \ldots, d \right\},$$

(10)
of order $t \in \mathbb{N}_0$, the total degree index set

$$\mathcal{S}^{TD}_t = \left\{ \iota = (\iota_k)_{k=1}^d : \sum_{k=1}^d \iota_k \leq t \right\},$$

(11)
of order $t$, and the hyperbolic cross index set

$$\mathcal{S}^{HC}_t = \left\{ \iota = (\iota_k)_{k=1}^d : \prod_{k=1}^d (\iota_k + 1) \leq t + 1 \right\},$$

(12)
of order $t$. Note that these index sets are subsets of $\mathbb{N}_0^d$, and therefore suitable for Examples 2 and 3. We define analogous subsets of $\mathbb{Z}^d$ for Example 1 simply by replacing $\iota_k$ by its absolute value $|\iota_k|$ in (10)–(12).

The choices (10)–(12) are commonly used in low to moderate dimensional problems when selecting the index set $S$ in the setting of Problem 1. However, their respective cardinalities grow rapidly with dimension; this is in particular true of (10), whose cardinality is $(n+1)^d$. This makes their applicability limited in higher dimensions, as, for a fixed maximum cardinality $s$, it may be impossible to achieve high orders, which are generally necessary and correspond to better accuracy. Indeed, in higher dimensions, it is often important to incorporate anisotropy into the index set $S$ to take into account different rates of variation of the function in different coordinate directions. By contrast, the index sets (10)–(12) are isotropic: indices in $S$ remain in $S$ when their entries are permuted. While it is possible to define anisotropic versions of each of these index sets (see, for example, [17]), the challenge becomes to set the anisotropy parameters in an a priori manner without knowledge of the underlying function $f$. Instead, we adopt the setting of Problem 2 and suppose $f$ has a sparse representation in some unknown index set $S$ contained within a larger, but finite index set of the above form – the goal then being to compute an approximation achieving a similar error as that of the sparse representation, without necessarily computing $S$ itself.
3 Sparse approximation via (weighted) least squares

We first suppose that Problem 1 holds and also that \( m \geq s \). Let \( S \subseteq \mathcal{S} \), \(|S| \leq s\), be the corresponding subset, and define the resulting subspace

\[
P_S;V = \left\{ \sum_{i \in S} c_i \phi_i : c_i \in V \right\} \subset L_2^p(D;V).
\]

Note that if \( V = \mathbb{C} \), we simply write \( P_S \) for subspace of complex-valued functions \( P_S = P_S;C = \left\{ \sum_{i \in S} c_i \phi_i : c_i \in \mathbb{C} \right\} \subset L_2^p(D) \). Next, we recall the discretized subspace \( V_h \) of \( V \) and the noisy samples (8). With this in hand, we follow a similar approach of [32] (which considers only the real scalar-valued case) and define the weighted least-squares approximation to \( f \) as:

\[
\hat{f} \in \arg\min_{p \in P_S;V_h} \left\{ \frac{1}{m} \sum_{i=1}^{m} w(y_i) \| f(y_i) + n_i - p(y_i) \|_2^2 \right\}.
\] (13)

Here \( w \) is the weight function specified in (6). Notice that we form an approximation in the subspace \( P_S;V_h \), as opposed to \( P_S;V \), since we generally cannot perform computations over the infinite-dimensional Hilbert space \( V \). In the scalar-valued case, we simply have \( V_h = V = (\mathbb{C}, \| \cdot \|) \).

3.1 Computation of the least-squares approximation

We first describe the computation of the approximation (13). Since any \( p \in P_S;V_h \) can be expressed \( p = \sum_{i \in S} c_i \phi_i \) with \( c_i \in V_h \), we can rewrite (13) as

\[
\hat{f} = \sum_{i \in S} \hat{c}_i \phi_i, \quad \hat{c} = (\hat{c}_i)_{i \in S} \in \arg\min_{\hat{c} \in V_h} \| A \hat{c} - v \|_2^2,
\] (14)

where

\[
A = \frac{1}{\sqrt{m}} \left( \sqrt{w(y_i)} \phi_i(y_i) \right)_{i \in [m], j \in [s]} \in \mathbb{C}^{m \times s},
\]

\[
v = \frac{1}{\sqrt{m}} \left( \sqrt{w(y_i)} (f(y_i) + n_i) \right)_{i \in [m]} \in \mathbb{R}^m,
\] (15)

and \( \{t_1, \ldots, t_s\} \) is an enumeration of the indices in \( S \). Note that we consider \( A \) both as an \( m \times s \) matrix and as a mapping \( \mathbb{V}^s \to \mathbb{V}^m \) defined in the obvious way, i.e. \( Ac = (\sum_{j \in [s]} A_{ij}c_j)_{i \in [m]} \in \mathbb{V}^m \) for \( c = (c_j)_{j \in [s]} \in \mathbb{V}^s \). In the case of scalar-valued function approximation (i.e. \( V = \mathbb{V} = \mathbb{C} \)), the problem (14) is a standard algebraic least-squares problem.
Moreover, in the general Hilbert-valued case, it is a straightforward exercise to show that a solution of (14) is given by

\[ \hat{c} = A^\dagger v. \]  

(16)

Here \( A^\dagger \) is the pseudoinverse of \( A \), or more precisely, its extension in the above manner to a mapping \( \mathbb{V}^m \to \mathbb{V}^s \). In particular, if \( A \) is full rank, then this is the unique solution of (14). Now recall the basis \( \{ \psi_i \}_{i=1}^k \) for \( \mathbb{V}_h \). Observe that we can write the \( r \)th component \( v_i \) of the \( \mathbb{V}_h \)-valued vector \( v \) defined above as

\[ v_i = \sum_{j \in [k]} V_{ij} \psi_j, \quad i \in [m], \]

for scalar coefficients \( V_{ij} \in \mathbb{C} \). Likewise, we can also write

\[ \hat{c} = (\hat{c}_i)_{i \in [s]}, \quad \hat{c}_i = \sum_{j \in [k]} \hat{C}_{ij}, \quad i \in [s], \]

for scalar coefficients \( \hat{C}_{ij} \in \mathbb{C} \), so that \( \hat{f} \) can be expressed as

\[ \hat{f} = \sum_{i \in \mathcal{S}} \sum_{j \in [k]} \hat{C}_{ij} \theta_i \otimes \psi_j. \]

Letting \( \hat{C} = (C_{ij})_{i \in [s], j \in [k]} \) and \( V = (V_{ij})_{i \in [m], j \in [k]} \) and using (16), we see that

\[ \hat{C} = A^\dagger V. \]

Hence, the coefficients \( \hat{C} \) can be computed by first computing the pseudoinverse \( A^\dagger \) and then performing the above matrix-matrix multiplication, for a total of \( O(ms^2 + msk) \) floating point operations. Alternatively, one could solve \( k \) standard algebraic least-squares problems for the columns of \( V \). If conjugate gradients are used, for example, the cost of obtaining a residual error of size \( \eta \) is \( O(\text{cond}(A)msk \log(\eta^{-1})) \), where \( \text{cond}(A) \) is the condition number of \( A \). This may be more efficient in the case where \( k \ll s \); in particular, the scalar-valued case, where \( k = 1 \).

### 3.2 Accuracy, stability and sample complexity

In this and the next several subsections, we investigate Questions 1–3. We commence with Question 1. Accuracy and stability of the approximation (13) is governed by the existence of a norm equivalence over \( \mathbb{P}_3 \). Specifically, we assume that

\[ \alpha \| p \|_{L_2^2(D)}^2 \leq \frac{1}{m} \sum_{i=1}^m w(y_i) |p(y_i)|^2 \leq \beta \| p \|_{L_2^2(D)}^2, \quad \forall p \in \mathbb{P}_3, \]  

(17)
for constants $0 \leq \alpha \leq \beta < \infty$. In other words, the functional $p \mapsto \sqrt{\frac{1}{m} \sum_{i=1}^{m} w(y_i)|p(y_i)|^2}$ is an equivalent norm over $P_S$ to the $L^2_P$-norm. We remark also that (17) is a condition for the space $P_S = P_S;\mathbb{C}$ consisting of scalar-valued functions. As the next theorem shows, however, such a condition also determines accuracy and stability for the approximation of Hilbert-valued functions in the space $P_S;\mathbb{V}_h$. With this in hand, we now define the discrete inner product

$$\langle g, h \rangle_{\text{disc}} = \frac{1}{m} \sum_{i=1}^{m} w(y_i)\langle g(y_i), h(y_i) \rangle_{\mathbb{V}}, \quad g, h \in L^2_P(D;\mathbb{V}),$$

and corresponding discrete semi-norm $\|g\|_{\text{disc}} = \sqrt{\langle g, g \rangle_{\text{disc}}}, g \in L^2_P(D;\mathbb{V})$.

**Theorem 2 (Accuracy and stability of weighted least squares).** Let $f \in L^2_P(D)$, $0 < \alpha \leq \beta < \infty$, $\{y_i\}_{i=1}^{m} \subseteq D$, $w : D \rightarrow \mathbb{R}$ be such that $w(y_i)$ is well defined for all $i$, and suppose that (17) holds. Then the approximation $\hat{f}$ in (13) is unique, and satisfies

$$\|f - \hat{f}\|_{L^2_P(D;\mathbb{V})} \leq \inf_{p \in P_S;\mathbb{V}} \left\{ \|f - p\|_{L^2_P(D;\mathbb{V})} + \left(1 + \frac{1}{\sqrt{\alpha}}\right)\|f - p\|_{\text{disc}} \right\}$$

$$+ \|f - \mathcal{P}_h(f)\|_{L^2_P(D;\mathbb{V})} + \frac{1}{\sqrt{\alpha}}\|e\|_{L^2(m;\mathbb{V})}$$

where $e = \frac{1}{\sqrt{m}} \left(\sqrt{w(y_i)n_i}\right)_{i=1}^{m} \in \mathbb{V}^m$.

This result (see §3.7 for its proof) asserts that the error for the learned approximation $\hat{f}$ splits into three quantities. First, a best approximation error term in the subspace $P_S;\mathbb{V}$. Second, a space discretization error, which accounts for the fact that the least-squares problem is formulated over $\mathbb{V}_h$ as opposed to $\mathbb{V}$, and is equal to the projection error $f - \mathcal{P}_h(f)$. And third, a term depending on the measurement noise values $n_i$. Note that this theorem does not require the points $\{y_i\}_{i=1}^{m}$ to be random. It holds for any fixed set of sample points whenever (17) also holds.

**Remark 1.** This result has several disadvantages. First, the noise terms $n_i$ are multiplied by the weight factors $\sqrt{w(y_i)}$, meaning that noise terms corresponding to large values of $w$ are weighted more heavily. Second, the best approximation error mixes the $L^2_P(D;\mathbb{V})$-norm (which is the norm in which the error $f - \hat{f}$ is measured) with the discrete norm $\|f - p\|_{\text{disc}}$. When the sample points $y_i$ are random variables (as they will be below), one can use this fact to slightly modify the approximation $\hat{f}$ in a way in which error bounds involving only $\|f - p\|_{L^2_P(D;\mathbb{V})}$ can be obtained. We omit the details. See [28, 32, 33] for further information in the scalar-valued case.

**Remark 2.** The reader will notice that Theorem 2 does not involve the upper constant $\beta$ in (17). While not strictly needed for this theorem, this constant plays a role in the computation of the least-squares approximation. Indeed, it is straightforward to show that the condition number $\text{cond}(A)$ is bounded by $\sqrt{\beta/\alpha}$ whenever $\{\phi_i :
$\iota \in S$ forms an orthonormal basis for $P_S$. Hence, when the ratio $\beta/\alpha$ is small, the least-squares system can be solved more efficiently (when employing conjugate gradients) and its output is less affected by floating point errors.

This property is relevant to Examples 1 and 2 since they involve orthonormal bases. On the other hand, the least-squares matrix $A$ will be poorly conditioned whenever the system $\{\phi_\iota : \iota \in S\}$ is near-linear dependent. This occurs notably in Example 3 [12]. Perhaps counter-intuitively, this does not necessarily lead to substantial errors in the resulting least-squares approximation. In fact, whenever the infinite system of functions forms a frame (as it does Example 3), this property endows the problem with sufficient structure to ensure accurate and stable (regularized) least-squares approximations. See [12] for further discussion.

We now progress to the matter of sample complexity, which will lead to answers to Questions 2 and 3. As shown in [32] (see also [10]) Sample complexity of the least-squares scheme is determined by the existence of a so-called weighted Nikolskii-type inequality over $P_S$. Specifically, let $N(P_S, w)$ be the smallest constant such that

$$\|p\|_{L^\infty(D)} \leq N(P_S, w)\|p\|_{L^2(D)}, \quad \forall p \in P_S. \quad (18)$$

Again, we observe that this inequality is formulated for the space $P_S = P_S; C$ of scalar-valued functions. We remark also that $N(P_S, w)$ is related to the Christoffel function $K(P_S)$ of the subspace $P_S$. Specifically,

$$N(P_S, w) = \left\| \sqrt{w(\cdot)K(P_S)(\cdot)} \right\|_{L^\infty(D)}, \quad (19)$$

where $K(P_S)$ is the reciprocal of the Christoffel function of $P_S$. Let $\{v_\iota\}_{\iota \in S} \subset L^2(D)$ be any orthonormal basis for $P_S$. Then this function has the explicit expression

$$K(P_S)(y) = \sum_{\iota \in S} |v_\iota(y)|^2. \quad (20)$$

**Theorem 3 (Sample complexity of weighted least squares).** Let $0 < \varepsilon < 1$, $0 < \delta < 1$, $\mu_1, \ldots, \mu_m$ be probability measures on $D$ satisfying Assumption 7 and $y_1, \ldots, y_m$ be independent with $y_i \sim \mu_i$ for $i = 1, \ldots, m$. Suppose that

$$m \geq c_\delta \cdot (N(P_S, w))^2 \cdot \log(2s/\varepsilon), \quad c_\delta = \left( (1 - \delta) \log(1 - \delta) + \delta \right)^{-1}, \quad (21)$$

where $w$ is the weight function specified in [6]. Then (17) holds with $1 - \delta \leq \alpha \leq \beta \leq 1 + \delta$, with probability at least $1 - \varepsilon$.

This theorem (see §3.7 for the proof) states that the sample complexity is dominated by the behaviour of the weighted Nikolskii constant $N(P_S, w)$. Observe that

$$(N(P_S, w))^2 \geq s, \quad (22)$$

for any choice of $w$. Indeed, $(N(P_S, w))^2 \geq w(y)K(P_S)(y)$ for almost every $y$, and therefore
Towards optimal sampling for learning sparse approximations in high dimensions

\[ (\mathcal{N}(P_S, w))^2 \int_D \frac{1}{w(y)} \, d\rho(y) \geq \int_D K(P_S)(y) \, d\rho(y). \]

The left-hand side is equal to \( (\mathcal{N}(P_S, w))^2 \) due to (7), and the right hand side is equal to \( s \), due to the relation (20) and the fact that the \( \psi_i \)'s are orthonormal.

### 3.3 Monte Carlo sampling

We are now ready to discuss the first part of Question 3 in the context of the examples introduced in §2.3. Recall that Monte Carlo sampling corresponds to setting \( \mu_1 = \ldots = \mu_m = \rho \).

In this case, it follows from (6) that the function \( w(y) \equiv 1 \). Hence, \( \hat{f} \) is a standard unweighted least-squares approximation. As shown by Theorem 3, the sample complexity

\[ m \geq c_s \cdot (\mathcal{N}(P_S))^2 \cdot \log(2s/\varepsilon), \quad (23) \]

is governed by the unweighted Nikolskii constant

\[ \mathcal{N}(P_S) = \| \sqrt{K(P_S)}(\cdot) \|_{L_\infty(D)}. \quad (24) \]

We are interested in the behaviour of \( (\mathcal{N}(P_S))^2 \) in relation to \( s = |S| \). Clearly, there are instances where \( (\mathcal{N}(P_S))^2 \) attains the optimal value \( (\mathcal{N}(P_S))^2 = s \) (recall (22)). Indeed, the functions \( \phi_i \) of Example 1 are orthonormal and equal to one in absolute value. Hence \( K(P_S) \equiv s \) by (20), and (23) yields the sample estimate \( m \gtrsim s \cdot \log(2s/\varepsilon) \), which is optimal up to constants and log factors.

Unfortunately, this desirable property does not hold in general. As the next result attests, the constant \( \mathcal{N}(P_S) \) can generally be arbitrarily large in comparison to \( s \):

**Lemma 1.** There exists a probability space \((D, \mathcal{F}, \rho)\) such that following holds. For every \( s \in \mathbb{N} \) and \( C > 0 \) there exists a subspace \( P \subset L_2^2(D) \) of dimension \( s \) such that \( \mathcal{N}(P) \geq C \).

**Proof.** We consider Example 2 in the case \( d = 1 \). The classical Legendre polynomial \( P_1 \) attains its maximum value at \( y = 1 \) and takes value \( P_1(1) = 1 \). Hence,

\[ \| P_1 \|_{L_2([-1,1])} = P_1(1) = \sqrt{2t + 1}. \quad (25) \]

It follows that for any subspace \( P = P_S \), we have

\[ (\mathcal{N}(P_S))^2 = \| K(P_S)(\cdot) \|_{L_\infty(D)} = K(P_S)(1) = \sum_{i \in S} (2i + 1). \]

Since \( S \subset \mathbb{N}_0^d \), \( |S| = s \) can be arbitrary, we now choose it so that the right-hand side exceeds \( C \). \( \square \)
This lemma and its proof suggest that Monte Carlo sampling may be highly sub-optimal in the setting of Example 2 (and therefore Example 3 as well) if the indices in the target set $S$ are allowed to become arbitrarily large. One way to mitigate this is to impose additional structure on $S$. A common structure is that of lower sets:

**Definition 1.** A multi-index set $I \subseteq \mathbb{N}_0^d$ is lower if, whenever $\iota \in I$ and $\kappa \leq \iota$ (this inequality is understood componentwise), then $\kappa \in I$.

Note that many common index sets used in polynomial approximation are lower. For example, the sets (10)–(12) are all lower. In general, lower sets are known to be good candidates for the support sets of polynomial coefficients of smooth functions in high dimensions [2, 8, 9, 22, 25, 25, 33]. In particular, this is true for solutions to parametric PDEs, where the lower set sparsity has been studied and variously exploited to construct effective polynomial approximations [9, 21–24, 29, 33]. Motivated by Example 1, we observe that it is also straightforward to define lower subsets of $\mathbb{Z}^d$. In this case, we replace the inequality by $|\kappa| \leq |\iota|$, where, for a multi-index $\iota = (\iota_k)_{k=1}^d \in \mathbb{Z}^d$, $|\iota| = (|\iota_k|)_{k=1}^d$ is the multi-index of its absolute values.

In the case of Example 2, it is known that when $S \subset \mathbb{N}_0^d$, $|S| \leq s$, is a lower set, one has

$$(N(P_S))^2 \leq s^2.$$  

See [22,23]. Furthermore, this bound is sharp, in the sense that there exists a lower set $S$ of size $s$ — specifically, the set $S = \{(t,0,\ldots,0) : s = 0,\ldots,s-1\}$ — for which $(N(P_S))^2 = s^2$. Hence, imposing a lower set structure reduces the sample complexity for Monte Carlo sampling to at worst quadratic in $s$, up to log factors.

**Remark 3.** In view of Example 3, we remark in passing that this quadratic bound also holds for arbitrary lower sets and large classes of irregular domains [12], up to a domain-dependent constant. Moreover, this also holds for any Lipschitz domain in the case where $S = \mathcal{S}^{\text{TD}}_n$ is the total degree index set [11, 38]. On the other hand, for domains with $C^2$ boundary and $S = \mathcal{S}^{\text{TD}}_n$, one has a better scaling in higher dimensions; namely, $(N(P_S))^2 \leq c_D s^{d+1/d}$, where $c_D > 0$ is a constant depending on the domain $D$ only [38].

### 3.4 Optimal sampling

We now answer the second part of Question 3 in the affirmative. Our aim is to choose the weight function $w$ to minimize $N(P_S, w)$, and then choose the measures satisfying Assumption 1. To do this, we appeal to (19) and, keeping in mind the normalization (7), set

$$w(y) = \left(\frac{1}{s} K(P_S)(y)\right)^{-1} = \left(\frac{1}{s} \sum_{\iota \in S} |\upsilon_\iota(y)|^2\right)^{-1}.$$  

Notice that this yields, via (19), the optimal Nikolskii constant
Towards optimal sampling for learning sparse approximations in high dimensions

\[ N(P_S, w) = \sqrt{s}. \]

In particular, the sample complexity estimate (21) becomes

\[ m \geq c_8 \cdot s \cdot \log\left(\frac{2s}{\varepsilon}\right), \]

which is optimal up to the log factor.

Having chosen \( w \), we now choose the measures \( \mu_i \) so that (6) holds. We consider two possibilities. The first we term nonhierarchical, and is given simply by

\[ \mu_1 = \ldots = \mu_m = \mu, \quad d\mu(y) = \frac{1}{w(y)} d\rho(y) = \frac{1}{s} \sum_{i \in S} |\nu_i(y)|^2 d\rho(y). \tag{26} \]

Clearly, (6) holds in this case. The second scheme is hierarchical. In this scheme, we suppose that \( m = ks \) for some \( k \in \mathbb{N} \). Then we define

\[ \mu_i = |\nu_{ij}(y)|^2 d\rho(y), \quad (j-1)k < i \leq jk, \quad j = 1, \ldots, s, \tag{27} \]

where \( \{t_1, \ldots, t_s\} \) is an enumeration of the indices in \( S \). Notice that

\[ \frac{1}{m} \sum_{i=1}^{m} d\mu_i(y) = \frac{k}{m} \sum_{j=1}^{s} |\nu_{ij}(y)|^2 d\rho(y) = \frac{1}{s} \sum_{s \in S} |\nu_s(y)|^2 d\rho(y). \]

Therefore (6) also holds in this case.

The nonhierarchical scheme (26) was introduced in [32] and is suitable for learning an approximation in a fixed subspace \( S \). However, as discussed in [16, 67], it is not well suited to the problem where one seeks to learn a sequence of approximations \( \hat{f}_1, \hat{f}_2, \ldots \) in a hierarchy of nested subspaces \( S_1 \subseteq S_2 \subseteq \cdots \). The issue is that as the subspace \( S = S_i \) changes, the measure \( \mu \) defined in (26) changes, hence the existing samples are effectively drawn from the wrong distribution for the purposes of constructing an approximation in the new subspace \( S_{i+1} \). The hierarchical scheme (27), introduced in [67], overcomes this problem; see also [16] for a different approach. We refer to [10, 67] for further information.

### 3.5 Practical optimal sampling via discrete measures

Unfortunately, generating samples from either the measure (26) or the measures (27) may not be straightforward, since it requires an orthonormal basis \( \{\nu_k\}_{k \in S} \) of \( P_S \). This may not be available in practice, and even if it is, drawing samples from the resulting measures may be computationally challenging. See [10, 16, 32, 68] for further information on this issue, as well as [49, 71] for the specific case of tensor-product polynomial approximation.

A remedy to this situation was proposed in [10, 68]. The idea is to replace \( \rho \) (which is typically a continuous measure) by a discrete measure, supported on a
finite grid, so that both constructing an orthonormal basis and sampling from the corresponding measures are automatically straightforward. Let \( Z = \{ z_i \}_{i=1}^k \subset D \) be a finite grid. We consider the discrete uniform measure given by
\[
\tau = \frac{1}{k} \sum_{i=1}^k \delta_{z_i},
\] (28)
The idea is now to replace \( \rho \) by \( \tau \) throughout. Consider the nonhierarchical scheme for simplicity. Then, doing so, we deduce that if
\[
m \gtrsim s \cdot \log\left(\frac{2s}{\epsilon}\right),
\] (29)
then the error bound
\[
\| f - \hat{f} \|_{L^2_\tau(D;V)} \lesssim \inf_{p \in P_{S,V}} \left\{ \| f - p \|_{L^2_\tau(D;V)} + \| f - p \|_{\text{disc}} \right\} + \| f - \mathcal{P}_h(f) \|_{L^2_\tau(D;V)} + \| e \|_{\ell^2([m];V)},
\] (30)
holds probability at least \( 1 - \epsilon \), where
\[
w(y) = \left( \frac{1}{s} \sum_{i \in S} | u_i(y) |^2 \right)^{-1}, \quad y \in \text{supp}(\tau) = Z,
\]
\[
d\mu(y) = \frac{1}{s} \sum_{i \in S} | u_i(y) |^2 \, d\tau(y), \quad y \in \text{supp}(\tau),
\] (31)
and \( \{ u_i \}_{i \in S} \subset L^2_\tau(D) \) is an orthonormal basis for \( P_S \) with respect to \( \tau \).

Since \( \tau \) is a discrete measure, this orthonormal basis can be constructed via straightforward linear algebra. Indeed, define the matrix
\[
B = \left( \phi_{ij}(z_i)/\sqrt{k} \right)_{i \in [k], j \in [s]} \in \mathbb{C}^{k \times s},
\]
and suppose that it has the QR-factorization \( B = QR \), where \( Q \in \mathbb{C}^{k \times s} \) has orthonormal columns and \( R \in \mathbb{C}^{s \times s} \) is upper triangular. Then the orthonormal basis \( \{ u_i \}_{i \in S} \) is given by
\[
u_i(y) = \sum_{j=1}^s (R^{-\top})_{ij} \phi_{ij}(y), \quad i \in [s].
\]
In particular, its values on the grid \( Z \) are precisely
\[
u_i(z_i) = \sqrt{k} Q_{ij}, \quad i \in [k], j \in [s].
\]
Substituting this into (31) and recalling the definition of \( \tau \), we see that the discrete measure \( \mu \) is given by
Towards optimal sampling for learning sparse approximations in high dimensions

\[ d\mu(y) = \sum_{i \in [k]} \left( \frac{1}{s} \sum_{j \in [s]} |Q_{ij}|^2 \right) d\delta_i(y). \]

Hence, sampling from \( \mu \) is now trivial. Indeed, \( y \sim \mu \) if

\[ P(y = z_i) = \frac{1}{s} \sum_{j \in [s]} |Q_{ij}|^2, \quad i \in [k]. \]

The reader will have no doubt noticed that the error bound (30) is with respect to the discrete measure \( \tau \). It is often preferable to also have an error bound over the original measure \( \rho \). Such an error bound is guaranteed whenever the \( L_2^\tau \)-norm is equivalent to the \( L_2^\rho \)-norm over \( P_S \), i.e.

\[ \alpha \|p\|_{L_2^\tau(D)}^2 \leq \|p\|_{L_2^\rho(D)}^2 \leq \beta \|p\|_{L_2^\tau(D)}^2, \quad \forall p \in P_S. \]  

Indeed, recall that the sampling condition (29) and the choices of \( \mu \) and \( w \) imply a norm equivalence between the \( L_2^\tau \)-norm and the discrete norm over the sample points, i.e.

\[ (1 - \delta) \|p\|_{L_2^\tau(D)}^2 \leq \frac{1}{m} \sum_{i=1}^m w(y_i) |p(y_i)|^2 \leq (1 + \delta) \|p\|_{L_2^\tau(D)}^2, \quad \forall p \in P_S. \]

Hence, we deduce that

\[ (1 - \delta) \alpha \|p\|_{L_2^\tau(D)}^2 \leq \frac{1}{m} \sum_{i=1}^m w(y_i) |p(y_i)|^2 \leq (1 + \delta) \beta \|p\|_{L_2^\tau(D)}^2, \quad \forall p \in P_S. \]

Therefore the norm equivalence (17) with respect to the original \( L_2^\rho \)-norm also holds, meaning that an error bound in this norm follows immediately from Theorem 2 (with constant \( \alpha = (1 - \delta) \alpha' \)).

Remark 4. A simple means to ensure (32) is to construct \( Z = \{z_i\}_{i=1}^k \) as a random Monte Carlo grid (independently of the sample points \( y_i \)). That is, we let the \( z_i \) be independent random variables drawn according to the measure \( \rho \). Observe that (32) is precisely

\[ \alpha' \|p\|_{L_2^\tau(D)}^2 \leq \frac{1}{k} \sum_{i=1}^k |p(z_i)|^2 \leq \beta' \|p\|_{L_2^\tau(D)}^2, \quad \forall p \in P_S. \]

This is nothing more than the special case of (17) for the grid \( Z \) (recall that \( w \equiv 1 \) for Monte Carlo sampling). Hence, (32) is ensured by Theorem 3. In particular, it holds with \( 1 - \delta \leq \alpha' \leq \beta' \leq 1 + \delta \), provided

\[ k \geq c_{\delta} \cdot \left( \mathcal{N}(P_S) \right)^2 \cdot \log(2s/\epsilon), \]
where, as in §3.3, \( \mathcal{N}(P_S) = \|K(P_S)(\cdot)\|_{L^\infty(D)} \) is the unweighted Nikolskii constant. Of course, this constant may be very large depending on the choice of \( P_S \); recall the discussion §3.3. Yet, this grid is only used to define the optimal sampling measure \( \mu \). Therefore, the number of grid points \( k \) only affects the computational cost for generating the sample points. It does not affect the sample complexity of the weighted least-squares approximation, which is \( m \gtrsim s \log(2s/\epsilon) \) for the optimal measure.

Having said this, a practical problem is that estimates for \( \mathcal{N}(P_S) \) may not be available, or if they are, they may not be particularly tight, thus leading to overly large grids. In [38] an empirical strategy is described to mitigate this issue, based on independently drawing an auxiliary grid that is used to test the quality of the grid \( Z \).

### 3.6 Numerical examples

We conclude this discussion on least-squares approximation with several numerical examples. In these and other examples considered later in this chapter, we consider the scalar-valued functions

\[
\begin{align*}
  f_1(y) &= \exp \left( -\frac{1}{d} \sum_{k=1}^{d} y_k \right), \\
  f_2(y) &= \prod_{\substack{1 \leq k \leq \lfloor d/2 \rfloor + 1 \leq \lfloor d/2 \rfloor \leq y_k/4^k}} \cos(16y_k/2^k) \\
  &\quad \times \prod_{\substack{1 \leq k \leq \lfloor d/2 \rfloor \leq 1 - y_k/4^k}} (1 - y_k/4^k), \\
  f_3(y) &= \prod_{i=1}^{d} \sqrt{4 + (y_i + (-1)^{i+1}/(i+1))^2}, \\
  f_4(y) &= \frac{1}{\sum_{i=1}^{d} \sqrt{|y_i|}}.
\end{align*}
\]

Since our goal is to compare Monte Carlo sampling with the optimal sampling procedures described above, we focus on Examples 2 and 3 (recall that Monte Carlo sampling is optimal, up to the log term, in the case of Example 1). To this end, we consider the domains

\[
\begin{align*}
  D_1 &= [-1, 1]^d, \\
  D_2 &= \{ y \in \mathbb{R}^d : 1/4 \leq y_1^2 + \ldots + y_d^2 \leq 1 \}, \\
  D_3 &= \{ y \in [-1, 1]^d : y_1 + \ldots + y_d \leq 1 \}.
\end{align*}
\]

We follow the approach of §3.5 and, in particular, Remark 4, to generate a Monte Carlo grid \( Z \) and corresponding discrete measure \( \tau \) as in (28). Here, \( k = 30s_{\text{max}} \), where \( s_{\text{max}} \) is the maximum size of \( s \) used in the given experiment. For the error, we compute the relative \( L^\infty_\tau \)-norm error, i.e.
Towards optimal sampling for learning sparse approximations in high dimensions

\[ \| f - \hat{f} \|_{L_\infty(D)} \leq \beta \| f \|_{L_\infty(D)} \]  

(35)

We perform a total of \( T = 50 \) trials. In the Monte Carlo and optimal nonhierarchical schemes, each trial corresponds to a single draw of the sample points \( y_1, \ldots, y_m \) at each value of \( m \) considered. For the optimal hierarchical scheme, a single trial is a full set of points \( y_1, \ldots, y_{m_{\text{max}}} \), where \( m_{\text{max}} \) is the maximum value of \( m \) considered. In all cases, we report the log-average of the error (35) over these trials, with the shaded regions corresponding to one log-standard deviation (see [9, App. A] for further information).

In Fig. 1, we compare Monte Carlo with both the hierarchical and nonhierarchical optimal sampling schemes. In two dimensions, typical sample points generated by these schemes for different domains are shown in Fig. 2. As we see from Fig. 1, Monte Carlo sampling leads to worse performance compared to both optimal sampling schemes, especially in lower-dimensional problems. It is notable that Monte Carlo sampling also leads to an increasing approximation error in several cases, since the number of samples is chosen to scale log-linearly with \( s \), rather than log-quadratically (recall the discussion in §3.3). This is corroborated in Fig. 3, where we plot the constant \( \alpha \) for the different sampling schemes. On the other hand, we observe that the relative performance of Monte Carlo sampling improves in higher dimensions, where it offers similar approximation errors to the optimal schemes. We see this effect consistently throughout this work. Finally, we remark in passing that there is virtually no difference between the nonhierarchical and hierarchical versions of the optimal sampling scheme.

3.7 Proofs of Theorems 2 and 3

The proofs of Theorems 2 and 3 follow ideas that are now well established in the literature (see, for example, [10, 12, 28, 32, 68]). They are included for completeness. We commence with Theorem 2. We now observe the following:

**Lemma 2.** Suppose that (17) holds. Then

\[ \alpha \| p \|^2_{L_p^2(D;V)} \leq \| p \|_{\text{disc}}^2 \leq \beta \| p \|^2_{L_p^2(D;V)} \quad \forall p \in P_{S;V_h}, \]  

(36)

where \( \| \cdot \|_{\text{disc}} \) is as in §3.2.

Simply put, this lemma states that if there is a norm equivalence in the scalar-valued case over \( P_S = P_{S;C} \) between the continuous \( L_p^2 \)-norm and the discrete norm defined by the sample points, then there is also the same norm equivalence in the Hilbert-valued case over \( P_{S;V} \).

**Proof.** First, let \( \{ \psi_i \}_{i=1}^K \) be an orthonormal basis of \( V_h \subseteq V \). Let \( p \in P_{S;V} \), and observe that it has the unique expression
We compare Monte Carlo sampling (left), discrete optimal nonhierarchical sampling (middle) and discrete optimal hierarchical sampling (right) for \( f = f_1, D = D_2 \) (top row), \( f = f_2, D = D_3 \) (middle row), and \( f = f_3, D = D_2 \) (bottom row). In all experiments, \( S = \mathcal{H}_t \) is the hyperbolic cross index set \( \mathcal{H}_t \) and, for each value of \( t \), \( m \) is chosen as the smallest integer such that
\[
m \geq s \log(s),
\]
where \( s = |\mathcal{H}_t| \). Let \( p = \sum_{i \in [s]} \sum_{j \in [K]} c_{ij} \phi_i \otimes \psi_j, \quad c_{ij} \in \mathbb{C}, \ i \in [s], \ j \in [K], \)
where \( \{t_1, \ldots, t_s\} \) is an enumeration of the indices in \( S \). Let \( q_j = \sum_{i \in [s]} c_{ij} \phi_i \) and observe that \( q_j \in \mathbb{P}_S \). Notice also that
\[
\|p(y)\|_V^2 = \sum_{j \in [K]} \left| \sum_{i \in [s]} c_{ij} \phi_i(y) \right|^2 = \sum_{j \in [K]} |q_j(y)|^2, \quad \forall y \in D.
\]
In particular, this implies that
\[
\|p\|_{L^2(D;\mathbb{V})} = \int_D \|p(y)\|_V^2 \, d\rho(y) = \int_D \sum_{j \in [K]} |q_j(y)|^2 \, d\rho(y) = \sum_{j \in [K]} \|q_j\|_{L^2(D)}^2. \quad (37)
\]
Fig. 2: The domains $D_1$, $D_2$ and $D_3$ (left to right) in $d = 2$ dimensions. The grey dots are the finite Monte Carlo grid $Z = \{z_i\}_{i=1}^k$ with $k = 20,000$. The top row shows Monte Carlo sampling with $m = 1,824$ points. The second row shows the same number of points generated from the discrete optimal nonhierarchical sampling measure based on $\mathcal{F} = \mathcal{F}_{t-1}$, where $t = 68$ and $s = 308$.

Fig. 3: The constant $1/\sqrt{\alpha}$ versus $s$ in $d = 2$ dimensions for Monte Carlo sampling (left), discrete optimal nonhierarchical sampling (middle) and discrete optimal hierarchical sampling (right) and for $D_1$ (top row) and $D_2$ (bottom row), where $\mathcal{F} = \mathcal{F}_{t-1}$.

and also that

$$\|p\|_{\text{disc}}^2 = \frac{1}{m} \sum_{i \in [m]} w(y_i) \|p(y_i)\|_V^2 = \sum_{j \in [K]} \left( \frac{1}{m} \sum_{i \in [m]} w(y_i) |q_j(y)|^2 \right).$$
Hence, by the scalar-valued norm equivalence (17), we deduce that
\[
\alpha \sum_{j \in [K]} \|q_j\|_{L^2(D)}^2 \leq \|p\|_{L^2(D)}^2 \leq \beta \sum_{j \in [K]} \|q_j\|_{L^2(D)}^2.
\]

The result now follows immediately from (37). \(\square\)

Proof of Theorem 2

Since \(\hat{f}\) is a solution of the least-squares problem, it is also a solution of the variational equations

find \(\hat{f} \in P_{S;V_h}\) such that \(\langle \hat{f}, q \rangle_{\text{disc}} = \langle f, q \rangle_{\text{disc}} + \frac{1}{m} \sum_{i \in [m]} w(y_i) \langle n_i, q(y_i) \rangle_{V}, \forall q \in P_{S;V_h}.\)

Uniqueness of \(\hat{f}\) now follows immediately from Lemma 2 since \(\langle \cdot, \cdot \rangle_{\text{disc}}\) forms an inner product on \(P_{S;V}\), and therefore \(P_{S;V_h}\).

We now derive the desired error bound. First, observe that since \(q \in P_{S;V_h}\), these equations are equivalent to

find \(\hat{f} \in P_{S;V_h}\) such that \(\langle \hat{f}, q \rangle_{\text{disc}} = \langle \mathcal{P}_h(f), q \rangle_{\text{disc}} + \frac{1}{m} \sum_{i \in [m]} w(y_i) \langle n_i, q(y_i) \rangle_{V}, \forall q \in P_{S;V_h}.\)

Now let \(q \in P_{S;V_h}\) be arbitrary. Then these equations give

\[
\|\hat{f} - q\|_{\text{disc}}^2 = \langle \mathcal{P}_h(f) - q, \hat{f} - q \rangle_{\text{disc}} + \frac{1}{m} \sum_{i \in [m]} w(y_i) \langle n_i, \hat{f}(y_i) - q(y_i) \rangle_{V},
\]

and applying the Cauchy–Schwarz inequality several times to the right-hand side we deduce that

\[
\|\hat{f} - q\|_{\text{disc}} \leq \|\mathcal{P}_h(f) - q\|_{\text{disc}} + \|e\|_{L^2([m];V)}.
\]

Here, we also recall the definition of \(e\). Further, using (36) and the fact that \(q \in P_{S;V_h}\) we get

\[
\|\hat{f} - q\|_{L^2(D;V)} \leq \frac{1}{\sqrt{\alpha}} \left( \|\mathcal{P}_h(f) - q\|_{\text{disc}} + \|e\|_{L^2([m];V)} \right).
\]

Now let \(p \in P_{S;V}\) be arbitrary and write \(q = \mathcal{P}_h(p)\). Therefore,

\[
\|\hat{f} - f\|_{L^2(D;V)} \leq \|\hat{f} - \mathcal{P}_h(p)\|_{L^2(D;V)} + \|\mathcal{P}_h(p) - \mathcal{P}_h(f)\|_{L^2(D;V)} + \|\mathcal{P}_h(f) - f\|_{L^2(D;V)}
\]
\[
\leq \frac{1}{\sqrt{\alpha}} \left( \|\mathcal{P}_h(f) - \mathcal{P}_h(p)\|_{\text{disc}} + \|e\|_{L^2([m];V)} \right)
\]
\[
+ \|\mathcal{P}_h(f - p)\|_{L^2(D;V)} + \|\mathcal{P}_h(f) - f\|_{L^2(D;V)}
\]
\[
\leq \frac{1}{\sqrt{\alpha}} \left( \|f - p\|_{\text{disc}} + \|e\|_{L^2([m];V)} \right)
\]
\[
+ \|f - p\|_{L^2(D;V)} + \|\mathcal{P}_h(f) - f\|_{L^2(D;V)}.
\]
By construction, these matrices are independent and positive semidefinite. Also, it therefore suffices to show that
\[ \lambda_{\min}(X_i) \geq 1 - \delta \]
which gives
\[ \lambda_{\max}(X_i) \leq 1 + \delta \]  
This completes the proof.

We now consider Theorem 4. This is most commonly established using the matrix Chernoff bound [93, Thm. 1.1], which we restate here for convenience:

**Theorem 4 (Matrix Chernoff bound).** Let \( X_1, \ldots, X_m \) be independent, self-adjoint random matrices of size \( s \times s \). Assume that \( X_i \) is positive semidefinite and \( \lambda_{\max}(X_i) \leq R \) almost surely for each \( i = 1, \ldots, m \), and define
\[ \mu_{\min} = \lambda_{\min} \left( \sum_{i=1}^{m} E X_i \right), \quad \mu_{\max} = \lambda_{\max} \left( \sum_{i=1}^{m} E X_i \right). \]

Then, for \( 0 \leq \delta \leq 1 \),
\[ \mathbb{P} \left( \lambda_{\min} \left( \sum_{i=1}^{m} X_i \right) \leq (1 - \delta) \mu_{\min} \right) \leq s \cdot \exp \left( - \frac{\mu_{\min}((1 - \delta) \log(1 - \delta) + \delta)}{R} \right), \]
and, for \( \delta \geq 0 \),
\[ \mathbb{P} \left( \lambda_{\max} \left( \sum_{i=1}^{m} X_i \right) \geq (1 + \delta) \mu_{\max} \right) \leq s \cdot \exp \left( - \frac{\mu_{\max}((1 + \delta) \log(1 + \delta) - \delta)}{R} \right). \]

**Proof of Theorem 4.** Let \( \{ u_i \}_{i \in [s]} \) be an orthonormal basis of \( P_s \subset L^2_\rho(D) \) with respect to \( \rho \). Let \( p \in P_s \), \( p \neq 0 \), be arbitrary, and write \( p = \sum_{i \in [s]} c_i u_i \) and \( c = (c_i)_{i \in [s]} \).
Then
\[ \| p \|_{L^2_\rho(D)}^2 = \int_D \left| \sum_{i \in [s]} c_i u_i(y) \right|^2 d\rho(y) = \sum_{i \in [s]} |c_i|^2 = \| c \|_2^2, \]
and
\[ \frac{1}{m} \sum_{i=1}^{m} w(y_i) |p(y_i)|^2 = c^* G c, \]
where \( G = A^* A \in \mathbb{C}^{s \times s} \) is the self adjoint matrix with entries \( G_{j,k} = \frac{1}{m} \sum_{i=1}^{m} w(y_i) \langle u_j, u_k \rangle_{L^2_\rho(D)} \).
It therefore suffices to show that \( \lambda_{\min}(G) \geq 1 - \delta \) and \( \lambda_{\max}(G) \leq 1 + \delta \).
Write
\[ G = \sum_{i=1}^{m} X_i, \quad X_i = \left\{ \frac{1}{m} w(y_i) u_j(y_i) \overline{u_k(y_i)} \right\}_{j,k=1}^{m}. \]
By construction, these matrices are independent and positive semidefinite. Also,
\[ (\mathbb{E}(X_i))_{j,k} = \int_D v_j(y) \overline{v_k(y)} w(y) \frac{1}{m} d\mu_i(y), \]
which gives
\[
\left( \sum_{i=1}^{m} E(X_i) \right)_{j,k} = \int_D \nu_j(y) \overline{\nu_k(y)} w(y) d\mu(y) = \int_D \nu_j(y) \overline{\nu_k(y)} d\rho(y) = \delta_{j,k}.
\]

Hence \( \sum_{i=1}^{m} E(X_i) = I \) is the identity matrix. Moreover, for any \( c \in \mathbb{C} \) we have

\[
cXc = \frac{1}{m} \left| \sum_{j \in S} c_j \sqrt{w(y_i)} \nu_j(y_i) \right|^2 \leq \frac{\left( \mathcal{N}(P_S, w) \right)^2}{m} \left\| \sum_{j \in S} c_j \nu_j \right\|_{L^2(D)}^2 = \frac{\left( \mathcal{N}(P_S, w) \right)^2}{m} \| c \|_2^2
\]

Since these matrices are self adjoint and positive semidefinite, we deduce that

\[
\lambda_{\text{max}}(X_i) \leq \frac{\left( \mathcal{N}(P_S, w) \right)^2}{m}.
\]

We now apply the matrix Chernoff bound \( \square \) with \( s, R = \left( \mathcal{N}(P_S, w) \right)^2 / m \) and

\[
\mu_{\text{min}} = \lambda_{\text{min}} \left( \sum_{i=1}^{m} E(X_i) \right) = \lambda_{\text{min}}(I) = 1,
\]

and likewise \( \mu_{\text{max}} = 1 \). This gives

\[
\mathbb{P} \left( \lambda_{\text{min}}(G) \leq 1 - \delta \text{ or } \lambda_{\text{max}}(G) \geq 1 + \delta \right) \\
\leq \mathbb{P} \left( \lambda_{\text{max}}(G) \leq (1 + \delta) \right) + \mathbb{P} \left( \lambda_{\text{max}}(G) \geq (1 - \delta) \right) \\
\leq s \cdot \exp \left( - \frac{(1 - \delta) \log(1 - \delta) + \delta}{m^{-1} \left( \mathcal{N}(P_S, w) \right)^2} \right) + \exp \left( - \frac{(1 + \delta) \log(1 + \delta) - \delta}{m^{-1} \left( \mathcal{N}(P_S, w) \right)^2} \right)
\]

Note that \((1 + \delta) \log(1 + \delta) - \delta \leq (1 - \delta) \log(1 - \delta) + \delta \) for \(0 < \delta < 1\). Hence

\[
\mathbb{P} \left( \lambda_{\text{min}}(G) \leq 1 - \delta \text{ or } \lambda_{\text{max}}(G) \geq 1 + \delta \right) \leq 2s \cdot \exp \left( - \frac{-c \delta}{m^{-1} \left( \mathcal{N}(P_S, w) \right)^2} \right) \leq \varepsilon,
\]

where in the last step we use the condition on \( m \). This completes the proof. \( \square \)

### 4 Sparse approximation via \( \ell^1 \)-minimization

Having discussed the case of Problem 1 we now consider the substantially more challenging setting of Problem 2. In order to facilitate its solution, we now also make an additional assumption on the dictionary \( \Phi \): namely, the index set \( \mathcal{I} \) is finite, and the functions \( \phi_i, i \in \mathcal{I} \), are linearly independent. In what follows we write \( n = |\mathcal{I}| \). Note that, typically, \( n \gg m \), where \( m \) is the number of measurements. It is notable that the examples described in §2.3 correspond to cases where the index set is countable. In this case, one may define \( \mathcal{I} \) as a large, but finite truncated index.
set in which the target set $S$ in the sparse representation (5) is expected to lie. We shall return to this matter briefly in §6 (see Remark 12).

This aside, we now also assume that $\mu_1 = \ldots = \mu_m = \mu$ in Assumption 1, i.e. $\mu$ is a probability measure that is absolutely continuous with respect to $\rho$ and for which the Radon–Nikodym derivative is strictly positive almost everywhere. In this case, the corresponding weight function $w$ satisfies

$$d\mu(y) = \frac{1}{w(y)} d\rho(y).$$

(38)

This is done to simplify several of the arguments later. However, it is also possible to consider distinct measures $\mu_1, \ldots, \mu_m$ as in the previous section.

### 4.1 Formulation

Given $\Phi = \{\phi_t : t \in T\}$, we strive to exploit the fact that $f$ is assumed to have an approximate sparse representation in $\Phi$. In this section, we do this via minimizing the $\ell^1(\mathcal{S}; \mathcal{V})$-norm of the coefficients, while also promoting fidelity of the resulting approximation to the measurements (2). There are various ways to do this, including the (\mathcal{V}-valued) Quadratically-Constrained Basis Pursuit (QCBP)

$$\hat{f} \in \arg\min_{p \in P_{\mathcal{S}; \mathcal{V}}} \left\{ \lambda \|c\|_{\ell^1(\mathcal{S}; \mathcal{V})} : \frac{1}{m} \sum_{i=1}^{m} w(y_i) \|f(y_i) + n_i - p(y_i)\|_{\mathcal{V}}^2 \leq \eta^2 \right\},$$

(39)

the LASSO

$$\hat{f} \in \arg\min_{p \in P_{\mathcal{S}; \mathcal{V}}} \left\{ \lambda \|c\|_{\ell^1(\mathcal{S}; \mathcal{V})} + \frac{1}{m} \sum_{i=1}^{m} w(y_i) \|f(y_i) + n_i - p(y_i)\|_{\mathcal{V}}^2 \right\},$$

(40)

or the Square-Root LASSO (SR-LASSO)

$$\hat{f} \in \arg\min_{p \in P_{\mathcal{S}; \mathcal{V}}} \left\{ \lambda \|c\|_{\ell^1(\mathcal{S}; \mathcal{V})} + \sqrt{\frac{1}{m} \sum_{i=1}^{m} w(y_i) \|f(y_i) + n_i - p(y_i)\|_{\mathcal{V}}^2} \right\}.$$

(41)

Here, in all cases, $c = (c_t)_{t \in T} \in \mathbb{V}_h^m$ are the coefficients of $p = \sum_{t \in T} c_t \phi_t \in P_{\mathcal{S}; \mathcal{V}}$. The focus of this work is not the choice (39), (40) or (41). We remark that (41) enjoys a known advantage over the other problems in that the theoretically-optimal value of the tuning parameter $\lambda$ is independent of the noise (see also Theorem 5), which in this case also includes the typically unknown error $f - f_\mathcal{S}$. For further background and in-depth comparison of these optimization problems, see [3].

Now let
\[
A = \frac{1}{\sqrt{m}} \left( \sqrt{w(y_i) \phi_j(y_i)} \right)_{i \in [m], j \in [n]} \in \mathbb{C}^{m \times n}, \tag{42}
\]

\[
v = \frac{1}{\sqrt{m}} \left( \sqrt{w(y_i)(f(y_i) + n_i)} \right)_{i \in [m]} \in \mathbb{V}_h^m, \tag{43}
\]

where \( \{ t_1, \ldots, t_n \} \) is an enumeration of the indices in \( \mathcal{J} \). Notice that we use the same notation for this matrix as in the previous section (see (15)). However, it is important to note that this matrix is generally short (fat), since \( n \geq m \), whereas the matrix (15) is \( m \times s \), and therefore tall. Then \( \hat{f} = \sum_{i \in \mathcal{J}} \hat{c}_i \phi_i \) is a solution of (39), (40) or (41) if and only if \( \hat{c} = (\hat{c}_i)_{i \in \mathcal{J}} \in \mathbb{V}_h^n \) is given by

\[
\hat{c} \in \text{argmin}_{z \in \mathbb{V}_h^n} \left\{ \lambda \| z \|_\ell^1([n]; V) + \| A z - v \|_\ell^2([m]; V) \right\},
\]

\[
\hat{c} \in \text{argmin}_{z \in \mathbb{V}_h^n} \left\{ \lambda \| z \|_\ell^1([n]; V) + \| A z - v \|_\ell^2([m]; V) \right\},
\]

or

\[
\hat{c} \in \text{argmin}_{z \in \mathbb{V}_h^n} \left\{ \| z \|_\ell^1([n]; V) : \| A z - v \|_\ell^2([m]; V) \leq \eta \right\},
\]

respectively.

**Remark 5 (Algorithms for solving (39)–(41)).** The \( \mathbb{V} \)-valued versions of the QCBP, LASSO, and SR-LASSO problems can be solved by considering reformulations of standard methods for solving their real and complex-valued counterparts. For example, in [36] the LASSO problem was solved by extending Bregman iterations and forward-backward iterations to the \( \mathbb{V} \)-valued case, while in [7] the \( \mathbb{V} \)-valued SR-LASSO problem is solved via primal-dual iterations. We shall not describe algorithms for solving (39)–(41) in any further detail, and refer the interested reader to [7, 36].

### 4.2 Accuracy, stability and sample complexity

As in §3, our main assumption will be a condition of the form (17), but with two differences. First, since the target set from which the sparse representation of \( f \) is obtained is unknown, we require this to hold for all subsets, not just a fixed subset. Second, as we see in the theorem below, we also require for it to hold for some value \( t \geq s \). The precise condition is as follows:

\[
\alpha \| p \|_{L_2^2(D)}^2 \leq \frac{1}{m} \sum_{i=1}^{m} w(y_i) | p(y_i) |^2 \leq \beta \| p \|_{L_2^2(D)}^2, \quad \forall p \in P_T, T \subseteq \mathcal{J}, |T| \leq t. \tag{44}
\]

In addition to this, we also recall that \( \Phi \) is a finite dictionary consisting of linearly independent elements. Therefore it is a Riesz basis, meaning that
Let $\Phi = \{ \phi_i : i \in I \} \subset L^2(D) \times [\alpha, \beta]$ be a finite dictionary consisting of $n$ linearly-independent functions, with bounds $a, b > 0$ as in (45). Let $1 \leq s \leq n, 0 < \alpha \leq \beta < \infty$, $y_i \in D, w : D \to [0, \infty)$ be such that $w(y_i)$ is well defined for all $i$, and suppose that (44) holds with $\lambda = \min \{ n, 2[4b_D a \alpha^{10}] \}$. Let $f \in L^2(D; \mathbb{V})$ with measurements $\{ y_i \}$ and consider the problem (41) with $\lambda = \frac{15b_D a \alpha^{10}}{26}. Then any solution $\hat{f}$ of (41) satisfies
\[
\| f - \hat{f} \|_{L^2(D; \mathbb{V})} \leq c_1 \left( \| x \|_{L^2(F; \mathbb{V})} \right) + \left( \| f - f_{\text{disc}} \|_{\text{disc}} + \| f - \hat{f} \|_{\text{disc}} + \| e \|_{C([\alpha, \beta])} \right),
\]
where $c = (c_1)_i \in \mathbb{V}$, $f_{\text{disc}} = \sum_{i \in \mathcal{I}} c_i \phi_i$ is the orthogonal projection (best approximation) of $f$ in $P_{\mathcal{I}}$, $e = \frac{1}{\sqrt{m}} \left( \sqrt{\| w(y_i) \|} \right)_{i=1}^{m} \in \mathbb{V}$ and
\[
c_1 = 8\sqrt{b}, \quad c_2 = 8\sqrt{b} \left( \frac{1}{2\lambda \sqrt{s}} + \frac{1}{\sqrt{a\alpha}} \right) + 1.
\]

This result (see 45, 7 for its proof) shows stable and accurate recovery for the solution $\hat{f}$ of (41) (similar results can also be shown for (39) and (40) – see 5 and 11, Chpt. 6). Specifically, the error is bounded by a multiple of $\| f \|_{\text{disc}}$ and $\| f - f_{\text{disc}} \|_{\text{disc}}$, which together measure how well $f$ can be represented by an $s$-sparse representation in $\Phi$ (observe that these terms vanish when $f$ has an exact $s$-sparse representation). The other terms are the space discretization error and the noise error. As in the case of weighted least squares (see Remark 1), it is also possible to replace the $\| \cdot \|_{\text{disc}}$-norm by the $L^2(D)$-norm when the sample points are random variables. We also remark in passing that the factor 8 in the constants is somewhat arbitrary. Other numerical values could also be used, subject to changing the numerical values in the definition of $f$ and $\lambda$.

We next consider sample complexity. The following result is analogous to Theorem 5 for the case of compressed sensing.

**Theorem 6 (Sample complexity for (44)).** Let $\Phi = \{ \phi_i : i \in I \} \subset L^2(D)$ be a finite dictionary consisting of $n$ linearly-independent functions, with bounds $a, b > 0$ as in (45). Let $1 \leq s \leq n, 0 < a \leq b < \infty$, Finally, before stating the main result, we need some additional notation. We write
\[
\sigma_s(x)_{L^2(F; \mathbb{V})} = \inf \left\{ \| x - z \|_{L^2(F; \mathbb{V})} : z \in \mathbb{V}^m \text{ is } s\text{-sparse} \right\}, \quad x \in \mathbb{V}^m.
\]
Here we recall that a Hilbert-valued vector $z = (z_i)_{i \in [m]} \in \mathbb{V}^m$ is $s$-sparse if it has at most $s$ nonzero entries, i.e. $| \{ i : z_i \neq 0 \} | \leq s$.

**Theorem 5 (Accuracy and stability of $\ell^1$-minimization).** Let $\Phi = \{ \phi_i : i \in I \} \subset L^2(D)$ be a finite dictionary consisting of $n$ linearly-independent functions, with bounds $a, b > 0$ as in (45). Let $1 \leq s \leq n, 0 < \alpha \leq \beta < \infty$, $\{ y_i \}_{i=1}^{m} \subset D, w : D \to [0, \infty)$ be such that $w(y_i)$ is well defined for all $i$, and suppose that (44) holds with $t = \min \{ n, 2[4b_D a \alpha^{10}] \}$. Let $f \in L^2(D; \mathbb{V})$ with measurements $\{ y_i \}$ and consider the problem (41) with $\lambda = \frac{15b_D a \alpha^{10}}{26}. Then any solution $\hat{f}$ of (41) satisfies
\[
\| f - \hat{f} \|_{L^2(D; \mathbb{V})} \leq c_1 \left( \| x \|_{L^2(F; \mathbb{V})} \right) + \left( \| f - f_{\text{disc}} \|_{\text{disc}} + \| f - \hat{f} \|_{\text{disc}} + \| e \|_{C([\alpha, \beta])} \right),
\]
where $c = (c_1)_i \in \mathbb{V}$, $f_{\text{disc}} = \sum_{i \in \mathcal{I}} c_i \phi_i$ is the orthogonal projection (best approximation) of $f$ in $P_{\mathcal{I}}$, $e = \frac{1}{\sqrt{m}} \left( \sqrt{\| w(y_i) \|} \right)_{i=1}^{m} \in \mathbb{V}$ and
\[
c_1 = 8\sqrt{b}, \quad c_2 = 8\sqrt{b} \left( \frac{1}{2\lambda \sqrt{s}} + \frac{1}{\sqrt{a\alpha}} \right) + 1.
\]
0 as in (45). Let $\mu$ be a probability measure satisfying Assumption $\square$ for some universal constant $0 < \delta^* < 1$, $0 < \varepsilon < 1$, $0 < \alpha \leq \beta < \infty$ and $y_1, \ldots, y_m$ be independent with $y_i \sim \mu$ for $i = 1, \ldots, m$. Define

$$ \Gamma = \Gamma(\Phi, w) = \left\| \max_{t \in \mathcal{F}} \left\{ \sqrt{w(t)} | \phi_t(t) | \right\} \right\|_{L^2_w(D)}, \tag{47} $$

where $w$ is the weight function specified in (38), and suppose that

$$ m \geq C \cdot \delta^{-2} \cdot (\Gamma^2/a) \cdot t \cdot (\log(en) \cdot \log^2(e(G^2/a) \cdot \delta) + \log(2/\delta)),$$

for some universal constant $C > 0$. Then (44) holds with $1 - \delta \leq \alpha \leq \beta \leq 1 + \delta$, with probability at least $1 - \varepsilon$.

As in the least-squares case, this result reduces the question of sample complexity to the matter of estimating a certain constant $\Gamma(\Phi, w)$ depending on the system $\Phi$ and the weight function $w$. Observe that $\Gamma^2 > a$ for any $\Phi$ and $w$. Indeed, $\Gamma^2 \geq w(y) | \phi_t(y) |^2$ almost everywhere, and therefore (7) and (45) give $\Gamma^2 \geq \| \phi_t \|^2_{L^2_w(D)} \geq a$.

Combining this with Theorem $\square$ we deduce the following:

**Corollary 1 (Sample complexity of $\ell^1$-minimization).** Let $\Phi = \{ \phi_t : t \in \mathcal{F} \} \subset L^1(D)$ be a finite dictionary consisting of $n$ linearly-independent functions, with bounds $a, b > 0$ as in (45). Let $1 \leq s \leq n$, $0 < \varepsilon < 1$, $\mu$ be a probability measure satisfying Assumption $\square$ for $i = 1, \ldots, m$ and $\Gamma$ be as in (47). Suppose that

$$ m \geq (b/a) \cdot (\Gamma^2/a) \cdot s \cdot (\log(en) \cdot \log^2(e(b/a)(\Gamma^2/a)s) + \log(2/\varepsilon)).$$

Then the following holds with probability at least $1 - \varepsilon$. Let $f \in L^1(D; \mathbb{V})$ with measurements (38) and consider the problem (44) with $A$ and $b$ as in (42) and $\lambda = c \sqrt{a/s}$ for some $0 < c < C$, where $C > 0$ is a universal constant. Then any solution $\hat{f}$ of (47) satisfies

$$ \| f - \hat{f} \|_{L^2(D; \mathbb{V})} \lesssim \sqrt{b} \cdot \frac{\sigma_c(e)^{\ell_1[|n|; \mathbb{V}]}}{\Gamma^2/a} + \sqrt{b/a} \left( \| f - \hat{f} \|_{\text{disc}} + \| f - \mathcal{P}_h(f) \|_{\text{disc}} + \| e \|_{\ell^2[|n|; \mathbb{V}]} \right).$$

**Proof.** Theorem $\square$ and the condition on $m$ show that (44) holds with $\tilde{\delta} = \delta^*/2$, where $t = \min\{ n, 2 \cdot \max(b_1, b_2) \}$. We now apply Theorem $\square$ noting that $c_1 \lesssim \sqrt{b}$ and $c_2 \lesssim (1/c + 1) \sqrt{b/a} + 1 \lesssim c \sqrt{b/a}$ in this case.

**Remark 6 (The Riesz basis constants $a, b$).** On closer inspection of the proofs, it is evident that it is possible to somewhat relax the assumption (45) by requiring only sparse subsets of $\Phi$ to form Riesz bases (with the same bounds). In Theorem $\square$, for example, it is possible to replace (45) with the weaker condition...
Towards optimal sampling for learning sparse approximations in high dimensions

\[ a\|c\|_{\ell_2(F)}^2 \leq \left\| \sum_{i \in F} c_i \phi_i \right\|_{L_2^2(D)}^2 \leq b\|c\|_{\ell_2(F)}^2, \quad c = (c_i)_{i \in F} \in \mathbb{C}^n \text{ is } t \text{-sparse}, \quad (48) \]

where \( t = \min\{n, 2\left\lceil 4s^{\beta/\alpha} \right\rceil \}. \) Even when \( \Phi \) forms a Riesz basis, when \( s \ll n \) the corresponding constants in (45) may be significantly better behaved than the Riesz basis constants in (45). In (45) we see an example where the lower constant \( a \) in (45) is extremely small, yet recovery is still possible from a reasonable number of measurements. This suggests that it may be important to use (48) instead of (45) in some scenarios.

4.3 Monte Carlo sampling

We now discuss the case of Monte Carlo sampling, which corresponds to the choice \( \mu = \rho \), i.e. \( w(y) \equiv 1 \). Corollary 1 shows that the sample complexity of \( \ell_1 \)-minimization is determined by the constant \( \Gamma \) defined in (47). In this case, we have

\[ \Gamma = \max_{i \in F} \|\phi_i\|_{L_\infty^2(D)} := \Theta = \Theta(\Phi), \]

which leads to the sample complexity bound

\[ m \gtrsim (b/a) \cdot (\Theta^2/a) \cdot s \cdot (\log(en) \cdot \log^2(1/e) + \log(2/\epsilon)). \quad (49) \]

It is worth comparing this bound with the least squares bound discussed in §3.3. Let \( S \subseteq \mathcal{F}, |S| \leq s \) and \( p \in \mathcal{P}_S \) be arbitrary. Write \( p = \sum_{i \in S} c_i \phi_i \). Then

\[ \|p\|_{L_\infty^2(D)} \leq \sum_{i \in S} |c_i| \|\phi_i\|_{L_\infty^2(D)} \leq \Theta \sqrt{s} \sqrt{\sum_{i \in S} |c_i|^2} \leq \Theta \sqrt{s} \|p\|_{L_2^2(D)}/\sqrt{a}. \]

Hence, (18) gives that

\[ (\mathcal{N}(\mathcal{P}_S))^2 \leq \Theta^2 s/a. \]

Therefore, and unsurprisingly, the sample complexity for \( \ell_1 \)-minimization in the setting of Problem 2 is always at least as large as least squares in the setting of Problem 1.

On the other hand, there are clearly instances where both sample complexities are the same, at least up to log terms. Recall that the functions \( \phi_i \) of Example 1 are equal to one in absolute value. Therefore, \( \Theta = 1 \) and, as discussed previously, \( (\mathcal{N}(\mathcal{P}_S))^2 = s \). Since this is an orthonormal basis, we also have \( a = b = 1 \) in this case. Hence, (49) reads

\[ m \gtrsim s \cdot (\log(en) \cdot \log^2(es) + \log(2/\epsilon)). \]
We conclude that Monte Carlo sampling in combination with least squares (in the setting of Problem 1) or $\ell^1$-minimization (in the setting of Problem 1) is near-optimal for sparse approximation via trigonometric polynomials.

By contrast, in the case of Example 2, the size of $\Theta$ depends on the choice of the finite index set $\mathcal{I}$. Using (9) and (25) we see that

$$\|\phi_1\|_{L^2(D)}^2 = \prod_{k=1}^d (2t_k + 1), \quad t = (t_k)_{k=1}^d,$$

and therefore

$$\Theta^2 = \max_{\mathbf{i} \in \mathcal{I}} \left\{ \prod_{k=1}^d (2i_k + 1) \right\}.$$  \hspace{1cm} (50)

Hence, if, for example,

$$\mathcal{I} = \mathcal{I}_{s}^{TP},$$

is the tensor product index set of order $s$ (see (10)) then it follows immediately that

$$\Theta^2 = (2s + 1)^d.$$  

Thus, the sample complexity bound behaves like $s(2s + 1)^d$, up to log factors. This grows exponentially with $d$, and always substantially exceeds $s$. Further, since $n = |\mathcal{I}| = (s + 1)^d$ in this case, this means that the sample complexity bound actually exceeds $n$; a situation that is, naturally, undesirable.

This situation can be ameliorated by choosing a truncated set $\mathcal{I}$ with fewer high-order polynomial indices, at the potential cost that important terms may be missed in the truncation. For example, let

$$\mathcal{I} = \mathcal{I}_{s-1}^{HC}$$

be the hyperbolic cross index set of order $s - 1$ (recall (12)). Then (50) gives

$$\Theta^2 \leq 2^d \max_{\mathbf{i} \in \mathcal{I}} \left\{ \prod_{k=1}^d (i_k + 1) \right\} \leq 2^d s.$$  

Hence, the sample complexity behaves like $2^d s$, up to log terms – in other words, substantially better than in the case of the tensor-product index set, but still exponentially large in $d$. Note that this bound is well suited when $d$ is comparatively small in relation to $s$. In the setting where $d$ is large, one can also show that

$$s^{\log 3 / \log 2} / 3 \leq \Theta^2 \leq s^{\log 3 / \log 2} + 1 \leq s^{2.58} \leq s^{2d}.$$

See [25, Lem. 3.5]. Thus, for large $d$, the same complexity bound scales like $s^{\log 3 / \log 2 + 1} = s^{2.58}$, up to log terms – in other words, polynomial in $s$, independently of $d$, albeit with a scaling that is substantially bigger than the optimal linear in $s$ scaling.
4.4 ‘Optimal’ sampling

With this in mind, we now consider how to choose the sampling measure to obtain a smaller sample complexity. Following ideas of [50], our aim is to minimize the quantity $\Gamma$ defined in (47). This is achieved by setting

$$\\left( w(y) \right)^{-1} = \frac{\max_{\iota \in I} |\phi_\iota(y)|^2}{\theta^2}, \quad \theta = \theta(\Phi) := \left( \int_{D} \max_{\iota \in I} |\phi_\iota(y)|^2 d\rho(y) \right)^{1/2}. \quad (51)$$

Notice that $\Gamma = \theta$ for this choice of $w$, which gives the sample complexity bound

$$m \gtrsim \left( \frac{b}{a} \right) \cdot \left( \frac{\theta^2}{a} \right) \cdot s \cdot \left( \log(en) \cdot \log^2(e(\theta^2/a)(b/a)s) + \log(2/\epsilon) \right), \quad (52)$$

provided $\mu$ satisfies (38), i.e.

$$d\mu(y) = \frac{\max_{\iota \in I} |\phi_\iota(y)|^2}{\theta^2} d\rho(y), \quad y \in \text{supp}(\rho). \quad (53)$$

Observe that the constant $\theta$ is always no larger than the constant $\Theta$ that appears in the Monte Carlo sampling estimate (49). Hence, we expect this choice of measures to be no worse than Monte Carlo sampling.

Remark 7 (The gap between Problem 1 and Problem 2). In the setting of Problem 1, we obtained a sampling measure in §3.4 that lead to near-optimal sample complexity, scaling linearly in $s$ for any $\Phi$ and subset $S$ of size $|S| = s$. Critically, this measure depended on the known, target set $S$. Conversely, in the setting of Problem 2, the measure defined above does not, in general, lead to near-optimal sample complexity. Indeed, it is not generally the case that $\theta = 1$. This constitutes a key gap between the two settings. That it exists should come of little surprise. The sampling measure used in the former setting depends completely on the target set. Yet, the whole purpose of the latter setting is to compute sparse approximations in the absence of this assumption. Hence, it is not unexpected that the sample measure defined above (which depends on $\mathcal{I}$ but not $S$) is not generally optimal.

Remark 8 (Arbitrarily-large improvements are possible). On the other hand, there are cases where the above sampling measure leads to substantial theoretical improvements. For example, let $D = [0, 1]$, $d\rho(y) = dy$ and

$$\phi_\iota(y) = y^{-\alpha} \exp(2\pi i y),$$

be the trigonometric polynomials scaled by a weight factor $y^{-\alpha}$ for some $0 < \alpha < 1/2$. Notice that $\{\phi_\iota : \iota \in \mathcal{I}\} \subset L_2^p(D)$ forms a Riesz basis for any finite $\mathcal{I}$. Clearly, in this case one has $\Theta = \infty$. On the other hand,

$$\theta^2 = \int_0^1 y^{-2\alpha} dy = \frac{1}{1-2\alpha},$$
is bounded, for any choice of \( \mathcal{I} \). The reason for this is that the functions \( \phi_i \) are all singular, yet their singularity occurs at the same place \( y = 0 \). The measure \( \mu \), which has the form

\[
d\mu(y) = \frac{\max_{i \in \mathcal{I}} |\phi_i(y)|^2}{\theta^2} \, d\rho(y) = \frac{y^{-2\alpha}}{1 - 2\alpha} \, dy,
\]
samples more densely near \( y = 0 \), thereby capturing the common singularity more efficiently. Note that a similar scenario occurs in the setting of algebraic polynomial approximation on the real line via Hermite polynomials. See [50, 59].

### 4.5 ‘Optimal’ sampling and discrete measures

As in the context of least squares, drawing samples from the measure (53) can be challenging. Fortunately, we can overcome this issue in the same way by introducing a finite grid. Let \( Z = \{z_i\}_{i=1}^k \) be such a grid and

\[
\tau = \frac{1}{k} \sum_{i=1}^k \delta_{z_i},
\]
(54)

be the discrete uniform measure supported on it. We then define the corresponding discrete measure \( \mu \) as

\[
d\mu(y) = \frac{\max_{i \in \mathcal{I}} |\phi_i(y)|^2}{\theta^2} \, d\tau(y),
\]

where

\[
\theta^2 = \int_D \max_{i \in \mathcal{I}} |\phi_i(y)|^2 \, d\tau(y) = \frac{1}{k} \sum_{i=1}^k \max_{i \in \mathcal{I}} |\phi_i(z_i)|^2.
\]
(55)

In other words,

\[
\mu = \sum_{i=1}^k \frac{\max_{i \in \mathcal{I}} |\phi_i(z_i)|^2}{\sum_{j=1}^k \max_{i \in \mathcal{I}} |\phi_i(z_j)|^2} \delta_{z_i}.
\]
(56)

Sampling from this measure is achieved as follows. Define the matrix

\[
B = \left( \frac{\phi_{ij}(z_i)}{\sqrt{k}} \right)_{i \in [k], j \in [n]} \in \mathbb{C}^{k \times n},
\]

and notice that

\[
\theta^2 = \sum_{i=1}^k \max_{j \in [n]} |B_{ij}|^2.
\]

Hence, \( y \sim \mu \) if

\[
\mathbb{P}(y = z_i) = \frac{\max_{j \in [n]} |B_{ij}|^2}{\sum_{j=1}^k \max_{j \in [n]} |B_{ij}|^2}, \quad i \in [k].
\]
As in the least-squares setting, sampling with respect to this measure, following the sample complexity bound (52) (with $\theta$ as in (55)), is sufficient to ensure an estimate with respect to the discrete measure $\tau$. From this, one can also obtain an error bound with respect to the original measure $\rho$, whenever the grid is sufficiently fine. Indeed, suppose that

$$\alpha'\|p\|_{L^2(\rho)}^2 \leq \|p\|_{L^2(\tau)}^2 \leq \beta'\|p\|_{L^2(\rho)}^2, \quad \forall p \in P,$$

(57)

for constants $\beta' \geq \alpha' > 0$ (note that this is analogous to (32), the difference being that we now require it to hold over $P$). Then (44) holds with respect to the $\rho$ measure with constants $\alpha'\alpha$ and $\beta'\beta$ whenever it holds with respect to the $\tau$ measure.

The key point is that the grid is required to satisfy essentially the same condition (57) – in other words, it should give rise to a discrete norm on $P$. One can construct such a grid exactly as in Remark 4.

### 4.6 Further discussion and numerical examples

We now numerically examine Question 3 in the context of the Examples 1 and 2 (we discuss Example 3 further in §5). Since the complex exponentials have absolute value equal to one, Example 1 is an instance where $\theta = \Theta = 1$, and where $\mu = \rho$. In other words, much as in the setting of Problem 1, sparse approximation in the case of Problem 2 using trigonometric polynomials is possible via Monte Carlo sampling with a number of samples that is proportional to $s$, multiplied by several log terms.

We next consider Example 2. Here, we recall from (50) that the relevant constant $\Theta^2$ for Monte Carlo sampling can become arbitrarily-large depending on the choice of $\mathcal{I}$. By contrast, we now show that this situation cannot occur when sampling according to the measure (53).

**Proposition 1.** Consider Example 2. Let $\mathcal{I}$ be a finite index set and $\Theta = \{\phi_\iota : \iota \in \mathcal{I}\}$, where the $\phi_\iota$ are as in (9). Then the constant $\theta = \theta(\Phi)$ defined by (51) satisfies

$$\theta^2 < 2^d.$$

In particular, when sampling from the corresponding measure (53), sample complexity estimate (52) is implied by

$$m \gtrsim 2^d \cdot s \cdot \left( \log(en) \cdot (d + \log(es))^2 + \log(2/\epsilon) \right).$$

(58)

**Proof.** The univariate Legendre polynomials $\phi_\iota$ satisfy the envelope bound

$$|\phi_\iota(y)| < \frac{2}{\sqrt{\pi}(1-y^2)^{1/4}}, \quad -1 < y < 1, \iota \in \mathbb{N}_0.$$
Tensor product: $\mathcal{I} = \mathcal{I}_{TP}^{\mathcal{J}}$

Total degree: $\mathcal{I} = \mathcal{I}_{TD}^{\mathcal{J}}$

Hyperbolic cross: $\mathcal{I} = \mathcal{I}_{HC}^{\mathcal{J}}$

Fig. 4: The constants $\Theta^2$ (top row), given by (50), and $\theta^2$ (bottom row), given by (55), against $n = |\mathcal{J}|$ in the case of Example 2 for different choices of $\mathcal{J}$ and dimensions $d$.

See, for example, [2, Eqn. (5.3)]. Observe that

$$\int_{-1}^{1} \frac{4}{\pi \sqrt{1-y^2}} \, dy = 2.$$ 

The result now follows by taking tensor products.

This result states that the sample complexity for the ‘optimal’ measure is linear in $s$, up to a constant that scales at worst like $2^d$. This is a marked improvement over the sample complexity bounds for Monte Carlo sampling. Moreover, as we see in the examples below, the constant $\theta^2$ can be substantially smaller than $2^d$ for certain choices of index set $\mathcal{J}$.

**Remark 9 (The preconditioning scheme).** The envelope bound (59) suggests an alternative strategy for choosing $w$, based on the choice

$$w(y) = \prod_{k=1}^{d} (\pi/2)(1-y_k^2)^{1/2}.$$ 

This is sometimes termed the *preconditioning* technique for sparse approximation with Legendre polynomials [59, 79]. The corresponding measure $\mu$ is precisely the arcsine (Chebyshev) measure. Because of (58), it leads to the same sufficient sample complexity bound (55) as sampling via the ‘optimal’ measure (51).

We now explore this effect numerically in the setting of Example 2. In order to avoid the difficulties of sampling from the continuous measure (53), we instead use the discrete measure (56) throughout. The fine grid consists of $k = 10n$ Monte Carlo points (recall Remark 4).
Towards optimal sampling for learning sparse approximations in high dimensions

In Figure 4, we plot the constants \( \Theta^2 \) and \( \theta^2 \) for several different choices of \( \mathcal{I} \). We notice several key effects. First, the constant \( \theta^2 \) is small – in fact, no greater than \( \approx 25 \) in all cases. It is much smaller than the bound \( 2^d \) shown above (notice that \( 2^{16} = 65,536 \)), which appears to be very pessimistic in practice. We also observe that \( \theta^2 \) is several times smaller than \( \Theta^2 \), suggesting better sample complexity when sampling from (56) instead of Monte Carlo sampling. On the other hand, the difference between the two quantities lessens in higher dimensions. This is not surprising, since the bad scaling of \( \Theta \) is caused by the presence of high polynomial indices. For fixed maximum size \( |\mathcal{I}| = n \), the index set \( \mathcal{I} \) contains fewer higher-order polynomials in higher dimensions than in lower dimensions. This suggests that Monte Carlo sampling may become more acceptable in higher dimensions. We show this effect in more detail next.

In Figures 5–7, we consider function approximation using the two sampling strategies. As we see, the ‘optimal’ strategy gives a nonnegligible benefit over Monte Carlo sampling in lower dimensions for \( f_1 \) and \( f_2 \). Yet, in higher dimensions this
benefit lessens. On the other hand, for \( f = f_3 \) the ‘optimal’ strategy yields no better performance, and actually a larger error than Monte Carlo sampling in high dimensions. The lessening benefit with increasing dimension is consistent with the results of Figure 4 wherein it is shown that the difference between the constants \( \Theta^2 \) and \( \theta^2 \) decreases as \( d \) increases. On the other hand, the observation that it can sometimes yield worse approximations is an important reminder that sampling strategy is designed to enhance the performance of sparse approximation in general, and may not therefore be the best strategy for any fixed function.

Remark 10. To elaborate on this previous comment, notice that the ‘optimal’ measure samples more densely near the boundary of the hypercube \([-1, 1]^d\), where the Legendre polynomials are larger, and therefore less densely near the origin. Hence, any function that varies most significantly in the interior of the domain is liable to be less well approximated by sampling from the ‘optimal’ measure. This is the case in particular for the function \( f = f_3 \), which is a product of one dimensional functions that are peaked around centres that get progressively closer to the origin with increasing index \( i \), and which are relatively flat away from their centre.
Towards optimal sampling for learning sparse approximations in high dimensions

\[(d, n, N) = (1, 399, 400)\]
\[(d, n, N) = (2, 152, 796)\]
\[(d, n, N) = (8, 22, 1843)\]
\[(d, n, N) = (16, 14, 4385)\]

Fig. 7: The same as Figure 5 but with \(f = f_3\) as in (33).

4.7 Proof of Theorems 5 and 6

Throughout this section, if \(x = (x_i)_{i=1}^n \in \mathbb{V}^n\) and \(S \subseteq [n]\) we use the notation \(x_S \in \mathbb{V}^n\) to denote the vector with \(i\)th entry equal to \(x_i\) if \(i \in S\) and 0 otherwise. Note that \(x_S\) is isomorphic to a vector in \(\mathbb{V}^{|S|}\). We will sometimes consider it as an element of this space. We now recall the following definition and lemma, which can be found in [6, Defn. 6 & Lem. 7]:

**Definition 2.** A matrix \(A \in \mathbb{C}^{m \times n}\) satisfies the robust Null Space Property (rNSP) of order \(1 \leq s \leq n\) over \(\mathbb{V}^n\) with constants \(0 < \rho < 1\) and \(\tau > 0\) if

\[
\|x_S\|_{\ell_2([|S|];\mathbb{V})} \leq \frac{\rho \|x_S\|_{\ell_1([|S|];\mathbb{V})}}{\sqrt{|S|}} + \tau \|Ax\|_{\ell_2([m];\mathbb{V})}, \quad \forall x \in \mathbb{V}^N,
\]

for any \(S \subseteq [n]\) with \(|S| \leq s\).

Here, in the second term, we recall that a matrix \(A \in \mathbb{C}^{m \times n}\) extends in the obvious way to a mapping \(\mathbb{V}^n \to \mathbb{V}^m\).
Lemma 3. Suppose that $A \in \mathbb{C}^{m \times n}$ has the rNSP of order $1 \leq s \leq n$ with constants $0 < \rho < 1$ and $\tau > 0$. Let $x \in \mathbb{V}^n$, $v = Ax + e \in \mathbb{V}^m$ and
\[
\lambda \leq \frac{C_1}{C_2\sqrt{s}};
\]
where $C_1 = \frac{(3\rho+1)(\rho+1)}{2(1-\rho)}$ and $C_2 = \frac{(3\rho+5)\tau}{2(1-\rho)}$. Then every minimizer $\hat{x} \in \mathbb{V}^n$ of the Hilbert-valued SR-LASSO problem
\[
\min_{z \in \mathbb{V}^n} \lambda \|z\|_{\ell^2 ([n]; \mathbb{V})} + \|Az - v\|_{\ell^2 ([m]; \mathbb{V})},
\]
satisfies
\[
\|\hat{x} - x\|_{\ell^2 ([n]; \mathbb{V})} \leq 2C_1 \frac{\mathbf{r}(x)_{\ell^2 ([n]; \mathbb{V})}}{\sqrt{s}} + \left( \frac{C_1}{\sqrt{5\lambda}} + C_2 \right) \|e\|_{\ell^2 ([m]; \mathbb{V})}.
\]

Proof of Theorem 5. We first claim that the matrix $A$ defined in (42) satisfies
\[
a \alpha \|c\|^2_{\ell^2 ([n]; \mathbb{V})} \leq \|Ac\|^2_{\ell^2 ([m]; \mathbb{V})} \leq b \beta \|c\|^2_{\ell^2 ([n]; \mathbb{V})},
\]
for all vectors $c \in \mathbb{V}^n$ that are $t$-sparse. Observe that any such vector $c$ corresponds to the coefficients of an element $p \in P_{T; \mathbb{V}}$ for $T \subseteq \mathcal{F}$, $|T| \leq t$, and also that $\|Ac\|^2_{\ell^2 ([m]; \mathbb{V})} = \frac{1}{t} \sum_{i=1}^{m} w(y_i) \|p(y_i)\|^2_{\mathbb{V}}$. We now recall from Lemma 2 that (44) also holds over $P_{T; \mathbb{V}}$ whenever it holds over $P_T \equiv P_{T; \mathcal{C}}$. Hence
\[
\alpha \|p\|^2_{\mathcal{L}_p^2 (D; \mathbb{V})} \leq \|Ac\|^2_{\ell^2 ([m]; \mathbb{V})} \leq \beta \|p\|^2_{\mathcal{L}_p^2 (D; \mathbb{V})}
\]
(61)
Now, using almost identical arguments to those used in the proof of Lemma 2, we find that the Riesz basis condition (45) also extends to the $\mathbb{V}$-valued case:
\[
a \|c\|^2_{\ell^2 (\mathcal{F}; \mathbb{V})} \leq \left\| \sum_{i \in \mathcal{F}} c_i \phi_i \right\|^2_{\mathcal{L}_p^2 (\mathbb{V})} \leq b \|c\|^2_{\ell^2 (\mathcal{F}; \mathbb{V})}, \quad \forall c = (c_i)_{i \in \mathcal{F}} \in \mathbb{V}^n.
\]
(62)
Since $p = \sum_{i \in \mathcal{F}} c_i \phi_i$, the claim now follows from this and (61).

We now show that $A$ satisfies the rNSP of order $s$ over $\mathbb{V}^n$ and derive values for the constants $\rho$ and $\tau$. The following argument is based on [11] Lem. 13.8. Let $c \in \mathbb{V}^n$ and $S \subseteq [n]$ with $|S| \leq s$. Suppose first that $t < n$, so that $t = \lfloor \frac{n-t}{\rho} \rfloor$ (we consider the case $t = n$ later). Define a partition $\Delta_1, \Delta_2, \ldots$ of $S^c$ as follows. First, let $\Delta_1$ be the index set of the largest $t' = t/2$ (notice that this is an integer, due to the definition of $t$) indices of the vector $\{\|c_j\|_{\mathbb{V}}\}_{j \in S'}$. $\Delta_2$ be the index set of the largest $t' = t/2$ indices of the vector $\{\|c_j\|_{\mathbb{V}}\}_{j \in (S \cup \Delta_1)^c}$ and so forth. This gives a partition of $S^c$ for which each set is of size $t'$, except possibly the final set (this is of no consequence to the argument). Consider the set $S \cup \Delta_1 \supseteq S$. Since $s \leq t'$ we have $|S \cup \Delta_1| \leq s + t' \leq 2t' = t$. Hence, we may apply (60) to obtain
Towards optimal sampling for learning sparse approximations in high dimensions

$$\|P_{S'}c\|_{\ell^2([n]:V)} \leq \|P_{S \cup \Delta_1}c\|_{\ell^2([n]:V)} \leq \frac{1}{\alpha \bar{\alpha}} \|AP_{S \cup \Delta_1}c\|_{\ell^2([n]:V)}.$$  (63)

We now write

$$\|AP_{S \cup \Delta_1}c\|_{\ell^2([n]:V)}^2 = \langle AP_{S \cup \Delta_1}c, Ac \rangle_{\ell^2([n]:V)} - \sum_{i \geq 2} \langle AP_{S \cup \Delta_1}c, AP_{\Delta_i}c \rangle_{\ell^2([n]:V)},$$

and then apply the Cauchy–Schwarz inequality and the (60) once more to get

$$\|AP_{S \cup \Delta_1}c\|_{\ell^2([n]:V)} \leq \|Ac\|_{\ell^2([n]:V)} + \sqrt{b \beta} \sum_{i \geq 2} \|P_{\Delta_i}c\|_{\ell^2([n]:V)}.$$  

Notice that, by construction of \(\Delta_i\) for each \(i \geq 2\), we have

$$\|P_{\Delta_i}c\|_{\ell^2([n]:V)}^2 \leq t' \max_{j \in \Delta_i} \|c_j\|_V^2 \leq t' \min_{j \in \Delta_{i-1}} \|c_j\|_V^2.$$

Since \(\|P_{\Delta_{i-1}}c\|_{\ell^1([n]:V)} \geq t' \min_{j \in \Delta_{i-1}} \|c_j\|_V\), this implies the following:

$$\|P_{\Delta_i}c\|_{\ell^1([n]:V)} \leq (t')^{-1/2} \|P_{\Delta_{i-1}}c\|_{\ell^1([n]:V)}, \quad i \geq 2.$$  

Using this and the previous expression, we deduce that

$$\|AP_{S \cup \Delta_1}c\|_{\ell^2([n]:V)} \leq \|Ac\|_{\ell^2([n]:V)} + \sqrt{\frac{b \beta}{t'}} \sum_{i \geq 1} \|P_{\Delta_i}c\|_{\ell^2([n]:V)}$$

$$= \|Ac\|_{\ell^2([n]:V)} + \sqrt{\frac{b \beta}{t'}} \|P_{S'}c\|_{\ell^1([n]:V)}.$$  

Substituting this back into (63) and noticing that \(b \beta / (a \alpha t') \leq 1/(2 \sqrt{3})\), we get

$$\|P_{S'}c\|_{\ell^2([n]:V)} \leq \frac{1}{\alpha \bar{\alpha}} \|Ac\|_{\ell^2([n]:V)} + \frac{1}{2 \sqrt{3}} \|P_{S'}c\|_{\ell^1([n]:V)},$$  (64)

for the case \(t < n\). Now suppose that \(t = n\). Since (60) now holds for any vector \(c \in Y^n\), we easily see that

$$\|P_{S'}c\|_{\ell^2([n]:V)} \leq \frac{1}{\alpha \bar{\alpha}} \|Ac\|_{\ell^2([n]:V)} \leq \frac{1}{\sqrt{\alpha \bar{\alpha}}} \|Ac\|_{\ell^2([n]:V)} + \frac{1}{2 \sqrt{3}} \|P_{S'}c\|_{\ell^1([n]:V)}.$$  

Hence, (64) also holds in this case as well. We deduce that \(A\) satisfies the rNSP of order \(s\) over \(Y^n\) with constants \(\rho = 1/2\) and \(\tau = 1/\sqrt{\alpha \bar{\alpha}}\).

Having shown this, we complete the proof by establishing the error bounds for \(\hat{f}\). Let \(\hat{c}\) be the coefficients of \(\hat{f}\). By the triangle inequality, we have
We deduce that
\[ 44 \quad B. \ Adcock, \ J. \ M. \ Cardenas, \ N. \ Dexter \ and \ S. \ Moraga \]

Now, observe that
\[ \text{Moreover, using (42) and (43), we see that} \]
\[ \text{Therefore, since} \ \mathcal{P}_h \ \text{is an orthogonal projection we have} \]
\[ \| f - \hat{f} \|_{L^2(D;V)} \leq \| f - \mathcal{P}_h(f) \|_{L^2(D;V)} + \| \mathcal{P}_h(f) - \mathcal{P}_h((u,f)) \|_{L^2(D;V)} + \| \mathcal{P}_h((u,f)) - \hat{f} \|_{L^2(D;V)}. \]

We first bound the term \( \| \mathcal{P}_h((u,f)) - \hat{f} \|_{L^2(D;V)} \). By (62), we have
\[ \| \mathcal{P}_h(p) - \hat{f} \|_{L^2(D;V)} = \left\| \sum_{i \in \mathcal{I}} (\mathcal{P}_h(c_i) - \hat{c}_i) \phi_i \right\|_{L^2(D;V)} \leq \sqrt{b} \| \mathcal{P}_h(c) - \hat{c} \|_{L^2([m];V)}, \]
where \( \mathcal{P}_h(c) \) is the vector \((\mathcal{P}_h(c_i))_{i \in \mathcal{I}}\). Using Lemma 3, we get
\[ \| \mathcal{P}_h(p) - \hat{f} \|_{L^2(D;V)} \leq 8\sqrt{b} \left( \frac{\sigma_i(\mathcal{P}_h(c))_{i \in \mathcal{I}}}{\sqrt{3}} + \left( \frac{1}{2\lambda\sqrt{3}} + \frac{1}{\sqrt{a\alpha}} \right) \| A \mathcal{P}_h(c) - b \|_{L^2([m];V)} \right). \]

Since \( \mathcal{P}_h \) is an orthogonal projection, we have
\[ \sigma_i(\mathcal{P}_h(c))_{i \in \mathcal{I}} \leq \sigma_i(c)_{i \in \mathcal{I}} \]
Moreover, using (42) and (43), we see that
\[ \| A \mathcal{P}_h(c) - v \|_{L^2([m];V)} = \left\| \frac{1}{\sqrt{m}} \left( \sqrt{w(y_i)} (\mathcal{P}_h(f)(y_i)) - f(y_i) - n_i \right)_{i \in \mathcal{I}} \right\|_{L^2([m];V)} \]
\[ \leq \left\| \frac{1}{\sqrt{m}} \left( \sqrt{w(y_i)} (\mathcal{P}_h(f)(y_i)) - \mathcal{P}_h(f)(y_i) \right)_{i \in \mathcal{I}} \right\|_{L^2([m];V)} + \left\| \frac{1}{\sqrt{m}} \left( \sqrt{w(y_i)} (f(y_i) - \mathcal{P}_h(f)(y_i)) \right)_{i \in \mathcal{I}} \right\|_{L^2([m];V)} + \| e \|_{L^2([m];V)}. \]

Now, observe that
\[ \left\| \sqrt{w(y_i)} (\mathcal{P}_h(f)(y_i)) - \mathcal{P}_h(f)(y_i) \right\|_{V} \leq \sqrt{w(y_i)} \| f_{\mathcal{I}}(y_i) - f(y_i) \|_{V}. \]

We deduce that
\[ \| A \mathcal{P}_h(c) - v \|_{L^2([m];V)} \leq \| f - f_{\mathcal{I}} \|_{L^2([m];V)} + \| f - \mathcal{P}_h(f) \|_{L^2([m];V)} + \| e \|_{L^2([m];V)}. \]
Substituting this into (65) and then combining with (66) now completes the proof.

We now prove Theorem 6. For this, we use the following result, which was shown in [20, Thm. 1.1]:

**Theorem 7.** There exist absolute constants \(c_0, c_1 > 0\) such that the following holds. Let \(X_1, \ldots, X_m\) be independent copies of a random vector \(X \in \mathbb{C}^n\) such that \(\|X\|_\infty \leq K\) almost surely for some \(K > 0\). Let \(\mathcal{T} \subseteq \{c \in \mathbb{C}^n : \|c\|_{t} \leq \sqrt{s}\}\), \(\delta \in (0, \kappa)\), \(0 < \varepsilon < 1\) and suppose that

\[
m \geq c_0 \cdot K^2 \cdot \delta^{-2} \cdot s \cdot (\log(en) \log^2(sK^2/\delta) + \log(2/\varepsilon)).
\]

Then, with probability at least \(1 - \varepsilon\), we have

\[
\sup_{c \in \mathcal{T}} \left| \frac{1}{m} \sum_{i=1}^{m} |\langle X_i, c \rangle|^2 - \mathbb{E}|\langle X, c \rangle|^2 \right| \leq c_1 \delta \left( 1 + \sup_{c \in \mathcal{T}} \mathbb{E}|\langle c, X \rangle|^2 \right).
\]

**Proof of Theorem 6.** Note that the result holds, provided

\[
\left| \frac{1}{m} \sum_{i=1}^{m} w(y_i)|p(y_i)|^2 - \|p\|_{L^2(D)}^2 \right| \leq \delta, \quad \forall p \in P_T, \|p\|_{L^2(D)} \leq 1, \ T \subseteq \mathcal{T}, \ |T| \leq t.
\]

Define the random vector \(X = (\sqrt{w(y)}\phi_j(y))_{j \in [N]}\), where \(y \sim \mu\). Let \(p \in P_T\) for some \(T \subseteq \mathcal{T}\) with \(|T| \leq t\), and write \(p = \sum_{i \in T} c_i \phi_i\). Then

\[
\frac{1}{m} \sum_{i=1}^{m} w(y_i)|p(y_i)|^2 = \frac{1}{m} \sum_{i=1}^{m} |\langle X_i, c \rangle|^2,
\]

and, since \(d\mu(y) = (w(y))^{-1}d\rho(y)\),

\[
\|p\|_{L^2(D)}^2 = \int_D |p(y)|^2 d\rho(y) = \int_D w(y) \left| \sum_{i \in T} c_i \phi_i(y) \right|^2 d\mu(y) = \mathbb{E}|\langle X, c \rangle|^2.
\]

Hence, it suffices to show that

\[
\left| \frac{1}{m} \sum_{i=1}^{m} |\langle X_i, c \rangle|^2 - \mathbb{E}|\langle X, c \rangle|^2 \right| \leq \delta, \quad \forall c \in \mathcal{T}, \quad (67)
\]

where

\[
\mathcal{T} = \{c \in \mathbb{C}^n : c \text{ is } t\text{-sparse and } \mathbb{E}|\langle X, c \rangle|^2 \leq 1\}.
\]

Notice that if \(c \in \mathcal{T}\) then, by (45),

\[
\|c\|_{t} \leq \sqrt{t}/\|c\|_{t} \leq \sqrt{t/a} \left\| \sum_{i} c_i \phi_i \right\|_{L^2(D)}^2 = \sqrt{t/a} \sqrt{\mathbb{E}|\langle X, c \rangle|^2} \leq \sqrt{t/a}.
\]
Hence \( \mathcal{T} \subseteq \{ c \in \mathbb{C}^n : ||c||_1 \leq \sqrt{t/a} \} \). Therefore, we may apply Theorem 7 to get that

\[
\left| \frac{1}{m} \sum_{i=1}^{m} |\langle X_i, c \rangle|^2 - \mathbb{E}[|\langle X, c \rangle|^2] \right| \leq c_1 \delta' \left( 1 + \sup_{c \in \mathcal{T}} \mathbb{E}[|\langle c, X \rangle|^2] \right) \leq 2c_1 \delta',
\]

for \( 0 < \delta' < \kappa \). This holds with probability at least \( 1 - \varepsilon \), provided

\[
m \geq c_0 \cdot K^2 \cdot (\delta')^{-2} \cdot (t/a) \cdot \left( \log(en) \log^2(tK^2/(a\delta')) + \log(2/e) \right).
\]

Observe that \( \|X\|_\infty = \max_{\phi \in \mathcal{F}} \{ \sqrt{w(y)|\phi_\delta(y)|} \} \), and therefore we may take \( K = \Gamma = \max_{\phi \in \mathcal{F}} \{ \sqrt{w(\cdot)|\phi_\delta(\cdot)|} \} \|_{L^2(D)} \). We now set \( \delta' = \delta/(2c_1) \) to deduce that (67) holds for \( 0 < \delta < \kappa/(2c_1) \) with probability at least \( 1 - \varepsilon \), provided

\[
m \geq 4c_0c_1^2 \cdot \Gamma^2 \cdot \delta^{-2} \cdot (t/a) \cdot \left( \log(en) \log^2(2c_1t\Gamma^2/(a\delta)) + \log(2/e) \right).
\]

To complete the proof, we simply notice that

\[
\log(2c_1t\Gamma^2/(a\delta)) \leq \log(2c_1\Gamma^2/(a\delta)) + \log(2c_1/e) \leq \log(2c_1\Gamma^2/(a\delta)),
\]

since \( t\Gamma^2/(a\delta) > 1 \).

5 A novel approach for sparse polynomial approximation on irregular domains

In this section, we focus on Example 3 in the context of \( \ell^1 \)-minimization. As we have seen in the previous section, the various measurement conditions depend on the Riesz basis constants \( a, b \) of the system \( \Phi = \{ \phi_t : t \in \mathcal{T} \} \). Unfortunately, in cases such as Example 3, these constants are often very poorly behaved, especially when the polynomial degree is large. Later, in Figs. 11–13 we compute \( a, b \) for various different domains. While we note that these constants may be pessimistic (in particular, recall Remark 6), the fact remains that dealing with poorly-conditioned dictionaries may well be problematic in the setting of \( \ell^1 \)-minimization for sparse approximation.

5.1 Method

Inspired by our earlier use of discrete measures to effect optimal sampling, in this section we consider a different approach for approximation on irregular domains in which the original Riesz basis \( \Phi \) is orthogonalized over the discrete grid. Specifically, let \( Z = \{ z_t \}_{t=1}^T \) be a finite grid and \( \tau \) be as in (54). Then we construct a new basis \( \Upsilon = \{ \upsilon_t : t \in \mathcal{T} \} \) that is orthonormal with respect to \( \tau \), and subsequently use this basis in which to construct the approximation \( \hat{f} \) to \( f \) via \( \ell^1 \)-minimization.

As in §3.5, the orthonormal basis \( \Upsilon \) is constructed via QR factorization. Let
Towards optimal sampling for learning sparse approximations in high dimensions 47

\[ B = \left( \phi_i(z_i) / \sqrt{k} \right)_{i \in [k], j \in [n]} \in \mathbb{C}^{k \times n}, \]

and suppose that \( B \) has QR factorization \( B = QR \), where \( Q \in \mathbb{C}^{k \times n} \) and \( R \in \mathbb{C}^{n \times n} \). Then, this orthonormal basis is given by

\[ \psi_i(y) = \sum_{j=1}^{i} (R^{-\top})_{ji} \phi_{ij}(y), \quad i \in [n]. \tag{68} \]

In what follows, we compare two sampling strategies for the basis \( \Upsilon \). First, Monte Carlo sampling from the underlying measure \( \tau \), i.e.

\[ P(y = z_i) = \frac{1}{k}, \quad i \in [k]. \tag{69} \]

Second, the ‘optimal’ sampling measure identified in §4.5. In this case, due to the orthogonalization, this is given by

\[ P(y = z_i) = \frac{\max_{j \in [n]} |Q_{ij}|^2}{\sum_{i=1}^{k} \max_{j \in [n]} |Q_{ij}|^2}, \quad i \in [k]. \tag{70} \]

As shown in the previous section, the sample complexity bounds for these two strategies depend on their respective constants \( \Theta \) and \( \theta \). Because of the previous definition of the basis, these are given by

\[ \Theta = \Theta(\Upsilon) = \sqrt{k} \max_{i \in [k]} |Q_{ij}|, \]

for the former, and

\[ \theta = \theta(\Upsilon) = \sqrt{\sum_{i \in [k]} \max_{j \in [n]} |Q_{ij}|^2}, \]

for the latter.

5.2 Orderings

Suppose that \( f \) has a sparse representation in the system \( \Phi \). Then it is not guaranteed to possess a sparse representation in the basis \( \Upsilon \), since the \( j \)th basis function \( \psi_{ij} \) is a linear combination of the functions \( \phi_{i1}, \ldots, \phi_{ij} \). In particular, the sparsity of \( f \) in the representation \( \Upsilon \) will be heavily influenced by the ordering of the basis \( \{\iota_1, \ldots, \iota_n\} \) of the indices in \( \mathcal{I} \).

To illustrate this, in Figs. we compare ordering the multi-indices in \( \mathcal{I} \) lexicographically to ordering them according to increasing total degree, i.e. the value \( t_1 + \ldots + t_d \) for \( i = (t_k)_{k=1}^d \). We remark in passing that ordering according maximum,
Fig. 8: The absolute values of the coefficients of the function \( f = f_1 \) over the domain \( D = D_2 \) with respect to the bases \( \Phi \) and \( \Upsilon \), where \( \mathcal{F} = \mathcal{F}^{HC} \) is the hyperbolic cross index set. The coefficients are sorted from largest in absolute value to smallest. In the top row, the multi-indices in \( \mathcal{F} \) are sorted lexicographically. In the bottom row, they are sorted according to increasing total degree (i.e. the value \( \sum_{k=1}^{d} \iota_k \) for \( \iota = (\iota_k)_{k=1}^{d} \)).

\[
(d, t, n) = (2, 152, 796) \quad (d, t, n) = (8, 22, 1843) \quad (d, t, n) = (16, 14, 4385)
\]

Fig. 9: The same as in Fig. 8 except for \( f = f_1 \) and \( D = D_3 \).

i.e. the value \( \max_{k=1, \ldots, d} \{ \iota_k \} \) produces similar results. To examine the sparsity in either basis, we plot the coefficients of a function \( f \) sorted from largest to smallest in absolute value. In particular, the more rapidly the coefficients decrease, the better \( f \) is approximated by a sparse representation in the given basis.

As is evident, lexicographic ordering always leads to a deterioration in sparsity when switching from the basis \( \Phi \) to the basis \( \Upsilon \). This is of little surprise. On the
other hand, using the total degree ordering can substantially improve the situation. In Fig. 8 it actually leads to better sparsity and in Fig. 9 it yields better sparsity in \( d = 2 \) dimensions, and similar sparsity in \( d = 8, 16 \) dimensions. Finally, in Fig. 10, while still leading to worse sparsity than in the original basis \( \Phi \), it is still generally better than lexicographic ordering.

For this reason, in our subsequent experiments, we employ the total degree ordering. Naturally, this discussion leads to the question of the optimal ordering. We anticipate this to be function dependent, and it is outside the scope of this work to discuss it further. In practice, we expect a good ordering could be estimated from a set of candidate orderings via cross validation.

### 5.3 Numerical examples

In Fig. 11 and 12 we compare the orthogonalization strategy \( \Upsilon \) against the original Legendre basis restricted to the irregular domain \( D \) (labelled \( \Phi \)). Several effects are notable. First, orthogonalizing the basis generally leads to better performance than using the original Legendre basis. This is consistent with the observation that the various sample complexity bounds depend on the Riesz basis constants \( a, b > 0 \) which, as noted and as shown numerically in these figure, can behave wildly for irregular domains. On the other hand, it is clear that the values constants are extremely pessimistic when it comes to predicting the actual performance. Even when the constant \( a = a_L \) is exceedingly small, the approximation based on the Legendre basis still offers a reasonable error in most cases. Further, as seen in Figs. 8 and 9 the orthogonalization strategy leads to slightly improved sparsity. Therefore, it
is unclear what property of orthogonalization is driving the better approximation, whether it be the smaller Riesz basis constants or the improvement in sparsity. On the other hand, in Fig. 13 we present an example where orthogonalization worsens the approximation. This we expect is due to the worse sparsity in the $\Upsilon$ basis, as shown in Fig. 10.

Second, we observe that the ‘optimal’ sampling procedures generally outperform Monte Carlo sampling in lower dimensions, while this improvement lessens in higher dimensions, or may actually lead to slightly worse performance. This is consistent with the observation made previously in §4.6. We also report the values of the constants $\theta$ and $\Theta$ in all cases. It is notable that, in the case of the orthogonalized basis, the corresponding constant $\Theta = \Theta_0$ is much larger than $\theta = \theta_0$, even in high dimensions. However, this is not reflected in the approximation errors for the two sampling strategies, which are similar, thus suggesting a gap between the theoretical guarantees and performance on actual function approximation problems.

6 Structured sparse approximation

Our main assumption throughout this chapter has been that $f$ admits an approximately sparse representation in a dictionary $\Phi = \{\phi_t : t \in \mathcal{I}\}$. We conclude this chapter with a brief discussion on several types of structured sparsity models. Our focus throughout is on the setting of Problem 2. Hence, as in §4, we assume that $\Phi$ is a finite set of $n$ linearly-independent functions.

6.1 Weighted sparsity and weighted $\ell^1$-minimization

Let $v = (v_t)_{t \in \mathcal{I}}$ be a vector of positive weights. For a set $S \subseteq \mathcal{I}$, we define its weighted cardinality as

$$|S|_v = \sum_{t \in S} v_t^2.$$  

In the weighted sparsity model, given a weights $v$ and a weighted sparsity $k > 0$, we assume that a function $f \in L_2^2(D; \mathcal{V})$ has a sparse representation $f_S$ of the form (5) for some set $S \subseteq \mathcal{I}$ with $|S|_v \leq k$. Note that, unlike the case of standard sparsity, the weighted sparsity parameter can take any positive value in this setting – hence our reason for using the notation $k$ instead of $s$.

Fortunately, promoting weighted sparsity structure is straightforward. Rather than $\ell^1$-minimization, i.e. (39), (40) or (41), we consider a weighted $\ell^1$-minimization problem. For example, we may replace (41) by

$$\hat{f} \in \arg\min_{p \in \mathcal{P}_f(V)} \left\{ \lambda \|e\|_{\ell_1(S; \mathcal{V})} + \sqrt{\frac{1}{m} \sum_{i=1}^m w(y_i) \|f(y_i) + n_i - p(y_i)\|_2^2} \right\}.  \tag{71}$$
Towards optimal sampling for learning sparse approximations in high dimensions

\[(d, t, n) = (1, 399, 400)\]
\[(\theta_2^L, \Theta_2^L) = (2.25, 303.73)\]
\[(\theta_2^Q, \Theta_2^Q) = (5.19, 768.17)\]
\[(a_L, b_L) = (5.29 \times 10^{-37}, 2.55)\]
\[(a_Q, b_Q) = (1, 1)\]

\[(d, t, n) = (2, 152, 796)\]
\[(\theta_2^L, \Theta_2^L) = (2.51, 30.26)\]
\[(\theta_2^Q, \Theta_2^Q) = (9.39, 508.06)\]
\[(a_L, b_L) = (3.69 \times 10^{-15}, 1.86)\]
\[(a_Q, b_Q) = (1, 1)\]

\[(d, t, n) = (8, 22, 1843)\]
\[(\theta_2^L, \Theta_2^L) = (1.99, 5.77)\]
\[(\theta_2^Q, \Theta_2^Q) = (66.99, 3659)\]
\[(a_L, b_L) = (9.31 \times 10^{-9}, 20.86)\]
\[(a_Q, b_Q) = (1, 1)\]

\[(d, t, n) = (16, 14, 4385)\]
\[(\theta_2^L, \Theta_2^L) = (1.74, 2.94)\]
\[(\theta_2^Q, \Theta_2^Q) = (132, 11406)\]
\[(a_L, b_L) = (2.88 \times 10^{-9}, 188.55)\]
\[(a_Q, b_Q) = (1, 1)\]

Fig. 11: The relative error \((35)\) versus \(m\) for \(\ell^1\)-minimization in the case of Example 3, where \(\mathcal{J} = \mathcal{J}_{HC}^{\ell^1}\) is the hyperbolic cross index set \((12)\) and \(f = f_1\) and \(D = D_2\) are as in \((33)\) and \((34)\), respectively. This figure compares the Legendre basis on \([-1, 1]^d\) restricted to \(d\) and the orthonormal basis on \(D\) constructed via \((68)\). In the former case, the sampling strategies are Monte Carlo sampling from the continuous uniform measure on \(D\) (labelled ‘LU’) and sampling from the discrete ‘optimal’ measure \((53)\) (‘LO’). In the latter cases, the sampling strategies are Monte Carlo sampling \((69)\) from the discrete uniform measure (‘QU’) and sampling from the discrete ‘optimal’ measure \((70)\) (‘QO’). We also report the values of the corresponding constants \(\theta^2\) and \(\Theta^2\) for both bases (labelled ‘L’ and ‘Q’, respectively), as well as the Riesz basis constants \(a, b\) with respect to the discrete measure \(\tau\).

Here, \(\|c\|_{\ell^1(\mathcal{J}, \mathcal{V})} = \sum_{I \in \mathcal{J}} v_I \|c_I\|_\mathcal{V}\) is the weighted \(\ell^1\)-norm of a Hilbert-valued vector \(c = (c_I)_{I \in \mathcal{J}}\).
Remark 11. Much like the lower set assumption (see §3.3), weighted sparsity is a natural assumption to consider when one expects the most significant coefficients of $f$ to correspond to lower-order terms. We discuss the relation between the two models later. This is typically the case for smooth function approximation using algebraic or trigonometric polynomials. As we see later, incorporating slowly-growing weights can lead to a significant improvement over unweighted $\ell^1$-minimization. Note that weighted sparsity and weighted $\ell^1$-minimization were first elaborated in [80], before further developments in [1, 2, 8, 25]. Other works on incorporating weights into sparse polynomial approximation include [5, 75, 99].
As in the case of standard sparsity, successful recovery via (71) follows from a norm equivalence similar to (44): namely,

$$\alpha \|p\|_{L^2(D)}^2 \leq \frac{1}{m} \sum_{i=1}^{m} w(y_i) |p(y_i)|^2 \leq \beta \|p\|_{L^2(D)}^2, \quad \forall p \in P_T, \ T \subseteq \mathcal{I}, \ |T| \leq t. \quad (72)$$

Under this condition, one obtains an error bound identical to Theorem 5, except with $\sigma_\ell(c, \ell_v(x, V))$ replaced by the weighted term

$$\sigma_{k,v}(x) = \inf \left\{ \|x - z\|_{\ell_v(x, V)} : z \in \mathbb{V}^n \text{ is weighted } (k,v)-\text{sparse} \right\}, \quad x \in \mathbb{V}^n,$$
and $s$ replaced by $k$. For the sake of succinctness, we omit the details, and refer to [8,30] (the results therein are given in the scalar-valued case, but readily extend to the Hilbert-valued case).

Since our primary focus is on the question of sample complexity, we now state a variant of Theorem 6 for (72) for the weighted sparse model:

**Theorem 8 (Sample complexity for (72)).** Let $\Phi = \{ \phi_t : t \in \mathcal{I} \} \subset L^2_{\rho}(D)$ be a finite dictionary consisting of $n$ linearly-independent functions, with bounds $a, b > 0$ as in (45). Let $\mu$ be a probability measure satisfying Assumption 1, $v = (v_t)_{t \in \mathcal{I}}$ be weights with

$$v_t \geq \| \sqrt{w}(\cdot)\phi_t(\cdot) \|_{L^\infty_{\rho}(\mathcal{D})}, \quad \forall t \in \mathcal{I},$$

where $w$ is as in (38), $1 \leq t \leq n$, $0 < \delta < \delta^*$ for some universal constant $0 < \delta^* < 1$, $0 < \varepsilon < 1$, $0 < \alpha \leq \beta < \infty$, and $y_1, \ldots, y_m$ be independent with $y_i \sim \mu$ for $i = 1, \ldots, m$. Suppose that

$$m \geq C \cdot \delta^{-2} \cdot a^{-1} \cdot t \cdot (\log(en) \cdot \log^2(ea^{-1}t/\delta) + \log(2/\varepsilon)),$$

for some universal constant $C > 0$. Then (72) holds with $1 - \delta \leq \alpha \leq \beta \leq 1 + \delta$, with probability at least $1 - \varepsilon$.

We omit the proof of this result, since it is similar to that of Theorem 7, the main difference being the use of [20, Thm. 2.13] instead of [20, Thm. 1.1].

### 6.2 Sparsity in lower sets

The lower set sparsity model differs from the weighted sparsity model in that it imposes a lower set structure, as opposed to a weighted sparsity structure. But in practice it can also be effected via weighted $\ell^1$-minimization with specific choices of weights. In the lower set sparsity model, we suppose that $f \in L^2_{\rho}(D; \mathcal{V})$ has a sparse representation $f_S$ of the form (5) for some set $S \subseteq \mathcal{I}$ with $|S| \leq s$ that is also lower.

In terms of sufficient conditions for lower set recovery, one’s first thought may be to consider a variant of (44) with the additional assumption that the sets $T$ be lower. Unfortunately, it is not known whether such an approach can work. The difficulty lies with the fact that it is unclear how to promote lower set structure directly via a convex penalty term such as the $\ell^1$-norm.

Instead, the approach originally proposed in [2,25] is to use weighted sparsity as a surrogate for lower set sparsity. This is done by choosing weights $u = (u_t)_{t \in \mathcal{I}}$ as small as possible so that Theorem 5 applies, namely,

$$u_t = \| \sqrt{w}(\cdot)\phi_t(\cdot) \|_{L^\infty_{\rho}(\mathcal{D})}, \quad \forall t \in \mathcal{I},$$

and defining the weighted sparsity as

$$k = k(s; w) = \max \{|S| : |S| \leq s, S \text{ lower}\}.$$
Towards optimal sampling for learning sparse approximations in high dimensions

Note that this ensures that every lower set of size $s$ has weighted cardinality at most $k$, i.e.

$$\{S : |S| \leq s, S \text{ lower}\} \subseteq \{S : |S|_w \leq k(s;w)\}.$$  

As a result, we can promote lower set sparsity by solving the weighted $\ell^1$-minimization problem (71) with weights $v = u$.

**Remark 12 (The choice of $\mathcal{J}$).** Working with lower sets also yields a strategy for choosing the large truncated set $\mathcal{J}$ (recall the discussion at the beginning of §4). Indeed, it is a straightforward exercise to show that the union of all lower sets of size at most $s$ is the hyperbolic cross $\mathcal{J}^{HC}$. Hence, the target lower set $S$ in the sparse representation (5) is guaranteed to lie within this index set, thus giving a clear rationale for choosing this set.

### 6.3 Sampling and numerical experiments

We now discuss the matter of sampling. Notice that, unlike in the case of Theorem 6, the sample complexity bound in Theorem 8 does not involve a constant $\Gamma$ depending on the basis $\Phi$ and weight function $w$. This dependence only arises in the minimum size condition (73) on the weights $v$. For Monte Carlo sampling ($w \equiv 1$), this condition may be quite stringent if the $L^\infty$-norms of the basis functions grow rapidly. Hence, this conditions suggests choosing $w$ to minimize the right-hand side of (73). This leads to the same choice (51) and (53) as in the standard sparsity setting considered previously.

The case of lower set sparsity allows for a more concrete discussion. Choosing weights $v = u$ as discussed above, and invoking Theorem 8 leads to a measurement condition of the form

$$m \gtrsim a^{-1} k(s; w) \cdot (\log(e) \cdot \log(4e^{a^{-1} k(s; w)} / \delta) + \log(2/\epsilon)),$$

for recovering functions with sparse representations in lower sets. Hence, the objective is to minimize $k(s; w)$ with respect to $w$. In the case of Monte Carlo sampling, we have

$$k(s; w) = k(s; 1) = \max \{|S|_w : |S| \leq s, S \text{ lower}\}.$$  

Consider, for illustration, Example 2. In this case, since the Legendre polynomials all attain their maximum at the same point $y = (1, \ldots, 1)^T$, we have

$$|S|_w = (\mathcal{N}(P_3))^2,$$

where $\mathcal{N}(P_3)$ is the unweighted Nikolskii constant (24). Recall from the discussion in §3.3 that $\mathcal{N}(P_3)$ satisfies the sharp bound $(\mathcal{N}(P_3))^2 \leq s^2$ for lower sets. In other words, $k(s; 1) = s^2$ in this case, leading to a sample complexity bound of the form

$$m \gtrsim s^2 \cdot (\log(en) \cdot \log^2(e$s) + \log(2/\epsilon))$$
In other words, the worst-case sample complexity for lower set recovery via Monte Carlo sampling is the same (up to constants and log factors) as that of least squares in the setting of Problem 1. See \cite{8} for further discussion.

Having considered Monte Carlo sampling for lower set recovery, one may also consider how to choose the weight function $w$ and corresponding sampling measure $\mu$ to improve the sample complexity. The best solution in this case involves choosing $w$ to minimize

$$
  k(s; w) = \max \left\{ \sum_{i \in S} \left\| \sqrt{w(\cdot)} \phi_i(\cdot) \right\|_{L^2(\mathcal{P}(\mathcal{W}))}^2 : \left| S \right| \leq s, S \text{ lower} \right\},
$$

over all strictly positive and finite almost everywhere weight functions on $\text{supp}(\rho)$ for which (7) holds. Unfortunately, even after resorting to a discrete measure as in §4.5, it is unclear how to compute such a $w$, since it seemingly involves enumerating all lower sets. As shown in \cite{34}, there are many lower sets in high dimensions (for example, at least $\binom{d}{s-1}$ when $s \leq d + 1$).

Since the optimal choice of $w$ (in the sense of minimizing $k(s; w)$) may not be available, it is natural to consider how one might choose a good $w$. One option involves the choice (51). This leads to the bound $k(s; w) \leq \theta^2 s$. This has the benefit of scaling linearly in $s$. But, it gives a sample complexity bound that is no better than the case of standard sparse recovery studied previously. Once more, the question of whether one can choose $w$ in such a way to ensure optimal recovery (scaling linearly in $s$ and at most logarithmically in $d$ and $n$) is currently unresolved.

We conclude with a number of numerical experiments. In Figs. 14-16 we consider Example 3 and employ the orthogonalization strategy of §5. We compare unweighted and weighted $\ell^1$-minimization, where in the latter we set the weights $u = (u_i)_{i \in I}$ to be

$$
  u_i = \| \psi_i \|_{L^\infty(\mathcal{Y})},
$$

where $\mathcal{Y}$ is the basis constructed via the approach of §5. In other words, these weights follow the approach discussed in §6.2 for promoting lower set sparsity.

For sampling, we consider Monte Carlo sampling (69) and the discrete ‘optimal’ measure (70). In all examples, we see that weighted $\ell^1$-minimization substantially outperforms unweighted $\ell^1$-minimization. This is consistent with the observation that the larger polynomial coefficients tend to occur at smaller multi-indices – a property that the weights (74) promote by assigning larger weights to higher multi-indices. In terms of sampling, we observe the sampling measure (70) outperforming Monte Carlo sampling (69), where, as per usual, the benefit tends to lessen in higher dimensions. As discussed above, we do not claim that (70) is an optimal sampling measure in the weighted case: in fact, unlike in the unweighted case, it does not necessarily minimize the corresponding term $k(s; w)$ in the sample complexity bound. Yet, these experiments appear to suggest that it is a useful strategy when combined with weights to further enhance recovery of smooth functions via polynomials.
Towards optimal sampling for learning sparse approximations in high dimensions

\[(d, n, N) = (1,399,400)\]

\[(d, n, N) = (2,152,796)\]

\[(d, n, N) = (8,22,1843)\]

\[(d, n, N) = (16,14,4385)\]

Fig. 14: The relative error \((35)\) versus \(m\) for \(\ell^1\)-minimization and weighted \(\ell^1\)-minimization in the case of Example 3, where \(I = \mathcal{I}^{HC}_{t-1}\) is the hyperbolic cross index set \((12)\) and \(f = f_1\) and \(D = D_2\) are as in \((33)\) and \((34)\), respectively. This figure compares Monte Carlo sampling \((69)\) from the discrete uniform measure (labelled ‘QU’) and sampling from the discrete ‘optimal’ measure \((70)\) (‘QO’).

7 Conclusions and challenges

The purpose of this chapter has been to explore the question of optimal sampling for learning sparse approximations in high dimensions. In the more straightforward setting of Problem 1, we showed how this can be almost entirely resolved by defining a sampling measure (or measures) in terms of the Christoffel function of the corresponding subspace. We remark in passing recent work that strives to go even further, by removing the log factor in the sample complexity bound. See [31] and references therein. We note, however, that such procedures may not be feasible in practice, or may not guarantee quasi-optimal error bounds.

In the more challenging setting of Problem 2, we explored the limitations of Monte Carlo sampling, and showed how to obtain a sampling measure that optimized the sufficient condition of the number of measurements. Empirically, this leads to improved approximation, especially in lower-dimensional problems. Fi-
nally, we discussed structured sparsity, via either weighted or lower set sparsity, both of which can be promoted by using weights. Although here we were not even able to find a sampling measure to optimize the sample complexity bound, we found empirically that the same sampling measure used previously worked well in practice when combined with weights.

The major open problem raised by this work is therefore: is it possible to design sampling measures for sparse approximation in dictionaries that are theoretically optimal, with sample complexity bounds that scale log-linearly in $s$ and logarithmically in $n$? Currently, we have no answer to this question. We note in passing that it may be important to take into account more refined structured of the dictionary. See [90] for recent work that uses the envelope bound (59) to derive improved sample complexity bounds in the case of Example 2 with Monte Carlo sampling. It is also notable that the various constants $\theta$, $\Theta$ and $a, b$ (in the case of irregular domains) often very poorly explain the observed performance. This is particularly notable in the case of $a, b$. This raises the question of a more refined analysis that avoids these terms.

Let us also mention several extensions. First, while this work has focused on standard dictionaries consisting of algebraic or trigonometric polynomials, it is per-
fectly applicable to much more general dictionaries. This includes dictionaries now arising commonly in machine learning settings, such as random feature models or learned dictionaries obtained from deep neural network training. We note recent work on learning sparse representations in random feature models [54]. An interesting question for future work involves applying the techniques considered herein to these models, to obtain better sampling strategies for such dictionaries. This may be highly relevant for applications using machine learning techniques that are data-starved. There is also the problem of combining sampling, via the strategies discussed herein, with learning the dictionary in an adaptive way to boost performance.

Second, we note that the sampling model explored in this work is simple pointwise evaluations. It is possible to extend this to much more general sampling models, many of which occur in practical settings. An example is the discrete-in-space-continuous-in-time model, which can occur when sensors in physical space take continuous recordings a time-dependent function $f(y, t)$. Another problem, which arises commonly in uncertainty quantification (see [13,44,76] and references therein), is the problem where one measures both the function $f(y)$ and its gradient $\nabla f(y)$ simultaneously at a sample point $y$. We anticipate that many of the key results of this work can be extended to substantially more general sampling models.
Acknowledgements

The authors wish to thank Simone Brugiapaglia for useful feedback. The authors acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC) through grant 611675 and the Pacific Institute for the Mathematical Sciences (PIMS).

References

1. B. Adcock. Infinite-dimensional $\ell^1$ minimization and function approximation from pointwise data. Constr. Approx., 45(3):343–390, 2017.
2. B. Adcock. Infinite-dimensional compressed sensing and function interpolation. Found. Comput. Math., 18(3):661–701, 2018.
3. B. Adcock, A. Bao, and S. Brugiapaglia. Correcting for unknown errors in sparse high-dimensional function approximation. Numer. Math., 142(3):667–711, 2019.
4. B. Adcock, A. Bao, J. D. Jakeman, and A. Narayan. Compressed sensing with sparse corruptions: fault-tolerant sparse collocation approximations. SIAM/ASA J. Uncertain. Quantif., 6(4):1424–1453, 2018.
5. B. Adcock and S. Brugiapaglia. Sparse approximation of multivariate functions from small datasets via weighted orthogonal matching pursuit. In S. Sherwin, D. Moxey, J. Peiró, P. Vincent, and C. Schwab, editors, Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2018, volume 134 of Lect. Notes Comput. Sci. Eng., pages 611–621, Cham, Switzerland, 2020. Springer.
6. B. Adcock, S. Brugiapaglia, N. Dexter, and S. Moraga. Deep neural networks are effective at learning high-dimensional Hilbert-valued functions from limited data. In J. Bruna, J. S. Hesthaven, and L. Zdeborová, editors, Proceedings of The Second Annual Conference on Mathematical and Scientific Machine Learning, volume 145 of Proc. Mach. Learn. Res. (PMLR), pages 1–36. PMLR, 2021.
7. B. Adcock, S. Brugiapaglia, N. Dexter, and S. Moraga. On efficient algorithms for computing near-optimal polynomial approximations of smooth, high-dimensional hilbert-valued functions from sample values. In Preparation, 2021.
8. B. Adcock, S. Brugiapaglia, and C. G. Webster. Compressed sensing approaches for polynomial approximation of high-dimensional functions. In H. Boche, G. Caire, R. Calderbank, M. März, G. Kutyniok, and R. Mathar, editors, Compressed Sensing and its Applications: Second International MATHEON Conference 2015, Appl. Numer. Harmon. Anal., pages 93–124. Birkhäuser, Cham, Switzerland, 2017.
9. B. Adcock, S. Brugiapaglia, and C. G. Webster. Sparse Polynomial Approximation of High-Dimensional Functions. Comput. Sci. Eng. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2022.
10. B. Adcock and J. M. Cardenas. Near-optimal sampling strategies for multivariate function approximation on general domains. SIAM J. Math. Data Sci., 2(3):607–630, 2020.
11. B. Adcock and A. C. Hansen. Compressive Imaging: Structure, Sampling, Learning. Cambridge University Press, Cambridge, UK, 2021.
12. B. Adcock and D. Huybrechs. Approximating smooth, multivariate functions on irregular domains. Forum Math. Sigma, 8:e26, 2020.
13. B. Adcock and Y. Sui. Compressive Hermite interpolation: sparse, high-dimensional approximation from gradient-augmented measurements. Constr. Approx., 50:167–207, 2019.
14. N. Alemazkoor and H. Meidani. Divide and conquer: an incremental sparsity promoting compressive sampling approach for polynomial chaos expansions. Comput. Methods Appl. Mech. Engrg., 318:937–956, 2017.
15. N. Alemazkoor and H. Meidani. A near-optimal sampling strategy for sparse recovery of polynomial chaos expansions. *J. Comput. Phys.*, 371:137–151, 2018.

16. B. Arras, M. Bachmayr, and A. Cohen. Sequential sampling for optimal weighted least squares approximations in hierarchical spaces. *SIAM J. Math. Data Sci.*, 1(1):189–207, 2019.

17. J. Bäck, F. Nobile, L. Tamellini, and R. Tempone. Stochastic spectral Galerkin and collocation methods for PDEs with random coefficients: a numerical comparison. In J. S. Hesthaven and E. M. Ronquist, editors, *Spectral and High Order Methods for Partial Differential Equations*, volume 76 of *Lect. Notes Comput. Sci. Eng.*, pages 43–62, Berlin, Heidelberg, Germany, 2011. Springer.

18. G. Blatman and B. Sudret. Adaptive sparse polynomial chaos expansion based on least angle regression. *J. Comput. Phys.*, 230:2345–2367, 2011.

19. J.-L. Bouchot, H. Rauhut, and C. Schwab. Multi-level compressed sensing Petrov-Galerkin discretization of high-dimensional parametric PDEs. *arXiv:1701.01671*, 2017.

20. S. Brugiapaglia, S. Dirksen, H. C. Jung, and H. Rauhut. Sparse recovery in bounded Riesz systems with applications to numerical methods for PDEs. *Appl. Comput. Harmon. Anal.*, 53:231–269, 2021.

21. A. Chkifa, A. Cohen, R. DeVore, and C. Schwab. Sparse adaptive Taylor approximation algorithms for parametric and stochastic elliptic PDEs. *ESAIM. Math. Model. Numer. Anal.*, 47(1):253–280, 2013.

22. A. Chkifa, A. Cohen, G. Migliorati, F. Nobile, and R. Tempone. Discrete least squares polynomial approximation with random evaluations – application to parametric and stochastic elliptic PDEs. *ESAIM. Math. Model. Numer. Anal.*, 49(3):815–837, 2015.

23. A. Chkifa, A. Cohen, and C. Schwab. High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs. *Found. Comput. Math.*, 14(4):601–633, 2014.

24. A. Chkifa, A. Cohen, and C. Schwab. Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs. *J. Math. Pures Appl.*, 103:400–428, 2015.

25. A. Chkifa, N. Dexter, H. Tran, and C. G. Webster. Polynomial approximation via compressed sensing of high-dimensional functions on lower sets. *Math. Comp.*, 87(311):1415–1450, 2018.

26. B. Choi, M. A. Iwen, and F. Krahmer. Sparse harmonic transforms: a new class of sublinear-time algorithms for learning functions of many variables. *Found. Comput. Math.*, 21(2):275–329, 2021.

27. O. Christensen. *An Introduction to Frames and Riesz Bases*. Appl. Numer. Harmon. Anal. Birkhäuser, Basel, Switzerland, 2nd edition, 2016.

28. A. Cohen, M. A. Davenport, and D. Leviatan. On the stability and accuracy of least squares approximations. *Found. Comput. Math.*, 13:819–834, 2013.

29. A. Cohen and R. DeVore. Approximation of high-dimensional parametric PDEs. *Acta Numer.*, 24:1–159, 2015.

30. A. Cohen, R. A. DeVore, and C. Schwab. Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE’s. *Anal. Appl. (Singap)*, 9(1):11–47, 2011.

31. A. Cohen and M. Döbeault. Optimal pointwise sampling for $L^2$ approximation. *arXiv:2105.05545*, 2021.

32. A. Cohen and G. Migliorati. Optimal weighted least-squares methods. *SMAI J. Comput. Math.*, 3:181–203, 2017.

33. A. Cohen and G. Migliorati. Multivariate approximation in downward closed polynomial spaces. In J. Dick, F. Y. Kuo, and H. Woźniakowski, editors, *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, pages 233–282. Springer, Cham, Switzerland, 2018.

34. A. Cohen, G. Migliorati, and F. Nobile. Discrete least-squares approximations over optimized downward closed polynomial spaces in arbitrary dimension. *Constr. Approx.*, 45:497–519, 2017.

35. D. Döng, V. Temlyakov, and T. Ullrich. *Hyperbolic Cross Approximation*. Adv. Courses Math. CRM Barcelona. Birkhäuser, Basel, Switzerland, 2018.
36. N. Dexter, H. Tran, and C. Webster. A mixed $\ell_1$ regularization approach for sparse simultaneous approximation of parameterized PDEs. *ESAIM. Math. Model. Numer. Anal.*, 53:2025–2045, 2019.
37. P. Diaz, A. Doostan, and J. Hampton. Sparse polynomial chaos expansions via compressed sensing and D-optimal design. *Comput. Methods Appl. Mech. Engrg.*, 336:640–666, 2018.
38. M. Dolbeault and A. Cohen. Optimal sampling and Christoffel functions on general domains. *arXiv:2010.11040*, 2020.
39. A. Doostan and H. Owhadi. A non-adapted sparse approximation of PDEs with stochastic inputs. *J. Comput. Phys.*, 230(8):3015–3034, 2011.
40. O. G. Ernst, A. Mugler, H.-J. Starkloff, and E. Ullmann. On the convergence of generalized polynomial chaos expansions. *ESAIM. Math. Model. Numer. Anal.*, 46(2):317–339, Mar. 2012.
41. N. Fajraoui, S. Marelli, and B. Sudret. Sequential design of experiment for sparse polynomial chaos expansions. *SIAM/ASA J. Uncertain. Quantif.*, 5(1):1061–1085, 2017.
42. R. Ghanem, D. Higdon, and H. Owhadi, editors. *Handbook of Uncertainty Quantification*. Springer, Switzerland, 2017.
43. L. Guo, Y. Liu, and L. Yan. Sparse recovery via $\ell_Q$-minimization for polynomial chaos expansions. *Numer. Math. Theor. Meth. Appl.*, 10(4):775–797, 2017.
44. L. Guo, A. Narayan, and T. Zhou. A gradient enhanced $\ell_1$-minimization for sparse approximation of polynomial chaos expansions. *J. Comput. Phys.*, 367:49–64, 2018.
45. L. Guo, A. Narayan, and T. Zhou. Constructing least-squares polynomial approximations. *SIAM Rev.*, 62(2):483–508, 2020.
46. L. Guo, A. Narayan, T. Zhou, and Y. Chen. Stochastic collocation methods via $\ell_q$ minimization using randomized quadratures. *SIAM J. Sci. Comput.*, 39(1):A333–A359, 2017.
47. C. Haberstich, A. Nouy, and G. Perrin. Boosted optimal weighted least-squares. *arXiv:1912.07075*, 2019.
48. M. Hadigol and A. Doostan. Least squares polynomial chaos expansion: a review of sampling strategies. *Comput. Methods Appl. Mech. Engrg.*, 332:382–407, 2018.
49. J. Hampton and A. Doostan. Coherence motivated sampling and convergence analysis of least squares polynomial chaos regression. *Comput. Methods Appl. Mech. Engrg.*, 290:73–97, 2015.
50. J. Hampton and A. Doostan. Compressive sampling of polynomial chaos expansions: convergence analysis and sampling strategies. *J. Comput. Phys.*, 280:363–386, 2015.
51. J. Hampton and A. Doostan. Compressive sampling methods for sparse polynomial chaos expansions. In R. Ghanem, D. Higdon, and H. Owhadi, editors, *Handbook of Uncertainty Quantification*, pages 827–855. Springer, Cham, Switzerland, 2017.
52. J. Hampton and A. Doostan. Basis adaptive sample efficient polynomial chaos (BASE-PC). *J. Comput. Phys.*, 371:20–49, 2018.
53. M. Hansen and C. Schwab. Analytic regularity and nonlinear approximation of a class of parametric semilinear elliptic PDEs. *Math. Nachr.*, 286(8-9):832–860, 2013.
54. A. Hashemi, H. Schaeffer, R. Shi, U. Topcu, G. Tran, and R. Ward. Generalization bounds for sparse random feature expansions. *arXiv:2103.03191*, 2021.
55. L. S. T. Ho, H. Schaeffer, G. Tran, and R. Ward. Recovery guarantees for polynomial coefficients from weakly dependent data with outliers. *J. Approx. Theory*, 259:105472, 2020.
56. V. H. Hoang and C. Schwab. Regularity and generalized polynomial chaos approximation of parametric and random second-order hyperbolic partial differential equations. *Anal. Appl. (Singap.)*, 10(3):295–326, 2012.
57. J. D. Jakeman, M. S. Eldred, and K. Sargsyan. Enhancing $\ell_1$-minimization estimates of polynomial chaos expansions using basis selection. *J. Comput. Phys.*, 289:18–34, 2015.
58. J. D. Jakeman, F. Franzelin, A. Narayan, M. Eldred, and D. Plüger. Polynomial chaos expansions for dependent random variables. *Comput. Methods Appl. Mech. Engrg.*, 351:643–666, 2019.
59. J. D. Jakeman, A. Narayan, and T. Zhou. A generalized sampling and preconditioning scheme for sparse approximation of polynomial chaos expansions. *SIAM J. Sci. Comput.*, 39(3):A1114–A1144, 2017.
Towards optimal sampling for learning sparse approximations in high dimensions

60. I. A. Kougioumtzoglou, I. Petromichelakis, and A. F. Psaros. Sparse representations and compressive sampling approaches in engineering mechanics: a review of theoretical concepts and diverse applications. *Probabilistic Eng. Mech.*, 61:103082, 2020.

61. O. Le Maître and O. M. Knio. *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*. Sci. Comput. Springer, Dordrecht, Netherlands, 2010.

62. Y. Liu and L. Guo. Stochastic collocation via $l_1$-minimisation on low discrepancy point sets with application to uncertainty quantification. *East Asian J. Appl. Math.*, 6(2):171–191, 2016.

63. N. Lüthen, S. Marelli, and B. Sudret. Automatic selection of basis-adaptive sparse polynomial chaos expansions for engineering applications. *arXiv*:2009.04800, 2021.

64. N. Lüthen, S. Marelli, and B. Sudret. Sparse polynomial chaos expansions: literature survey and benchmark. *SIAM/ASA J. Uncertain. Quantif.*, 9(2):593–649, 2021.

65. L. Mathelin and K. A. Gallivan. A compressed sensing approach for partial differential equations with random input data. *Commun. Comput. Phys.*, 12(4):919–954, 2012.

66. G. Migliorati. *Polynomial approximation by means of the random discrete $L^2$ projection and application to inverse problems for PDEs with stochastic data*. PhD thesis, Politecnico di Milano, 2013.

67. G. Migliorati. Adaptive approximation by optimal weighted least squares methods. *SIAM J. Numer. Anal.*, 57(5):2217–2245, 2019.

68. G. Migliorati. Multivariate approximation of functions on irregular domains by weighted least-squares methods. *IMA J. Numer. Anal.*, 41(2):1293–1317, 2021.

69. G. Migliorati and F. Nobile. Analysis of discrete least squares on multivariate polynomial spaces with evaluations in low-discrepancy point sets. *J. Complexity*, 31:517–542, 2015.

70. G. Migliorati, F. Nobile, E. von Schwerin, and R. Tempone. Analysis of the discrete $L^2$ projection on polynomial spaces with random evaluations. *Found. Comput. Math.*, 14:419–456, 2014.

71. A. Narayan. Computation of induced orthogonal polynomial distributions. *Electron. Trans. Numer. Anal.*, 50:71–97, 2018.

72. A. Narayan, J. D. Jakeman, and T. Zhou. A Christoffel function weighted least squares algorithm for collocation approximations. *Math. Comp.*, 86:1913–1947, 2017.

73. A. Narayan and T. Zhou. Stochastic collocation on unstructured multivariate meshes. *Commun. Comput. Phys.*, 18(1):1–36, 2015.

74. L. W.-T. Ng and M. Eldred. Multifidelity uncertainty quantification using nonintrusive polynomial chaos and stochastic collocation. In 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, volume 45. AIAA, 2012.

75. J. Peng, J. Hampton, and A. Doostan. A weighted $\ell_1$-minimization approach for sparse polynomial chaos expansions. *J. Comput. Phys.*, 267:92–111, 2014.

76. J. Peng, J. Hampton, and A. Doostan. On polynomial chaos expansion via gradient-enhanced $\ell_1$-minimization. *J. Comput. Phys.*, 310:440–458, 2016.

77. G. Plonka, D. Potts, G. Steidl, and M. Tasche. *Numerical Fourier Analysis*. Appl. Numer. Harmon. Anal. Birkhäuser, Cham, Switzerland, 2018.

78. H. Rauhut and C. Schwab. Compressive sensing Petrov-Galerkin approximation of high dimensional parametric operator equations. *Math. Comp.*, 86:661–700, 2017.

79. H. Rauhut and R. Ward. Sparse Legendre expansions via $\ell_1$-minimization. *J. Approx. Theory*, 164(5):517–533, 2012.

80. H. Rauhut and R. Ward. Interpolation via weighted $\ell_1$ minimization. *Appl. Comput. Harmon. Anal.*, 40(2):321–351, 2016.

81. Y. Shin and D. Xiu. Correcting data corruption errors for multivariate function approximation. *SIAM J. Sci. Comput.*, 38(4):A2492–A2511, 2016.

82. Y. Shin and D. Xiu. Nonadaptive quasi-optimal points selection for least squares linear regression. *SIAM J. Sci. Comput.*, 38(1):A385–A411, 2016.

83. R. C. Smith. *Uncertainty Quantification: Theory, Implementation, and Applications*. Comput. Sci. Eng. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013.
84. C. Soize and R. Ghanem. Physical systems with random uncertainties: chaos representations with arbitrary probability measure. SIAM J. Sci. Comput., 26(2):395–410, Jan. 2004.
85. T. J. Sullivan. Introduction to Uncertainty Quantification, volume 63 of Texts Appl. Math. Springer, Cham, Switzerland, 2015.
86. G. Tang. Methods for high dimensional uncertainty quantification: regularization, sensitivity analysis, and derivative enhancement. PhD thesis, Stanford University, 2013.
87. G. Tang and G. Iaccarino. Subsampled Gauss quadrature nodes for estimating polynomial chaos expansions. SIAM/ASA J. Uncertain. Quantif., 2(1):423–443, 2014.
88. T. Tang and T. Zhou. On discrete least-squares projection in unbounded domain with random evaluations and its application to parametric uncertainty quantification. SIAM J. Sci. Comput., 36(5):A2272–A2295, 2014.
89. V. Temlyakov. Multivariate approximation, volume 32. Cambridge University Press, 2018.
90. H. Tran and C. Webster. Analysis of sparse recovery for Legendre expansions using envelope bound. arXiv:1810.02926, 2018.
91. H. Tran and C. Webster. A class of null space conditions for sparse recovery via nonconvex, non-separable minimizations. Results Appl. Math., 3:100011, 2019.
92. H. Tran, C. G. Webster, and G. Zhang. Analysis of quasi-optimal polynomial approximations for parameterized PDEs with deterministic and stochastic coefficients. Numer. Math., 137(2):451–493, 2017.
93. J. A. Tropp. User-friendly tail bounds for sums of random matrices. Found. Comput. Math., 12:389–434, 2012.
94. P. Tsilifis, X. Huan, C. Safta, K. Sargsyan, G. Lacaze, J. C. Oefelein, H. N. Najm, and R. G. Ghanem. Compressive sensing adaptation for polynomial chaos expansions. J. Comput. Phys., 380:29–47, 2019.
95. Y. Xu, A. Narayan, H. Tran, and C. Webster. Analysis of the ratio of $\ell_1$ and $\ell_2$ norms in compressed sensing. arXiv:2004.05873, 2020.
96. Z. Xu and T. Zhou. On sparse interpolation and the design of deterministic interpolation points. SIAM J. Sci. Comput., 36(4):1752–1769, 2014.
97. L. Yan, L. Guo, and D. Xiu. Stochastic collocation algorithms using $\ell_1$-minimization. Int. J. Uncertain. Quantif., 2(3):279–293, 2012.
98. L. Yan, Y. Shin, and D. Xiu. Sparse approximation using $\ell_1 – \ell_2$ minimization and its application to stochastic collocation. SIAM J. Sci. Comput., 39(1):A229–A254, 2017.
99. X. Yang and G. E. Karniadakis. Reweighted $\ell_1$ minimization method for stochastic elliptic differential equations. J. Comput. Phys., 248:87–108, 2013.
100. X. Yang, H. Lei, N. A. Baker, and G. Lin. Enhancing sparsity of Hermite polynomial expansions by iterative rotations. J. Comput. Phys., 307:94–109, 2016.
101. X. Yang, W. Li, and A. Tartakovsky. Sliced-inverse-regression–aided rotated compressive sensing method for uncertainty quantification. SIAM/ASA J. Uncertain. Quantif., 6(4):1532–1554, 2018.
102. X. Yang, X. Wan, L. Lin, and H. Lei. A general framework for enhancing sparsity of generalized polynomial chaos expansions. Int. J. Uncertain. Quantif., 9(3):221–243, 2019.
103. S. Zein, B. Colson, and P. Glineur. An efficient sampling method for regression-based polynomial chaos expansion. Commun. Comput. Phys., 13(4):1173–1188, 2013.
104. T. Zhou, A. Narayan, and D. Xiu. Weighted discrete least-squares polynomial approximation using randomized quadratures. J. Comput. Phys., 298:787–800, 2015.
105. T. Zhou, A. Narayan, and Z. Xu. Multivariate discrete least-squares approximations with a new type of collocation grid. SIAM J. Sci. Comput., 36(5):A2401–A2422, 2014.