A Unified Framework for Learning with Nonlinear Model Classes from Arbitrary Linear Samples

Ben Adcock [1] **Juan M. Cardenas** [2] **Nick Dexter** [3]

Abstract

This work considers the fundamental problem of learning an unknown object from training data using a given model class. We introduce a framework that allows for objects in arbitrary Hilbert spaces, general types of (random) linear measurements as training data and general types of nonlinear model classes. We establish a series of learning guarantees for this framework, which provide explicit relations between the amount of training data and the model class to ensure near-best generalization bounds. In doing so, we introduce the key notion of the *variation* of a model class with respect to a distribution of sampling operators. We show that this framework can accommodate many different types of well-known problems of interest, such as matrix sketching by random sampling, compressed sensing with isotropic vectors, active learning in regression and compressed sensing with generative models. In all cases, known results become straightforward corollaries of our general theory. Hence, this work provides a powerful framework for studying and analyzing many different types of learning problems.

1. Introduction

Learning an unknown object from a finite set of training data is a fundamental problem in computer science. Typically, in modern settings, one seeks to learn an approximate representation in a nonlinear model class (also known as an approximation space or hypothesis set). It is also common to generate the training data randomly according to some distribution. Of critical importance in this endeavour is the question of learning guarantees. In other words: *how much* ***training data*** *suffices to ensure* ***good generalization****, and how is this influenced by the choice of* ***model class*** *and the* ***random process generating the training data****?*

This question has often been addressed for specific types of training data. For instance, in the case of regression, the training data consists of pointwise evaluations of some target function, or in the case of computational imaging, the training data may consist of samples of the Fourier or Radon transform of some target image. It is also commonly studied for specific model classes, e.g., polynomial spaces (Adcock et al., 2022c) or spaces of sparse vectors (Foucart & Rauhut, 2013), spaces of low-rank matrices or tensors (Davenport & Romberg, 2016), spaces defined by generative models (Bora et al., 2017), single- (Gajjar et al., 2023) or multilayer neural networks (Adcock & Dexter, 2021; Adcock et al., 2021; 2022b), Fourier sparse functions (Erdelyi et al., 2020), (sparse) random feature models (Avron et al., 2017; Hashemi et al., 2023) and many more.

In this paper, we introduce a unified framework for learning with nonlinear model classes from arbitrary linear samples. The main features of this framework are:

(i) the object is an element of a separable Hilbert space;
(ii) each measurement is taken randomly and independently according to a random linear operator;
(iii) the measurements may be scalar- or vector-valued, or, in general, may take values in a Hilbert space;
(iv) the measurements can be *multimodal*, i.e., generated by different distributions of random linear operators, as long as a certain *nondegeneracy* condition holds;
(v) the model class can be linear or nonlinear, but should be contained in (or covered by) a union of finite-dimensional subspaces;
(vi) the resulting learning guarantees are given in terms of the *variation* of the model class with respect to the sampling distributions.

We present a series of examples to highlight the generality of this framework. In various cases, our learning guarantees either include or improve known results.

[1]Department of Mathematics, Simon Fraser University, Burnaby, BC, Canada [2]Ann and H. J. Smead Department of Aerospace Engineering Sciences, University of Colorado Boulder, Boulder, Colorado, USA [3]Department of Scientific Computing, Florida State University, Tallahassee, Florida, USA. Correspondence to: Nick Dexter <nick.dexter@fsu.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1.1. The framework

The setup we consider in this paper is the following.

- $\mathbb{X}$ is a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathbb{X}}$ and norm $\|\cdot\|_{\mathbb{X}}$.
- $\mathbb{X}_0 \subseteq \mathbb{X}$ is a semi-normed vector space, termed the *object space*, with semi-norm $\|\cdot\|_{\mathbb{X}_0}$.
- $x \in \mathbb{X}_0$ is the unknown target object.
- For each $i = 1, \dots, m$, $\mathbb{Y}_i$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathbb{Y}_i}$ and norm $\|\cdot\|_{\mathbb{Y}_i}$ termed the ith *measurement space*.
- For each $i = 1, \dots, m$, $\mathcal{A}_i$ is a distribution of bounded linear operators $(\mathbb{X}_0, \|\cdot\|_{\mathbb{X}_0}) \to (\mathbb{Y}_i, \|\cdot\|_{\mathbb{Y}_i})$. We term $A_i \sim \mathcal{A}_i$ the ith *sampling operator*. We also write $\mathcal{B}(\mathbb{X}_0, \mathbb{Y}_i)$ for the space of bounded linear operators, so that $A_i \in \mathcal{B}(\mathbb{X}_0, \mathbb{Y}_i)$.
- We assume that the family $\{\mathcal{A}_i\}_{i=1}^m$ of distributions is *nondegenerate*. Namely, there exist $0 < \alpha \leq \beta < \infty$ such that, for all $x \in \mathbb{X}_0$,

$$\alpha \|x\|_{\mathbb{X}}^2 \leq \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{A_i \sim \mathcal{A}_i} \|A_i(x)\|_{\mathbb{Y}_i}^2 \leq \beta \|x\|_{\mathbb{X}}^2. \quad (1.1)$$

- $\mathbb{U} \subseteq \mathbb{X}_0$ is a set, termed the *model class*. Our aim is to learn an approximation $x \approx \hat{x} \in \mathbb{U}$.

Now let $A_i \in \mathcal{B}(\mathbb{X}_0, \mathbb{Y}_i)$, $i = 1, \dots, m$, be independent realizations from the distributions $\mathcal{A}_1, \dots, \mathcal{A}_m$. We consider (noisy) training data of the form

$$\{(A_i, b_i := A_i(x) + e_i)\}_{i=1}^m, \quad (1.2)$$

where $(A_i, b_i) \in \mathcal{B}(\mathbb{X}_0, \mathbb{Y}_i) \times \mathbb{Y}_i$. Our aim is to learn x from this data, and we do this via empirical least squares, i.e.,

$$\hat{x} \in \operatorname*{argmin}_{u \in \mathbb{U}} \frac{1}{m} \sum_{i=1}^m \|b_i - A_i(u)\|_{\mathbb{Y}_i}^2. \quad (1.3)$$

Later, we also allow for $\hat{x}$ to be an approximate minimizer, to model the more practical scenario where the minimization problem is solved inexactly. Note that in this work we consider the *agnostic learning* setting, where $x \notin \mathbb{U}$ and the noise e_i can be adversarial (but small in norm).

Before describing our main theoretical contributions, it is worth showing how our framework includes the standard function regression problem as a special case.

Example 1.1 (Function regression from i.i.d. samples). Let $D \subseteq \mathbb{R}^d$ be a domain with a measure ρ and consider the problem of learning an unknown function $f \in L^2_\rho(D)$ from data $\{(z_i, f(z_i))\}_{i=1}^m$, where $z_i \sim_{\text{i.i.d.}} \mu$ for some probability measure μ on D. To cast this problem in this framework, we make the (mild) assumption that μ is absolutely continuous with respect to ρ and $\nu := \mathrm{d}\mu / \mathrm{d}\rho > 0$ a.e.. Now let $\mathbb{X} = L^2_\rho(D)$, $\mathbb{X}_0 = C(\overline{D})$ (here $\overline{D}$ denotes the closure of D), $\mathbb{Y}_i = \mathbb{Y} = \mathbb{R}$, $\forall i$, with the Euclidean inner product and $\mathcal{A}_i = \mathcal{A}$, $\forall i$, be defined by $A \sim \mathcal{A}$ if $A(f) = f(z)/\sqrt{\nu(z)}$ for $z \sim \mu$. A short calculation shows that nondegeneracy (1.1) holds with $\alpha = \beta = 1$ in this case. Now, given an approximation space $\mathbb{U} \subseteq C(\overline{D})$, the least-squares problem (1.3) becomes the (nonlinear) weighted-least squares fit

$$\hat{f} \in \operatorname*{argmin}_{u \in \mathbb{U}} \frac{1}{m} \sum_{i=1}^m \frac{1}{\nu(z_i)} |f(z_i) + \sqrt{\nu(z_i)} e_i - u(z_i)|^2. \quad (1.4)$$

Note that it is common to set $\mu = \rho$ in such problems, in which case $\nu = 1$ and (1.4) is an unweighted least-squares fit. However, this more general setup allows one to consider the active learning setting, where the sampling measure μ is chosen judiciously in term of $\mathbb{U}$ to improve the learning performance of $\hat{f}$. We discuss this in §4.

As we explain in §2, this framework also contains many other problems of interest as special cases. In many of these cases, as well as in Example 1.1 above, the training data is sampled from the same distribution, i.e., $\mathcal{A}_i = \mathcal{A}$, $\forall i$. However, having different distributions allows us to consider *multimodal* sampling problems, where data is obtained from different random processes. In §2 we also discuss situations where this arises.

1.2. Contributions

Besides the general framework described above, our main contributions are a series of learning guarantees that relate the amount of training data m to properties of the sampling distributions $\{\mathcal{A}_i\}_{i=1}^m$. We now present a simplified version of our main result covering the case where $\mathcal{A}_i = \mathcal{A}$, $\forall i$. The full case is presented in §3.

A key quantity in this analysis is the so-called *variation* of a (nonlinear) set $\mathbb{V} \subseteq \mathbb{X}_0$ with respect to a distribution $\mathcal{A}$ of bounded linear operators in $\mathcal{B}(\mathbb{X}_0, \mathbb{Y})$. We define this as the smallest constant $\Phi = \Phi(\mathbb{V}; \mathcal{A})$ such that

$$\|A(v)\|_{\mathbb{Y}}^2 \leq \Phi, \ \forall v \in \mathbb{V}, \quad \text{a.s. } A \sim \mathcal{A}. \quad (1.5)$$

See also Definition 3.1. As we discuss in Example 3.2, the variation is effectively a generalization of the notion of *coherence* in classical compressed sensing (see, e.g., (Candès & Plan, 2011) or (Adcock & Hansen, 2021, Chpt. 5)). It also generalizes various generalized notions of coherence, such as the 'local coherence' of (Krahmer & Ward, 2013), the 'local coherence in levels' of (Adcock & Hansen, 2021) and the 'block coherence' of (Bigot et al., 2016). Moreover, as we discuss in Appendix A, it is directly related to the leverage score function in the case of matrix sketching. In the following, we also define $S(\mathbb{V}) = \{v \in \mathbb{V} : \|v\|_{\mathbb{X}} = 1\}$ as the unit sphere of $\mathbb{V}$ in $\mathbb{X}$.

Simplified Result. Consider the setup of §1.1 with $\mathcal{A}_i = \mathcal{A}$, $\forall i$, let $\mathbb{U} \subseteq \mathbb{X}_0$ and suppose that $\mathbb{U}' := \mathbb{U} - \mathbb{U}$ is such that

(i) $\mathbb{U}'$ is a cone (i.e., $tu' \in \mathbb{U}'$ for $t \geq 0$ and $u' \in \mathbb{U}'$), and
(ii) $\mathbb{U}' \subseteq \mathbb{V}_1 \cup \cdots \cup \mathbb{V}_d =: \mathbb{V}$, where each $\mathbb{V}_i \subseteq \mathbb{X}_0$ is a subspace of dimension at most n.

Suppose that, for some $0 < \epsilon < 1$, either

(a) $m \gtrsim \alpha^{-1} \cdot \Phi(S(\mathbb{U}' - \mathbb{U}'); \mathcal{A}) \cdot [\log(2d/\epsilon) + n]$, or
(b) $m \gtrsim \alpha^{-1} \cdot \Phi(S(\mathbb{V}); \mathcal{A}) \cdot \log(2nd/\epsilon)$.

Let $x \in \mathbb{X}_0$, $\theta \geq \|x\|_{\mathbb{X}}$ and $\check{x} = \min\{1, \theta/\|\hat{x}\|_{\mathbb{X}}\}\hat{x}$ for any minimizer $\hat{x}$ of (1.3) with noisy measurements (1.2). If $N = \frac{1}{m}\sum_{i=1}^m \|e_i\|_{\mathbb{Y}_i}^2$, then

$$\mathbb{E}\|x - \check{x}\|_{\mathbb{X}}^2 \lesssim \frac{\beta}{\alpha} \inf_{u \in \mathbb{U}} \|x - u\|_{\mathbb{X}}^2 + \theta^2\epsilon + \frac{N}{\alpha}. \tag{1.6}$$

A few remarks are in order. First, we note that (i) is a standard assumption, and holds in many cases of interest. Assumption (ii) holds for many sparse or 'structured' sparse model classes (e.g., joint or block sparse vectors, sparse in levels vectors, tree sparse vectors, cosparse vectors and so forth). As we discuss in §5, it also holds for model classes defined by generative models. Conditions (a) and (b) are slightly different, in that they involve the variation over different sets. In particular, condition (b) evaluates the variation over $S(\mathbb{V})$, and, by doing so, obtains a better scaling with respect to n than (a)-(b). Finally, we note that this result is a simplified version of the results presented in §3. In the full results, we also consider inexact minimizers of (1.3). Moreover, we provide several refinements of conditions (a) and (b), as well as a variation involving an additional assumption on $\mathbb{U}'$ which leads to a sharper bound. This bound is particularly relevant to sparse and structured sparse model classes.

Having presented these in §3, we next describe how this framework unifies, generalizes and, in some cases, improves known results. In particular, we recover known bounds for *compressed sensing with isotropic vectors* and for *matrix sketching by random sampling* (see Appendices A and B, respectively). We then discuss the application of this framework to two different problems.

(a) *Active learning in regression (§4).* Here, we extend and improve the recent work (Adcock et al., 2023) by applying our main results to derive a random (importance) sampling strategy based on a certain *Christoffel function* (also known as the *leverage score function*). We show that this strategy, termed *Christoffel sampling*, leads to near-optimal sample complexity bounds in a number of important settings.

(b) *Compressed sensing with generative models (§5).* Here we extend and improve results from the recent work (Berk et al., 2023a) by obtaining learning guarantees for general types of measurements in the case where $\mathbb{U}$ is the range of a ReLU generative model. Subsequently, we applied these results to subsampled unitary matrices, as in (Adcock et al., 2023; Berk et al., 2023b), and use them to derive an optimal sampling strategy.

1.3. Related work

Our sampling framework is inspired by previous work in compressed sensing, notably compressed sensing with *isotropic vectors* (Candès & Plan, 2011) (see also (Adcock & Hansen, 2021, Chpt. 12) and Example 2.3). This work considers the case where $\mathbb{X}_0 = \mathbb{C}^N$, $\mathbb{Y}_1 = \cdots = \mathbb{Y}_m = \mathbb{C}$ and $\mathbb{U}$ is the set of s-sparse vectors, i.e., the target object is an (approximately) sparse vector and the sampling operators are linear functionals. Note that isotropy would correspond to the case $\alpha = \beta = 1$ in (1.1). We relax this to allow $\alpha \neq \beta$. Within the compressed sensing literature, there are a number of works that allow for non-scalar valued measurements. See (Bigot et al., 2016; Boyer et al., 2019) for an instance of vector-valued measurements ('block sampling') and (Traonmilin et al., 2017) as well as (Adcock et al., 2022a; Dexter et al., 2019) for Hilbert-valued measurements. The latter arise in an array of important applications in computational science, such as parametric Differential Equations (DEs) in computational Uncertainty Quantification (UQ), see (Adcock et al., 2022c, Section 1.2.2).

Recovery guarantees in compressed sensing are generally derived for specific model classes, such as sparse vectors or various generalizations (e.g., the aforementioned joint or block sparse vectors, sparse in levels vectors, tree sparse vectors, cosparse vectors and so forth (Adcock et al., 2017; Baraniuk et al., 2010; Bourrier et al., 2014; Davenport et al., 2012; Duarte & Eldar, 2011; Traonmilin & Gribonval, 2018)). Guarantees for general model classes are usually only presented in the case of (sub)Gaussian random measurements (see e.g., (Baraniuk et al., 2010; Dirksen, 2016)). These, while mathematically elegant, are typically not useful in practice. Our framework provides a unified set of recovery guarantees for very general types of measurements. It contains subsampled unitary matrices (a well-known measurement modality in compressed sensing, with practical relevance – see Example 2.4) as a special case, but also many others, including vector-valued measurements.

Active learning is a large topic. For function regression in linear spaces, a now well-known solution involves sampling according to the *Christoffel function* (Cohen & Migliorati, 2017) or *leverage score function* (Avron et al., 2017; Chen et al., 2016; Chen & Price, 2019; Derezinski et al., 2018; Erdelyi et al., 2020; Gajjar et al., 2023; Ma et al., 2015) of the subspace. A number of these works have extended this to certain nonlinear spaces, such as Fourier sparse functions (Erdelyi et al., 2020) and single-neuron models (Gajjar et al., 2023). In the work (Adcock et al., 2023), this was extended to more general nonlinear spaces and other types of measurements. As noted above, this work improves the

theoretical guarantees in (Adcock et al., 2023). In particular, we show the usefulness of Christoffel sampling for more general types of nonlinear model classes.

Compressed sensing with generative models was introduced in (Bora et al., 2017), and has proved useful in image reconstruction tasks such as Magnetic Resonance Imaging (MRI) (see (Jalal et al., 2021) and references therein). Initial learning guarantees for generative models involved (sub)Gaussian random measurements (Bora et al., 2017). Guarantees for randomly subsampled unitary matrices were established in (Berk et al., 2023a) for ReLU generative networks, and later extended in (Berk et al., 2023b) for the nonuniformly subsampled case. As noted above, our work refines and further generalizes this analysis to more general types of measurements.

2. Further examples

Having considered function regression Example 1.1, we now present a series of further examples that can be cast into our framework. We will return to these examples later after first stating our main learning guarantees in §3.

Example 2.1 (Matrix sketching for large least-squares problems). Let $X \in \mathbb{C}^{N\times n}$, $N \geq n$, be a given matrix and $y \in \mathbb{C}^N$ a target vector. In many applications, it may be infeasible (due to computational constraints) to find a solution to the 'full' least-squares problem $w \in \operatorname{argmin}_{z\in\mathbb{C}^n} \|Xz - x\|_{\ell^2}$. Therefore, one aims to instead find a *sketching matrix* $S \in \mathbb{C}^{m\times N}$ (a matrix with one nonzero per row) such that any minimizer of the sketched problem

$$\hat{w} \in \operatorname*{argmin}_{z\in\mathbb{C}^n} \|SXz - Sx\|_{\ell^2} \tag{2.1}$$

satisfies

$$\|X\hat{w} - y\|_{\ell^2}^2 \lesssim \|Xw - x\|_{\ell^2}^2. \tag{2.2}$$

A particularly effective way to do this involves constructing a random sketch. Formally, let $\pi = \{\pi_1, \dots, \pi_N\}$ be a discrete probability distribution on $\{1, \dots, N\}$ with $\pi_i > 0$ for all i. Then we draw $j_1, \dots, j_m \sim_{\text{i.i.d.}} \pi$ and set $S_{ij_i} = 1/\sqrt{\pi_{j_i}}$ and $S_{ij} = 0$ otherwise. Therefore, $SX \in \mathbb{C}^{m\times n}$ consists of m rows of X scaled by the probabilities $1/\sqrt{\pi_i}$. See (Malik et al., 2022; Woodruff, 2014) for further discussions on matrix sketching.

To cast this into the above framework, let $\mathbb{X} = \mathbb{X}_0 = \mathbb{C}^N$ and $\mathbb{Y} = \mathbb{C}$, both equipped with the Euclidean norm. Note that bounded linear operators $\mathbb{X}_0 \to \mathbb{Y}$ are equivalent to column vectors $a \in \mathbb{C}^N$ (via the relation $x \mapsto a^*x$). Hence we define $\mathcal{A}_i = \mathcal{A}$, $\forall i$, such that $a \sim \mathcal{A}$ if

$$\mathbb{P}(a = e_i/\sqrt{\pi_i}) = \pi_i, \quad i = 1, \dots, N.$$

Observe that (1.1) holds with $\alpha = \beta = 1$. If

$$\mathbb{U} = \{Xz : z \in \mathbb{C}^n\} \tag{2.3}$$

then we readily see that (1.3) (with $b_i = (Sx)_i$) is equivalent to (2.1) in the sense that $\hat{x} = X\hat{w}$ is a solution of (1.3) if and only if $\hat{w}$ is a solution of (2.1). In particular, $\|X\hat{w} - x\|_{\ell^2}^2 = \|\hat{x} - x\|_{\ell^2}^2$ is precisely the $\mathbb{X}$-norm error of the estimator $\hat{x}$.

Leverage score sampling is a near-optimal solution to the matrix sketching problem (Woodruff, 2014). Here, one sets

$$\pi_i = \tau(X)(i)/n, \quad i = 1, \dots, N, \tag{2.4}$$

where

$$\tau(X)(i) = \max_{\substack{z\in\mathbb{C}^n \\ Xz\neq 0}} \frac{|(Xz)_i|^2}{\|Xz\|_2^2}, \quad i = 1, \dots, N, \tag{2.5}$$

are the so-called *leverage scores* of the matrix X. In this case, (2.2) holds with high probability, provided

$$m \gtrsim n \cdot \log(2n/\epsilon). \tag{2.6}$$

In Appendix A, we show this bound is straightforward consequence of our general theory. Thus leverage score sampling is a specific case of our unified framework.

Example 2.2 (Function regression with vector-valued measurements). Regression problems in various applications call for learning vector- as opposed to scalar-valued functions. These are readily incorporated into this framework by modifying Example 1.1.

Hilbert-valued functions. Let $\mathbb{V}$ be a Hilbert space and consider the Hilbert-valued function $f : D \to \mathbb{V}$. As mentioned in §1.3, the problem of learning Hilbert-valued functions arises in various applications, including parametric DEs in computational UQ. Here, f represents the parameters-to-solution map of a DE involving parameters $x \in D$, which, for each $x \in D$, yields the Hilbert-valued output $f(x) \in D$ being the solution of the DE at those parameter values (see (Adcock et al., 2022c; Cohen & DeVore, 2015; Dexter et al., 2019) and references therein for discussion). To cast this problem into the general framework, we modify Example 1.1 by letting $\mathbb{X} = L^2_\rho(D;\mathbb{V})$ be the Bochner space of strongly ρ-measurable functions $D \to \mathbb{V}$, $\mathbb{X}_0 = C(\overline{D};\mathbb{V})$ be the space of continuous, $\mathbb{V}$-valued functions, $\mathbb{Y}_1 = \cdots = \mathbb{Y}_m = \mathbb{V}$ and $\mathcal{A}$ be the distribution of bounded linear operators $\mathbb{X}_0 \to \mathbb{V}$ with $A \sim \mathcal{A}$ if $A(f) = f(z)/\sqrt{\nu(z)} \in \mathbb{V}$ for $f \in \mathbb{X}_0$ and $z \sim \mu$.

Continuous-in-time sampling. Consider $f : D \times [0,T] \to \mathbb{R}$ depending on a spatial variable $x \in D$ and a time variable $t \in [0,T]$. In some examples, for instance, seismology, one assumes continuous (or dense) sampling in time with discrete (i.e., sparse) sampling in space. Thus, each measurement takes the form $\{f(z,t) : 0 \leq t \leq T\}$ for fixed $z \in D$. We can cast this problem into a Hilbert-valued function approximation problem by letting $\mathbb{V} = L^2_\sigma(0,T)$ so that $\mathbb{X} = L^2_\rho(D;\mathbb{V}) \cong L^2_{\rho\times\sigma}(D \times [0,T])$. Hence, continuous-in-time sampling is also covered by the general framework.

Gradient-augmented data. In this problem, each sample takes the form $(z, f(z), \nabla_z f(z))$, i.e., we obtain both the function and its gradient with respect at each sample point z. This problem arises in a number of applications, including parametric DEs and UQ (Alekseev et al., 2011; Peng et al., 2016; O'Leary-Roseberry et al., 2022; 2024), seismology and Physics-Informed Neural Networks (PINNs) for PDEs (Feng & Zeng, 2022; Yu et al., 2022) and deep learning (Czarnecki et al., 2017). Assuming the setup of Example 1.1, we cast this problem into the framework by letting $\mathbb{X}$ be the Sobolev space $H^1_\rho(D)$, $\mathbb{X}_0 = C^1(\overline{D})$, $\mathbb{Y}_i = \mathbb{Y} = \mathbb{V}^{d+1}$, $\forall i$, and defining $A \sim \mathcal{A}$ if $A(f) = (f(z), \nabla_z f(z))/\sqrt{\nu(z)} \in \mathbb{R}^{d+1}$ for $f \in \mathbb{X}_0$ and $z \sim \mu$.

In the next several examples, we show how this framework includes as special cases various general sampling models from the compressed sensing literature.

Example 2.3 (Compressed sensing with isotropic vectors). Classical compressed sensing concerns learning a sparse approximation to an unknown vector $f \in \mathbb{C}^N$ from m linear measurements. A well-known model involves sampling with *isotropic vectors* (Candès & Plan, 2011) (see also (Adcock & Hansen, 2021, Chpt. 11)). We can cast this in the above framework as follows. Let $\mathbb{X} = \mathbb{X}_0 = \mathbb{C}^N$ equipped with Euclidean inner product and $\mathbb{Y}_i = \mathbb{Y} = \mathbb{C}$, $\forall i$. As in Example 2.1, we consider $\mathcal{A}_i = \mathcal{A}$ to be a distribution of vectors in $\mathbb{C}^N$ that are *isotropic*, i.e.,

$$\mathbb{E}_{a\sim\mathcal{A}} aa^* = I. \tag{2.7}$$

Note that (1.1) holds with $\alpha = \beta = 1$ in this case. Moreover, the measurements (1.2) have the form

$$b_i = a_i^* x + e_i \in \mathbb{C}, \quad i = 1, \dots, m,$$

where $a_1, \dots, a_m \sim_{\text{i.i.d.}} \mathcal{A}$. In matrix-vector notation, we can rewrite this as

$$b = Ax + e, \quad \text{where } b = m^{-1/2}(b_i)_{i=1}^m \in \mathbb{C}^m, \tag{2.8}$$

$e = m^{-1/2}(e_i)_{i=1}^m \in \mathbb{C}^m$ and $A \in \mathbb{C}^{m\times N}$ has ith row a_i^*.

As discussed in (Candès & Plan, 2011), this model includes not only the well-known case of subgaussian random matrices, in which case $\mathcal{A}$ is a distribution of subgaussian random vectors, but also many other common sampling models used in signal and image processing applications. Moreover, if we slightly relax (2.7) to $\alpha I \preceq \mathbb{E}_{a\sim\mathcal{A}} aa^* \preceq \beta I$, so that (1.1) holds with the same values of α and β, then it also generalizes the bounded Riesz system model studied in (Brugiapaglia et al., 2021).

Example 2.4 (Compressed sensing with subsampled unitary matrices). A particular case of interest within the previous example is the class of *subsampled unitary* matrices. Let $U \in \mathbb{C}^{N\times N}$ be unitary, i.e., $U^*U = I$. Let $u_i = U^* e_i$, where e_i is the ith canonical basis vector, and $\pi = (\pi_1, \dots, \pi_N)$ be a discrete probability distribution on $\{1, \dots, N\}$ with $\pi_i > 0$, $\forall i$. Then we define the (discrete) distribution of vectors $\mathcal{A}$ by $a \sim \mathcal{A}$ if

$$\mathbb{P}(a = u_i/\sqrt{\pi_i}) = \pi_i, \quad i = 1, \dots, N.$$

It is readily checked that (2.7) holds in this case, making this family isotropic. The corresponding matrix A consists of (scaled) rows of U sampled with probability π.

Subsampled unitary matrices occur in various applications. For example, U may be the matrix of the Discrete Fourier Transform (DFT) in a Fourier sensing problem. This arises in applications such as MRI, NMR, Helium Atom Scattering and radio interferometry (see, e.g., (Adcock & Hansen, 2021) and references therein).

The reader will notice that in the previous examples, the distributions $\mathcal{A}_i$ were all equal. We now conclude with an example that motivates different distributions.

Example 2.5 (Multimodal data). Consider the general setup of §1.1. Rather than a single distribution $\mathcal{A}$ generating all the data, we now assume that there are $C > 1$ different types of data, with the cth type generated via a distribution $\mathcal{A}^{(c)}$, $c = 1, \dots, C$, of bounded linear operators in $\mathcal{B}(\mathbb{X}_0, \mathbb{Y}^{(c)})$. Let $m = m_1 + \cdots + m_C$ and define $\{\mathcal{A}_i\}_{i=1}^m$ by

$$\mathcal{A}_i = \mathcal{A}^{(c)} \quad \text{if } m_1 + \cdots + m_{c-1} < i \leq m_1 + \cdots + m_c.$$

Thus, the first m_1 samples are generated by $\mathcal{A}^{(1)}$, the next m_2 samples by $\mathcal{A}^{(2)}$, and so forth. Notice that nondegeneracy (1.1) is now equivalent to the condition

$$\alpha\|x\|_{\mathbb{X}}^2 \leq \sum_{c=1}^{C} \frac{m_c}{m} \mathbb{E}_{A\sim\mathcal{A}^{(c)}} \|A(x)\|_{\mathbb{Y}^{(c)}}^2 \leq \beta\|x\|_{\mathbb{X}}^2.$$

Multimodal data is important in many applications. It was previously considered in (Adcock et al., 2023), which, as noted in §1 is a special case of this work. As observed in (Adcock et al., 2023), multimodal data arises in various applications, such as multi-sensor imaging systems (Chun & Adcock, 2017) and PINNs for PDEs (Han et al., 2018; Raissi et al., 2019). This includes the important case of *parallel* MRI (McRobbie et al., 2006), which is used widely in medical practice. Another application involves an extension of the gradient-augmented learning problem in Example 2.2 where, due to cost or other constraints, one can only afford to measure gradients at some fraction of the total samples. See (Adcock et al., 2023, §B.7) for further details.

3. Learning guarantees

In this section, we present our main theoretical results. In these results, we consider approximate minimizers of (1.3). We say that $\hat{x} \in \mathbb{U}$ is a ζ-*minimizer* of (1.3) if

$$\frac{1}{m}\sum_{i=1}^m \|b_i - A_i(\hat{x})\|_{\mathbb{Y}_i}^2 \leq \min_{u\in\mathbb{U}} \frac{1}{m}\sum_{i=1}^m \|b_i - A_i(u)\|_{\mathbb{Y}_i}^2 + \zeta.$$

Also, given a set $\mathbb{V} \subseteq \mathbb{X}$, we define

$$S(\mathbb{V}) = \{v/\|v\|_{\mathbb{X}} : v \in \mathbb{V}\backslash\{0\}\}. \tag{3.1}$$

We also say that $\mathbb{V}$ is a *cone* if $tv \in \mathbb{V}$ for any $t \geq 0$, $v \in \mathbb{V}$. Notice that $S(\mathbb{V}) = \{v \in \mathbb{V} : \|v\|_{\mathbb{X}} = 1\}$ in this case.

3.1. Variation

We now formally introduce the concept of variation, which is crucial to our analysis.

Definition 3.1 (Variation with respect to a distribution)**.** Let $\mathbb{V} \subseteq \mathbb{X}_0$ and $\mathbb{Y}$ be a Hilbert space. Consider a distribution $\mathcal{A}$ of bounded linear operators in $\mathcal{B}(\mathbb{X}_0, \mathbb{Y})$. The *variation of* $\mathbb{V}$ *with respect to* $\mathcal{A}$ is the smallest constant $\Phi = \Phi(\mathbb{V}; \mathcal{A}) \in [0, \infty]$ such that

$$\|A(v)\|_{\mathbb{Y}}^2 \leq \Phi, \ \forall v \in \mathbb{V}, \quad \text{a.s. } A \sim \mathcal{A}. \tag{3.2}$$

Note that we specify Φ as the smallest constant to ensure that it is well defined. However, in all our results Φ can be *any* constant such that (3.2) holds. This is relevant in our examples, since it means we only need to derive upper bounds for the variation.

In the following example, we show how the variation extends to the classical notion of *coherence* in compressed sensing. Later, in the context of active learning in §4, we show that it also relates to the *Christoffel function* (also the *leverage score function*).

Example 3.2 (Classical compressed sensing and coherence). Coherence is a well-known concept in compressed sensing (see, e.g., (Candès & Plan, 2011)). Consider the setting of Example 2.3 and let $1 \leq s \leq N$. Classical compressed sensing considers the model class of s-sparse vectors

$$\mathbb{U} = \Sigma_s = \{x \in \mathbb{C}^N : x \text{ is } s\text{-sparse}\}. \tag{3.3}$$

In this case, $\Phi(S(\mathbb{U}); \mathcal{A})$ is the smallest constant such that

$$|a^* v|^2 \leq \Phi \|v\|_{\ell^2}^2, \quad \forall v \ s\text{-sparse}, \ \text{a.s. } a \sim \mathcal{A}.$$

Define the *coherence* $\mu = \mu(\mathcal{A})$ of the distribution $\mathcal{A}$ (see (Candès & Plan, 2011) or (Adcock & Hansen, 2021, Defn. 11.16)) as the smallest constant such that

$$\|a\|_{\ell^\infty}^2 \leq \mu(\mathcal{A}), \quad \text{a.s. } a \sim \mathcal{A}.$$

Then a short derivation gives that

$$\Phi(S(\Sigma_s); \mathcal{A}) \leq \mu(\mathcal{A}) s. \tag{3.4}$$

Thus, the variation is bounded by the coherence $\mu(\mathcal{A})$ multiplied by the sparsity s.

Coherence directly determines the learning guarantee for sparse vector recovery. A well-known measurement condition (see, e.g., (Adcock & Hansen, 2021, Cor. 13.15)) takes the form

$$m \gtrsim \mu(\mathcal{A}) \cdot s \cdot \left(\log^2(s) \log(N) + \log(\epsilon^{-1})\right). \tag{3.5}$$

In other words, the number of measurements that suffices for good generalization scales linearly in s, up to the coherence $\mu(\mathcal{A})$ and a polylogarithmic factor. In Appendix B we show that this bound is a straightforward corollary of our main theorems and (3.4) when applied to this problem. Hence our framework generalizes classical compressed sensing with isotropic vectors.

As described in §1.1, in this work we consider a collection of distributions, which we henceforth denote as $\bar{\mathcal{A}} = \{\mathcal{A}_i\}_{i=1}^m$. We define the variation of this collection as

$$\Phi(\mathbb{V}; \bar{\mathcal{A}}) = \max_{i=1,\dots,m} \Phi(\mathbb{V}; \mathcal{A}_i). \tag{3.6}$$

3.2. Main results

Theorem 3.3. *Consider the setup of §1.1, let $\mathbb{U} \subseteq \mathbb{X}_0$ and suppose that the difference set $\mathbb{U}' = \mathbb{U} - \mathbb{U}$ is such that*

(i) $\mathbb{U}'$ is a cone, and
(ii) $\mathbb{U}' \subseteq \mathbb{V}_1 \cup \cdots \cup \mathbb{V}_d =: \mathbb{V}$, where each $\mathbb{V}_i \subseteq \mathbb{X}_0$ is a subspace of dimension at most n.

Suppose that, for some $0 < \epsilon < 1$, either

(a) $m \gtrsim \alpha^{-1} \cdot \Phi(S(\mathbb{U}'); \bar{\mathcal{A}}) \cdot \left[n \log(2\gamma(\mathbb{U}'; \bar{\mathcal{A}})) + \log(2d/\epsilon)\right]$, where

$$\gamma(\mathbb{U}'; \bar{\mathcal{A}}) = \frac{\min\{\Phi(S(\mathbb{V}); \bar{\mathcal{A}}), \Phi(S(\mathbb{U}' - \mathbb{U}'); \bar{\mathcal{A}})\}}{\Phi(S(\mathbb{U}'); \bar{\mathcal{A}})}, \tag{3.7}$$

(b) $m \gtrsim \alpha^{-1} \cdot \Phi(S(\mathbb{U}' - \mathbb{U}'); \bar{\mathcal{A}}) \cdot [\log(2d/\epsilon) + n]$,
(c) or $m \gtrsim \alpha^{-1} \cdot \Phi(S(\mathbb{V}); \bar{\mathcal{A}}) \cdot \log(2nd/\epsilon)$.

Let $x \in \mathbb{X}_0$, $\theta \geq \|x\|_{\mathbb{X}}$ and $\zeta \geq 0$. If $N = \frac{1}{m} \sum_{i=1}^m \|e_i\|_{\mathbb{Y}_i}^2$, then

$$\mathbb{E}\|x - \check{x}\|_{\mathbb{X}}^2 \lesssim \frac{\beta}{\alpha} \cdot \inf_{u \in \mathbb{U}} \|x - u\|_{\mathbb{X}}^2 + \theta^2 \epsilon + \frac{\zeta^2}{\alpha} + \frac{N}{\alpha},$$

where $\check{x} = \min\{1, \theta/\|\hat{x}\|_{\mathbb{X}}\}\hat{x}$ for any ζ-minimizer $\hat{x}$ of (1.3) with noisy measurements (1.2).

As observed in §1.2, assumptions (i) and (ii) hold in many cases of interest. We note that condition (a) is weaker than (b) (see the proof of Theorem E.2). It depends linearly on the variation over the smaller set $S(\mathbb{U}')$, and only logarithmically on the variation over $S(\mathbb{U}' - \mathbb{U}')$. For the sparse model (Example 3.2), where $\mathbb{U} = \Sigma_s$, we have $\mathbb{U}' = \Sigma_{2s}$ and $\mathbb{U}' - \mathbb{U}' = \Sigma_{4s}$ (see Appendix B). In general, condition (c) has a better dependence on n than both conditions (a) and (b), at the expense of evaluating the variation over the generically larger set $\mathbb{V}$ defined in assumption (ii).

Theorem 3.3 is very general. However, as we discuss in Appendix B, it yields suboptimal bounds in cases such as Example 3.2. Fortunately, by making an additional assumption, we can resolve this shortcoming.

Theorem 3.4. *Consider the setup of §1.1 and let $\mathbb{U} \subseteq \mathbb{X}_0$ be such that assumptions (i) and (ii) of Theorem 3.3 hold and also that*

(iii) $\{u \in \mathbb{U}' : \|u\|_{\mathbb{X}} \leq 1\} \subseteq \mathrm{conv}(\mathbb{W})$, *where $\mathbb{W}$ is a finite set of size $|\mathbb{W}| = M$.*

Suppose that, for some $0 < \epsilon < 1$, either

(a) $m \gtrsim \alpha^{-1} \cdot \Phi(S(\mathbb{U}') \cup \mathbb{W}; \bar{\mathcal{A}}) \cdot L$, *where*

$$L = \log\left(2\Phi(S(\mathbb{U}') \cup \mathbb{W}; \bar{\mathcal{A}})/\alpha\right) \cdot \left[\log\left(2\gamma(\mathbb{U}'; \bar{\mathcal{A}})\right) + \log(2M) \cdot \log^2(\log(2d) + n)\right] + \log(\epsilon^{-1})$$

and $\gamma(\mathbb{U}'; \bar{\mathcal{A}})$ is as in (3.7);

(b) $m \gtrsim \alpha^{-1} \cdot \Phi(S(\mathbb{U}' - \mathbb{U}') \cup \mathbb{W}; \bar{\mathcal{A}}) \cdot L$, *where*

$$L = \log\left(2\Phi(S(\mathbb{U}') \cup \mathbb{W}; \bar{\mathcal{A}})/\alpha\right) \cdot \log(2M) \cdot \log^2(\log(2d) + n) + \log(\epsilon^{-1});$$

or

(c) $m \gtrsim \alpha^{-1} \cdot \Phi(S(\mathbb{V}) \cup \mathbb{W}; \bar{\mathcal{A}}) \cdot L$, *where*

$$L = \log\left(2\Phi(S(\mathbb{V}) \cup \mathbb{W}; \bar{\mathcal{A}})/\alpha\right) \cdot \log(2M) \cdot \log^2(\log(2d) + n) + \log(\epsilon^{-1}).$$

Let $x \in \mathbb{X}_0$, $\theta \geq \|x\|_{\mathbb{X}}$ and $\zeta \geq 0$. If $N = \frac{1}{m}\sum_{i=1}^m \|e_i\|_{\mathbb{Y}_i}^2$, then

$$\mathbb{E}\|x - \check{x}\|_{\mathbb{X}}^2 \lesssim \frac{\beta}{\alpha} \cdot \inf_{u \in \mathbb{U}} \|x - u\|_{\mathbb{X}}^2 + \theta^2\epsilon + \frac{\zeta^2}{\alpha} + \frac{N}{\alpha},$$

where $\check{x} = \min\{1, \theta/\|\hat{x}\|_{\mathbb{X}}\}\hat{x}$ for any ζ-minimizer $\hat{x}$ of (1.3) with noisy measurements (1.2).

We next consider several applications of these results. Before doing so, some additional comments are in order.

First, notice that (1.6) involves a truncation $\check{x}$ of a ζ-minimizer $\hat{x}$ of (1.3). This is a technical condition that is used to establish the expectation bound. See §E.3.5 and Remark E.14.

Second, the main difference between Theorem 3.3 and Theorem 3.4 is the additional assumption (iii), which states that the unit ball of $\mathbb{U}'$ should be contained in the convex hull of a set $\mathbb{W}$ that is not too large (since M enters logarithmically in the measurement condition) and has small variation (since the variation is taken over a set that includes $\mathbb{W}$). In practice, this assumption allows the dependence of the measurement condition on d and n to be reduced from essentially $\log(2d) + n$ in Theorem 3.3 to $\log^2(\log(2d) + n)$. We remark in passing that assumption (iii) always holds for some set $\mathbb{W}$ whenever assumptions (i) and (ii) of Theorem 3.3 hold. However, the resulting $\mathbb{W}$ may not lead to a good bound in the measurement condition. See Remark B.1. Nonetheless, in certain problems such as Example 3.2, we can indeed derive a suitable $\mathbb{W}$ that leads to a significantly better measurement condition than those implied by Theorem 3.3. See §B.1. Further, in §B.2 we discuss how (iii) can also be used for various structured sparse models.

Finally, as we show in the proof, assumption (i) in Theorem 3.3 can also be removed. The only difference comes in the definition of $\gamma(\mathbb{U}'; \bar{\mathcal{A}})$, which now involves a ratio of variations over slightly different sets. See Remark E.13.

4. Application to active learning in regression

In active learning, one has the flexibility to choose where to sample the ground truth so as to enhance the generalization performance of the learning algorithm. In the standard regression problem, where the ground truth is a function $f : D \to \mathbb{R}$, this means the training data takes the form

$$\{(z_i, f(z_i) + e_i)\}_{i=1}^m,$$

where one is free to choose the *sample points* z_i. Building on Example 1.1, we now show how our theoretical results lead naturally to an active learning strategy. This extends well-known *leverage score sampling* (Avron et al., 2017; Chen et al., 2016; Chen & Price, 2019; Derezinski et al., 2018; Erdelyi et al., 2020; Gajjar et al., 2023; Ma et al., 2015) to general nonlinear model classes.

Consider Example 1.1 and let $\mathbb{V} \subseteq \mathbb{X}_0$ be an arbitrary model class. In this case, we have

$$\Phi(S(\mathbb{V}); \mathcal{A}) = \operatorname*{ess\,sup}_{z \sim \rho} \sup_{v \in \mathbb{V}, v \neq 0} |v(z)|^2/(\nu(z)\|v\|_{L^2_\rho(D)}^2). \tag{4.1}$$

For convenience, we now let

$$K(\mathbb{V})(z) = \sup_{v \in \mathbb{V}, v \neq 0} |v(z)|^2/\|v\|_{L^2_\rho(D)}^2. \tag{4.2}$$

This is sometimes termed the *Christoffel function* of the set $\mathbb{V}$. Up to a scaling factor, it is the same as the *leverage score* function of $\mathbb{V}$ (see (Adcock et al., 2023, §A.2)). Notice that

$$\Phi(S(\mathbb{V}); \mathcal{A}) = \operatorname*{ess\,sup}_{z \sim \rho} \left\{K(\mathbb{V})(z)/\nu(z)\right\}. \tag{4.3}$$

Recall that the measurement conditions in Theorem 3.3 scales linearly with $\Phi(S(\mathbb{U}'); \mathcal{A})$, where $\mathbb{U}' = \mathbb{U} - \mathbb{U}$. Hence we now look to choose the sampling distribution μ to minimize this quantity. The following result is immediate.

Proposition 4.1 (Christoffel sampling)**.** *Suppose that $K(\mathbb{W})$ is integrable and positive almost everywhere. Then (4.1) is minimized by the choice*

$$\mathrm{d}\mu^\star(\mathbb{W})(z) = \frac{K(\mathbb{W})(z)}{\int_D K(\mathbb{W})(z)\,\mathrm{d}\rho(z)}\,\mathrm{d}\rho(z). \tag{4.4}$$

In this case, one has the optimal value

$$\Phi(S(\mathbb{W}); \mathcal{A}) = \int_D K(\mathbb{W})(z)\,\mathrm{d}\rho(z). \tag{4.5}$$

This result states that to obtain the best measurement condition over all possible sampling measures μ one should choose a measure that is proportional to the Christoffel function – an approach termed *Christoffel sampling* in (Adcock et al., 2023). Combining this with Theorem 3.3, we deduce the following learning guarantees for Christoffel sampling.

Corollary 4.2 (Learning guarantees for Christoffel sampling)**.** *Consider Example 1.1 and let $\mathbb{U} \subseteq \mathbb{X}_0$ and $\mathbb{U}' = \mathbb{U} - \mathbb{U}$ be such that (i) and (ii) of Theorem 3.3 hold. Suppose that, for some $0 < \epsilon < 1$, either*

(a) $\mu = \mu^\star(\mathbb{U}')$ in (4.4) and

$$m \gtrsim \int_D K(\mathbb{U}')(z)\,\mathrm{d}\rho(z)\cdot[\log(2d/\epsilon) + n\log(2\gamma(\mathbb{U}'))]\,,$$

where $\gamma(\mathbb{U}') = \min\{\gamma_1, \gamma_2\}$ and
$\gamma_1 = \operatorname{ess\,sup}_{z\sim\rho}\{K(\mathbb{V})(z)/\nu(z)\}$,
$\gamma_2 = \operatorname{ess\,sup}_{z\sim\rho}\{K(\mathbb{U}'-\mathbb{U}')(z)/\nu(z)\}$,

(b) $\mu = \mu^\star(\mathbb{U}'-\mathbb{U}')$ and

$$m \gtrsim \int_D K(\mathbb{U}'-\mathbb{U}')(z)\,\mathrm{d}\rho(z)\cdot[\log(2d/\epsilon) + n]\,,$$

(c) or $\mu = \mu^\star(\mathbb{V})$ and

$$m \gtrsim \int_D K(\mathbb{V})(z)\,\mathrm{d}\rho(z)\cdot\log(2dn/\epsilon).$$

Let $f \in C(\overline{D})$, $\theta \geq \|f\|_{L^2_\rho(D)}$ and $\zeta \geq 0$. Then

$$\mathbb{E}\|f - \check{f}\|^2_{L^2_\rho(D)} \lesssim \inf_{u\in\mathbb{U}} \|f-u\|^2_{L^2_\rho(D)} + \theta^2\epsilon + \zeta^2 + \frac{1}{m}\|e\|_2^2,$$

for any ζ-minimizer $\hat{f}$ of (1.4), where $e = (e_i)_{i=1}^m$ and $\check{f} = \min\{1, \theta/\|f\|_{L^2_\rho(D)}\}\hat{f}$.

This corollary describes three different variants of Christoffel sampling, depending on whether one chooses $\mathbb{U}'$, $\mathbb{U}'-\mathbb{U}'$ or $\mathbb{V}$ as the set over which to define the Christoffel function. Each choice leads to a slightly different measurement condition, but the key point is they all scale linearly in the integral of the corresponding Christoffel function. We note in passing that is also possible to develop a version of Corollary 4.2 under the additional assumption (iii) of Theorem 3.4. In this case, the relevant relevant set should also include $\mathbb{W}$.

A key question in active learning is how much one improves over the inactive case, i.e., $\mu = \rho$. In §C.1 we explain how this improvement can be significant whenever the Christoffel function is 'spiky'. In §C we also discuss several scenarios where this active learning strategy is provably near-optimal. In particular, in §C.4 we consider the case of *sparse regression*, where $\mathbb{U}$ is the set of s-sparse expansions in an orthonormal system. Fig. 1 shows a numerical experiment of active learning for sparse regression using multivariate polynomials. Here, the Christoffel function is highly spiky, and in correspondence with the discussion, Christoffel sampling leads a significant improvement over i.i.d. random sampling from ρ.

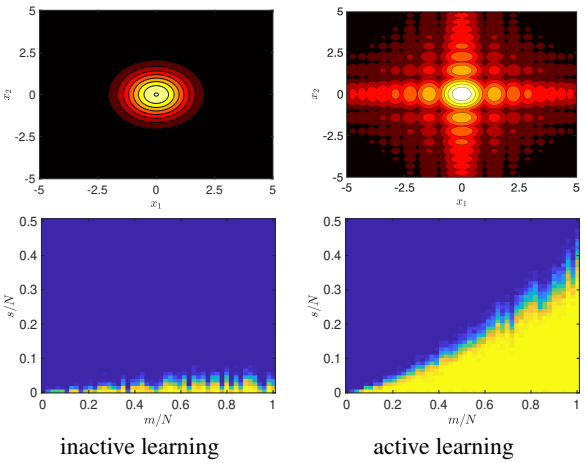

Figure 1: Active learning for sparse polynomial regression. This figure considers a sparse regression problem using orthonormal Hermite polynomials over $\mathbb{X} = L^2_\rho(\mathbb{R}^2)$, where ρ is the standard Gaussian measure. In this case, $\mathbb{U}$ is the set of all s-sparse polynomials of total degree at most 20. Top left: the density of Gaussian measure ρ. Top right: the density of the Christoffel sampling measure μ. Bottom row: Phase transition portraits showing the probability of successful recovery of an s-sparse polynomial for different values of s and m using either inactive learning, i.e., i.i.d. sampling from ρ (left) or μ (right).

5. Application to compressed sensing with generative models

Compressed sensing with generative models involves replacing the classical model class $\mathbb{U} = \Sigma_s$ in (3.3) with the range of a generative neural network that has been trained on some training data relevant to the learning problem at hand (e.g., brain images in the case of a MRI reconstruction problem). In this section, we demonstrate how our main results can be applied to this problem in the case of ReLU generative neural networks.

We consider the discrete setting, where $\mathbb{X}_0 = \mathbb{X} = \mathbb{R}^N$. Following (Berk et al., 2023a), we fix an ℓ-layer ReLU neural network $G : \mathbb{R}^n \to \mathbb{R}^N$ given by

$$G(z) = \sigma(A_\ell\sigma(A_{\ell-1}\sigma(\cdots A_2\sigma(A_1 z)\cdots))), \qquad (5.1)$$

where σ is the ReLU activation function, $A_i \in \mathbb{R}^{p_i\times p_{i-1}}$ and $p_0 = n$, $p_\ell = N$. Notice that this network has no bias terms, which is a standard assumption (Berk et al., 2023a). For sampling, we consider the general setup of §1.1, i.e., where the measurements are the result of arbitrary (scalar- or vector-valued) linear operators applied to the object $x \in \mathbb{R}^N$.

We next state a general learning guarantee for generative models. For this, we require the following from (Berk et al., 2023a). First, if $\mathbb{U} = \mathrm{Ran}(G)$, then the difference set $\mathbb{U} - \mathbb{U} = \mathrm{Ran}(G) - \mathrm{Ran}(G) = \cup_{i=1}^{d} \mathcal{C}_i$, where each $\mathcal{C}_i$ is a polyhedral cone of dimension at most $2n$ and $\log(d) \leq 2n \sum_{i=1}^{\ell-1} \log(2ep_i/n)$. See (Berk et al., 2023a, Lem. S2.2 & Rem. S2.3). Next, as in (Berk et al., 2023a, Defn. 5), we define the *piecewise linear expansion* of $\mathbb{U} - \mathbb{U}$ as

$$\Delta(\mathbb{U} - \mathbb{U}) = \mathrm{span}(\mathcal{C}_1) \cup \cdots \cup \mathrm{span}(\mathcal{C}_d). \tag{5.2}$$

Corollary 5.1. *Consider the setup of §1.1, where $\mathbb{X}_0 = \mathbb{X} = \mathbb{R}^N$ with the Euclidean inner product and norm. Let $\mathbb{U} = \mathrm{Ran}(G)$, where G is a ReLU generative neural network as in (5.1) with L layers and widths $n = p_0 \leq p_1, \ldots, p_{\ell-1}, p_\ell = N$. Suppose that, for some $0 < \epsilon < 1$,*

$$m \gtrsim \alpha^{-1} \cdot \Phi(S(\mathbb{U} - \mathbb{U}); \bar{\mathcal{A}}) \cdot n \cdot L, \tag{5.3}$$

where

$$L = \sum_{i=1}^{\ell-1} \log(2ep_i/n) + \log(2/\epsilon) + \log\left(2\frac{\Phi(\Delta(\mathbb{U} - \mathbb{U}); \bar{\mathcal{A}})}{\Phi(\mathbb{U} - \mathbb{U}; \bar{\mathcal{A}})}\right).$$

Let $x \in \mathbb{R}^N$, $\|x\|_{\mathbb{X}} \leq 1$ and $\zeta \geq 0$. If $N = \frac{1}{m}\sum_{i=1}^{m} \|e_i\|_{\mathbb{Y}_i}^2$, then

$$\mathbb{E}\|x - \check{x}\|_{\ell^2}^2 \lesssim \frac{\beta}{\alpha} \cdot \inf_{u \in \mathbb{U}} \|x - u\|_{\ell^2}^2 + \epsilon + \frac{\zeta^2}{\alpha} + \frac{N}{\alpha},$$

where $\check{x} = \min\{1, 1/\|\hat{x}\|_{\ell^2}\}\hat{x}$, for any ζ-minimizer $\hat{x}$ of (1.3) with noisy measurements (1.2).

This result shows that the sample complexity depends linearly on the dimension n of the latent space of the generative model, up to log terms, multiplied by the variations $\Phi(S(\mathbb{U} - \mathbb{U}); \bar{\mathcal{A}})$ and $\Phi(S(\Delta(\mathbb{U} - \mathbb{U})); \bar{\mathcal{A}})$. In particular, if these variations are small and the network widths $p_1, \ldots, p_{\ell-1} \lesssim N$ we obtain a measurement condition $m \gtrsim n(\ell \log(N/n) + \log(2/\epsilon))$ that scales linearly in n.

This result provides a guarantee for compressed sensing with generative models and very general types of data. As noted, existing guarantees consider either sampling with (sub)Gaussian random matrices (Bora et al., 2017) or subsampled unitary matrices (i.e., Example 2.4) (Berk et al., 2023b). In Appendix D we show how Corollary 5.1 implies the results of (Berk et al., 2023b) (in fact, improves them). In particular, it allows one to derive an optimal sampling strategy for sampling with subsampled unitary matrices (i.e., an optimal choice of the probability distribution π in Example 2.4). Such a sampling strategy was also derived in (Adcock et al., 2023), albeit without theoretical guarantees. To illustrate its effectiveness, in Fig. 2 we compare this sampling strategy against random sampling. This figure considers a synthetic MRI experiment, where the unknown objects are

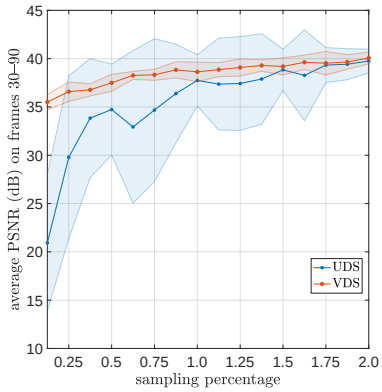

Figure 2: Improved sampling strategies for compressed sensing with generative models. This figure considers image recovery from discrete Fourier measurements, where G is a generative model trained on 3D brain images. The Variable Density Sampling (VDS) strategy implied by Corollary D.2 leads to a significant improvement over Uniform Density Sampling (UDS), i.e., random sampling with $\pi = (1/N, \ldots, 1/N)$. This figure is based on (Adcock et al., 2023, §C.6). See (Adcock et al., 2023, §C) for the full experimental details.

3D brain images and $U = F$ is a 3D discrete Fourier transform. As is evident, the sampling strategy obtained from the theory (namely, Corollary D.2, which is a consequence of Corollary 5.1) leads to a significant improvement.

6. Conclusions and limitations

We introduced a unified framework for learning unknown objects with nonlinear model classes from (multimodal) linear samples. We showed its versatility, obtaining results for matrix sketching and compressed sensing as corollaries of our results. We then used it to extend and improve recent results for active learning in regression and generative compressed sensing. Extensions to other structured sparsity models, e.g., joint and group sparsity, can also be made by adjusting the model class.

We also acknowledge several limitations. First, we overlook nonlinear measurements arising in nonlinear (classical or generative) compressed sensing. Recently (Chen et al., 2023) proposed a framework for such measurements with uniform recovery guarantees. Second, we focus on Hilbert measurement spaces. Many important applications operate in Banach spaces (Adcock et al., 2022b). Third, we do not address finding (local) minimizers of nonconvex problems. Fourth, our guarantees are uniform over objects and model classes, which is potentially pessimistic. Nonuniform analysis may be sharper (Trunschke, 2023). Finally, coherence is a necessary condition in standard compressed sensing (Candès & Plan, 2011). Determining whether variation is similarly fundamental in this framework is an open problem.

Acknowledgements

BA acknowledges the support of the Natural Sciences and Engineering Research Council of Canada of Canada (NSERC) through grant RGPIN-2021-611675. ND acknowledges the support of Florida State University through the CRC 2022-2023 FYAP grant program.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Adcock, B. and Brugiapaglia, S. Is Monte Carlo a bad sampling strategy for learning smooth functions in high dimensions? *arXiv:2208.09045*, 2022.

Adcock, B. and Dexter, N. The gap between theory and practice in function approximation with deep neural networks. *SIAM J. Math. Data Sci.*, 3(2):624–655, 2021.

Adcock, B. and Hansen, A. C. *Compressive Imaging: Structure, Sampling, Learning*. Cambridge University Press, Cambridge, UK, 2021.

Adcock, B., Hansen, A. C., Poon, C., and Roman, B. Breaking the coherence barrier: a new theory for compressed sensing. *Forum Math. Sigma*, 5:e4, 2017.

Adcock, B., Brugiapaglia, S., Dexter, N., and Moraga, S. Deep neural networks are effective at learning high-dimensional Hilbert-valued functions from limited data. In Bruna, J., Hesthaven, J. S., and Zdeborová, L. (eds.), *Proceedings of The Second Annual Conference on Mathematical and Scientific Machine Learning*, volume 145 of *Proc. Mach. Learn. Res. (PMLR)*, pp. 1–36. PMLR, 2021.

Adcock, B., Brugiapaglia, S., Dexter, N., and Moraga, S. On efficient algorithms for computing near-best polynomial approximations to high-dimensional, Hilbert-valued functions from limited samples. *arXiv:2203.13908*, 2022a.

Adcock, B., Brugiapaglia, S., Dexter, N., and Moraga, S. Near-optimal learning of Banach-valued, high-dimensional functions via deep neural networks. *arXiv:2211.12633*, 2022b.

Adcock, B., Brugiapaglia, S., and Webster, C. G. *Sparse Polynomial Approximation of High-Dimensional Functions*. Comput. Sci. Eng. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2022c.

Adcock, B., Cardenas, J. M., Dexter, N., and Moraga, S. *Towards optimal sampling for learning sparse approximation in high dimensions*, chapter 2, pp. 9–77. Springer Optimization and Its Applications. Springer, 2022d.

Adcock, B., Cardenas, J. M., and Dexter, N. CS4ML: A general framework for active learning with arbitrary data based on Christoffel functions. *arXiv:2306.00945*, 2023.

Alekseev, A. K., Navon, I. M., and Zelentsov, M. E. The estimation of functional uncertainty using polynomial chaos and adjoint equations. *Internat. J. Numer. Methods Fluids*, 67(3):328–341, 2011.

Avron, H., Kapralov, M., Musco, C., Musco, C., Velingker, A., and Zandieh, A. Random Fourier features for kernel ridge regression: approximation bounds and statistical guarantees. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 253–262. PMLR, 2017.

Avron, H., Kapralov, M., Musco, C., Musco, C., Velingker, A., and Zandieh, A. A universal sampling method for reconstructing signals with simple Fourier transforms. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, pp. 1051–1063, New York, NY, USA, 2019. Association for Computing Machinery.

Baraniuk, R. G., Cevher, V., Duarte, M. F., and Hedge, C. Model-based compressive sensing. *IEEE Trans. Inform. Theory*, 56(4):1982–2001, 2010.

Berk, A., Brugiapaglia, S., Joshi, B., Plan, Y., Scott, M., and Yilmaz, O. A coherence parameter characterizing generative compressed sensing with Fourier measurements. *IEEE J. Sel. Areas Inf. Theory*, 3(3):502–512, 2023a.

Berk, A., Brugiapaglia, S., Plan, Y., Scott, M., Sheng, X., and Yilmaz, O. Model-adapted Fourier sampling for generative compressed sensing. *arXiv:2310.04984*, 2023b.

Bigot, J., Boyer, C., and Weiss, P. An analysis of block sampling strategies in compressed sensing. *IEEE Trans. Inform. Theory*, 62(4):2125–2139, 2016.

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. In *International Conference on Machine Learning*, pp. 537–546, 2017.

Bourrier, A., Davies, M. E., Peleg, T., Pérez, P., and Gribonval, R. Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems. *IEEE Trans. Inform. Theory*, 60(12):7928–7946, 2014.

Boyer, C., Bigot, J., and Weiss, P. Compressed sensing with structured sparsity and structured acquisition. *Appl. Comput. Harmon. Anal.*, 46(2):312–350, 2019.

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

Candès, E. J. and Plan, Y. A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inform. Theory*, 57 (11):7235–7254, 2011.

Chen, J., Scarlett, J., Ng, M., and Liu, Z. A unified framework for uniform signal recovery in nonlinear generative compressed sensing. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Chen, S., Varma, R., Singh, A., and Kovačcević, J. A statistical perspective of sampling scores for linear regression. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 1556–1560, 2016.

Chen, X. and Price, E. Active regression via linear-sample sparsification. In Beygelzimer, A. and Hsu, D. (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 663–695. PMLR, 2019.

Chun, I.-Y. and Adcock, B. Compressed sensing and parallel acquisition. *IEEE Trans. Inform. Theory*, 63(8):4860–4882, 2017.

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

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

Czarnecki, W. M., Osindero, S., Jaderberg, M., Swirszcz, G., and Pascanu, R. Sobolev training for neural networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Davenport, M. A. and Romberg, J. An overview of low-rank matrix recovery from incomplete observations. *IEEE J. Sel. Topics Signal Process.*, 10(4):602–622, 2016.

Davenport, M. A., Duarte, M. F., Eldar, Y. C., and Kutyniok, G. Introduction to compressed sensing. In Eldar, Y. C. and Kutyniok, G. (eds.), *Compressed Sensing: Theory and Applications*, pp. 1–64. Cambridge University Press, Cambridge, UK, 2012.

Derezinski, M., Warmuth, M. K. K., and Hsu, D. J. Leveraged volume sampling for linear regression. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Dexter, N., Tran, H., and Webster, C. A mixed ℓ_1 regularization approach for sparse simultaneous approximation of parameterized PDEs. *ESAIM Math. Model. Numer. Anal.*, 53:2025–2045, 2019.

Dirksen, S. Dimensionality reduction with subgaussian matrices: a unified theory. *Found. Comput. Math.*, 16: 1367–1396, 2016.

Dolbeault, M. and Cohen, A. Optimal sampling and christoffel functions on general domains. *Constructive Approximation*, 56(1):121–163, 2022.

Donoho, D. L. and Tanner, J. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. Trans. Roy. Soc. A*, 367(1906):4273–4293, 2009.

Duarte, M. F. and Eldar, Y. C. Structured compressed sensing: from theory to applications. *IEEE Trans. Signal Process.*, 59(9):4053–4085, 2011.

Eigel, M., Schneider, R., and Trunschke, P. Convergence bounds for empirical nonlinear least-squares. *ESAIM Math. Model. Numer. Anal.*, 56(1):79–104, 2022.

Erdelyi, T., Musco, C., and Musco, C. Fourier sparse leverage scores and approximate kernel learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 109–122. Curran Associates, Inc., 2020.

Feng, X. and Zeng, L. Gradient-enhanced deep neural network approximations. *J. Mach. Learn. Model. Comput.*, 3(4):73–91, 2022.

Foucart, S. and Rauhut, H. *A Mathematical Introduction to Compressive Sensing*. Appl. Numer. Harmon. Anal. Birkhäuser, New York, NY, 2013.

Gajjar, A., Musco, C., and Hegde, C. Active learning for single neuron models with Lipschitz non-linearities. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 4101–4113. PMLR, 2023.

Genzel, M., Kutyniok, G., and März, M. ℓ^1-analysis minimization and generalized (co-)sparsity: When does recovery succeed? *Appl. Comput. Harmon. Anal.*, 52:82–140, 2021. ISSN 1063-5203.

Han, J., Jentzen, A., and E, W. Solving high-dimensional partial differential equations using deep learning. *Proc. Natl. Acad. Sci. U.S.A.*, 115(34):8505–8510, 2018.

Hashemi, A., Schaeffer, H., Shi, R., Topcu, U., Tran, G., and Ward, R. Generalization bounds for sparse random feature expansions. *Appl. Comput. Harmon. Anal.*, 62: 310–330, 2023.

Jalal, A., Arvinte, M., Daras, G., Price, E., Dimakis, A., and Tamir, J. Robust compressed sensing MR imaging with deep generative priors. In *NeurIPS 2021 Workshop on Deep Learning and Inverse Problems*, 2021.

Kabanava, M. and Rauhut, H. Cosparsity in compressed sensing. In Boche, H., Calderbank, R., Kutyniok, G., and Vybíral, J. (eds.), *Compressed Sensing and its Applications: MATHEON Workshop 2013*, Applied and Numerical Harmonic Analysis, pp. 315–339. Birkhäuser, Cham, 2015.

Krahmer, F. and Ward, R. Stable and robust sampling strategies for compressive imaging. *IEEE Trans. Image Process.*, 23(2):612–622, 2013.

Krahmer, F., Needell, D., and Ward, R. Compressive sensing with redundant dictionaries and structured measurements. *SIAM J. Math. Anal.*, 47(6):4606–4629, 2015.

Ma, P., Mahoney, M. W., and Yu, B. A statistical perspective on algorithmic leveraging. *J. Mach. Learn. Res.*, 16:861–911, 2015.

Malik, O., Xu, Y., Cheng, N., Becker, S., Doostan, A., and Narayan, A. Fast algorithms for monotone lower subsets of Kronecker least squares problems. *arXiv:2209.05662*, 2022.

McRobbie, D. W., Moore, E. A., Graves, M. J., and Prince, M. R. *MRI: From Picture to Proton*. Cambridge University Press, Cambridge, 2nd edition, 2006.

Monajemi, H., Jafarpour, S., Gavish, M., Donoho, D. L., Ambikasaran, S., Bacallado, S., Bharadia, D., Chen, Y., Choi, Y., Chowdhury, M., et al. Deterministic matrices matching the compressed sensing phase transitions of Gaussian random matrices. *Proc. Natl. A. Sci. USA*, 110 (4):1181–1186, Jan. 2013.

Nam, S., Davies, M. E., Elad, M., and Gribonval, R. The cosparse analysis model and algorithms. *Appl. Comput. Harmon. Anal.*, 34(1):30–56, 2013.

O'Leary-Roseberry, T., Villa, U., Chen, P., and Ghattas, O. Derivative-informed projected neural networks for high-dimensional parametric maps governed by PDEs. *Comput. Methods Appl. Mech. Engrg.*, 388(1):114199, 2022.

O'Leary-Roseberry, T., Chen, P., Villa, U., and Ghattas, O. Derivative-Informed Neural Operator: An efficient framework for high-dimensional parametric derivative learning. *Journal of Computational Physics*, 496:112555, 2024. ISSN 0021-9991.

Peng, J., Hampton, J., and Doostan, A. On polynomial chaos expansion via gradient-enhanced l_1-minimization. *J. Comput. Phys.*, 310:440–458, 2016.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, 378:686–707, 2019.

Rauhut, H. and Ward, R. Interpolation via weighted ℓ^1 minimization. *Appl. Comput. Harmon. Anal.*, 40(2):321–351, 2016.

Tillmann, A. M. and Pfetsch, M. E. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Trans. Inform. Theory*, 60(2):1248–1259, 2014.

Traonmilin, Y. and Gribonval, R. Stable recovery of low-dimensional cones in Hilbert spaces: one RIP to rule them all. *Appl. Comput. Harmon. Anal.*, 45(1):170–205, 2018.

Traonmilin, Y., Puy, G., Gribonval, R., and Davies, M. E. Compressed sensing in Hilbert spaces. In Boche, H., Caire, G., Calderbank, R., März, M., Kutyniok, G., and Mathar, R. (eds.), *Compressed Sensing and its Applications: Second International MATHEON Conference 2015*, Applied and Numerical Harmonic Analysis, pp. 359–384. Birkhäuser, Cham, 2017.

Trunschke, P. Convergence bounds for local least squares approximation. *arXiv:2208.10954*, 2023.

van den Berg, E. and Friedlander, M. P. Probing the pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comput.*, 31(2):890–912, 2008.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, UK, 2018.

Woodruff, D. P. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

Yu, J., Lu, L., Meng, X., and Karniadakis, G. E. Gradient-enhanced physics-informed neural networks for forward and inverse PDE problems. *Comput. Methods Appl. Mech. Engrg.*, 393:114823, 2022.

A. Matrix sketching for large least-squares problems: additional discussion

Consider Example 2.1, where for convenience we also assume that X has full column rank. We now show that sample complexity bound (2.6) follow directly from our general learning guarantees.

Recall that $\mathbb{U}$ is the linear subspace (2.3) in this example with $\dim(\mathbb{U}) = n$. Hence we strive to apply condition (c) of Theorem 3.3 with $d = 1$. We first calculate its variation. Recalling the definition of $\mathcal{A} = \mathcal{A}_1 = \cdots = \mathcal{A}_m$ and $\mathbb{Y}$ in this case and letting $u = Xz$ be an arbitrary element of $\mathbb{U}$, we notice that

$$\|A(u)\|_{\mathbb{Y}}^2 \leq \max_{i=1,\dots,N} \frac{1}{\pi_i} |e_i^* X z|^2$$

Therefore,

$$\Phi(S(\mathbb{U}); \bar{\mathcal{A}}) = \Phi(S(\mathbb{U}); \mathcal{A}) = \max_{i=1,\dots,N} \left\{ \frac{\tau(X)(i)}{\pi_i} \right\},$$

where $\tau(X)(i)$ are the leverage scores of X, as in (2.5) (note that it is a short argument to show that $\tau(X)(i) = x_i^*(X^*X)^{-1}x_i$, where x_i is the ith row of X).

Having obtained an explicit expression for the variation, we can now choose a discrete probability distribution $\pi = \{\pi_1, \dots, \pi_N\}$ to minimize it. It is a short exercise to show that

$$\sum_{i=1}^{N} \tau(X)(i) = n.$$

Therefore, we set

$$\pi_i = \frac{\tau(X)(i)}{n}, \quad i = 1, \dots, N,$$

as in (2.4). We now apply condition (c) of Theorem 3.3 to obtain the near-optimal sample complexity bound

$$m \gtrsim n \cdot \log(2n/\epsilon).$$

B. Compressed sensing with isotropic vectors: additional discussion

In this section, we provide additional discussion on the problem of compressed sensing with isotropic vectors as studied in Example 2.3.

B.1. Classical compressed sensing

Consider the classical compressed sensing problem as in Example 2.3, with $\mathbb{U} = \Sigma_s$ as in Example 3.2. Our aim is to derive the measurement condition (3.5) as a corollary of our general theory.

In this case, the collection $\bar{\mathcal{A}} = \{\mathcal{A}_i\}_{i=1}^m$, where $\mathcal{A}_1 = \cdots = \mathcal{A}_m = \mathcal{A}$ with $\mathcal{A}$ as in Example 3.2. Recall that $\alpha = \beta = 1$ as $\mathcal{A}$ is isotropic (2.7). Observe that $\mathbb{U}' = \mathbb{U} - \mathbb{U} = \Sigma_{2s}$ in this case and moreover that assumption (i) of Theorem 3.3 holds, since tx is s-sparse for any $t \geq 0$ (in fact, $t \in \mathbb{R}$) whenever x is s-sparse. Also, assumption (ii) holds, since

$$\Sigma_{2s} = \bigcup_{\substack{S \subseteq [N] \\ |S| = 2s}} \{x \in \mathbb{C}^N : \mathrm{supp}(x) \subseteq S\}.$$

Here $\mathrm{supp}(x) = \{i : x_i \neq 0\}$ is the *support* of $x = (x_i)_{i=1}^N$ and $[N] = \{1, \dots, N\}$. Notice that this is a union of

$$d = \binom{N}{2s} \leq \left(\frac{\mathrm{e}N}{2s} \right)^{2s} \tag{B.1}$$

subspaces of dimension $2s$. Example 3.2 derives an expression (3.4) for the variation in terms of s and $\mu(\mathcal{A})$. Using this, we deduce that $\gamma(\mathbb{U}'; \bar{\mathcal{A}}) = 2$ and the measurement condition in Theorem 3.3(a) reduces to

$$m \gtrsim \mu(\mathcal{A}) \cdot s \cdot \left(s \log(\mathrm{e}N/s) + \log(\epsilon^{-1}) \right)$$

(one can readily verify that the conditions (b) and (c) give the same result, up to constants). This bound is suboptimal, since it scales quadratically in s. Fortunately, this can be overcome by using Theorem 3.4. It is well known that (see, e.g., (Adcock & Hansen, 2021, proof of Lem. 13.29) or (Foucart & Rauhut, 2013, proof of Lem. 12.37))

$$\{x \in \Sigma_s : \|x\|_2 \leq 1\} \subseteq \operatorname{conv}\left(\left\{\pm\sqrt{2s}e_i, \pm\sqrt{2s}\mathrm{i}e_i : i = 1, \dots, N\right\}\right) =: \operatorname{conv}(\mathbb{W}). \tag{B.2}$$

This set has size $M = |\mathbb{W}| = 4N$. In order to apply Theorem 3.4 we estimate the variation

$$\Phi(S(\mathbb{U}') \cup \mathbb{W}; \bar{\mathcal{A}}) = \max\left\{\Phi(S(\mathbb{U}'); \bar{\mathcal{A}}), \Phi(\mathbb{W}; \bar{\mathcal{A}})\right\}.$$

The first term is bounded by $2s\mu(\mathcal{A})$ (recall (3.4)). For the latter, we observe that, for any $x \in \mathbb{W}$, we have, for some $1 \leq i \leq N$,

$$|a^*x|^2 = 2s|a_i|^2 \leq 2s\mu(\mathcal{A}), \quad \text{a.s. } a \sim \mathcal{A}.$$

Hence $\Phi(\mathbb{W}; \bar{\mathcal{A}}) \leq 2s\mu(\mathcal{A})$ and therefore

$$\Phi(S(\mathbb{U}') \cup \mathbb{W}; \bar{\mathcal{A}}) \leq 2s\mu(\mathcal{A}).$$

Using this and the previous bounds for d and $\gamma(\mathbb{U}'; \bar{\mathcal{A}})$ we see that condition (a) of Theorem 3.4 (the same applies for conditions (b) and (c)) reduces to

$$m \gtrsim \mu(\mathcal{A}) \cdot s \cdot \left(\log(2s\mu(\mathcal{A})) \cdot \log^2(s\log(N/s))\log(N) + \log(\epsilon^{-1})\right).$$

Subject to the mild assumption that $s \gtrsim \log(N/s)$, we now obtain the well-known bound (3.5).

B.2. Extension to structured sparse models

We now discuss that assumptions (i), (ii) and, importantly, (iii) of Theorem 3.3-Theorem 3.4 hold for various structured sparse models that refine the standard sparse model considered in Example 3.2.

For the well-known *weighted sparsity* model, see (Rauhut & Ward, 2016) and (Adcock et al., 2022c, Chpt. 6). This model satisfies assumptions (i)–(iii). In particular, (Rauhut & Ward, 2016, Proof of Lem. 5.3) shows that assumption (iii) holds with $|\mathbb{W}| \leq 4N$. For the well-known *sparsity in levels* model, we refer to (Adcock & Hansen, 2021, Chpts. 11 & 13). In particular, assumption (iii) also holds with $|\mathbb{W}| \leq 4N$, as was was shown in (Adcock & Hansen, 2021, proof of Lem. 13.29).

Another common structured sparsity model is the *cosparse* model. This arises in analysis compressed sensing problems (Genzel et al., 2021; Nam et al., 2013; Kabanava & Rauhut, 2015), where the analysis operator is not an orthonormal basis. This model satisfies assumptions (i) and (ii). Moreover, if the analysis operator is a Parseval frame it also satisfies (iii) with $|\mathbb{W}| \leq 4N$. See (Krahmer et al., 2015, proof of Cor. 4.2).

Finally, we consider the case of *group sparsity* (Davenport et al., 2012; Duarte & Eldar, 2011) (which also includes the case of *joint sparsity*). In this model, we consider a fixed partition $\mathcal{G} = \{G_i\}_{i=1}^P$ of $[N]$ into nonoverlapping groups. Let $x \in \mathbb{C}^N$ and x_{G_i} be the restriction of x to the indices in the ith group G_i. We say that x is *s-group sparse* for some $1 \leq s \leq P$ if $|\{i : x_{G_i} \neq 0\}| \leq s$. Now define the set

$$\mathbb{U} = \mathbb{U}_{\mathcal{G},s} = \{x \in \mathbb{C}^N : x \text{ is } s\text{-group sparse}\}.$$

As before, $\mathbb{U}' = \mathbb{U} - \mathbb{U} = \mathbb{U}_{\mathcal{G},2s}$. Assumption (i) straightforwardly holds. Assumption (ii) also holds, as we may write

$$\mathbb{U}' = \bigcup_{\substack{S \subseteq [P] \\ |S| = 2s}} \{x \in \mathbb{C}^N : \operatorname{supp}_{\mathcal{G}}(x) \subseteq S\},$$

where $\operatorname{supp}_{\mathcal{G}}(x) = \{i \in [P] : x_{G_i} \neq 0\}$. Let $|G_i| = g_i$ and suppose without loss of generality that $g_1 \geq g_2 \geq \cdots$. Then this is a union of $\binom{P}{2s}$ subspaces of dimension at most $2s\max\{g_i\} = 2sg_1$. Hence assumption (ii) holds. We now verify assumption (iii). By a short argument, we have, for any $x \in \mathbb{U}'$,

$$\begin{aligned}\|x\|_1^* := \sum_{i=1}^N (|\operatorname{Re}(x_i)| + |\operatorname{Im}(x_i)|) \leq \sqrt{2}\sum_{i \in S}\|x_{G_i}\|_1 \leq \sqrt{2}\sum_{i \in S}\sqrt{g_i}\|x_{G_i}\|_2 \leq \sqrt{2}\sqrt{\sum_{i \in S} g_i}\|x\|_2 \\ \leq \sqrt{2(g_1 + \cdots + g_{2s})}\|x\|_2,\end{aligned}$$

(here $S = \mathrm{supp}_{\mathcal{G}}(x)$). Hence, assumption (iii) holds with $|\mathbb{W}| = 4N$, since

$$\{x \in \mathbb{U}' : \|x\|_2 \leq 1\} \subseteq \mathrm{conv}\left(\left\{\pm\sqrt{2(g_1 + \cdots + g_{2s})}e_i, \pm\sqrt{2(g_1 + \cdots + g_{2s})}\mathrm{i}e_i : i = 1, \ldots, N\right\}\right) =: \mathrm{conv}(\mathbb{W}).$$

Remark B.1 (Assumption (iii) always holds). Consider the general setup of §1.1 where $\mathbb{U} \subseteq \mathbb{X}_0$ satisfies (i) and (ii) of Theorem 3.3. Then

$$\{u \in \mathbb{U}' : \|u\|_{\mathbb{X}} \leq 1\} \subseteq \bigcup_{i=1}^{d} \{v \in \mathbb{V}_i : \|v\|_{\mathbb{X}} \leq 1\}.$$

Let $\{v_j^{(i)}\}_{j=1}^{n_i}$ be an orthonormal basis of $\mathbb{V}_i$, where $n_i = \dim(\mathbb{V}_i) \leq n$ by assumption, and define $\|c\|_1^* = \sum_{j=1}^{n_i}(|\mathrm{Re}\,(c_i)| + |\mathrm{Im}\,(c_i)|)$ for $c \in \mathbb{C}^{n_i}$. Let $v = \sum_{j=1}^{n_i} c_j v_j^{(i)} \in \mathbb{V}_i$ be arbitrary. Then

$$\|c\|_1^* \leq \sqrt{2}\|c\|_1 \leq \sqrt{2n_i}\|c\|_2 = \sqrt{2n_i}\|v\|_{\mathbb{X}}.$$

We deduce that

$$\{v \in \mathbb{V}_i : \|v\|_{\mathbb{X}} \leq 1\} \subseteq \mathrm{conv}(\mathbb{W}_i),$$

where $\mathbb{W}_i = \left\{\pm\sqrt{2n_i}v_j^{(i)}, \pm\mathrm{i}\sqrt{2n_i}v_j^{(i)} : j = 1, \ldots, n_i\right\}$ satisfies $|\mathbb{W}_i| = 4n_i \leq 4n$. Therefore,

$$\{u \in \mathbb{U}' : \|u\|_{\mathbb{X}} \leq 1\} \subseteq \bigcup_{i=1}^{d} \mathrm{conv}(\mathbb{W}_i) \subseteq \mathrm{conv}(\mathbb{W}), \quad \text{where } \mathbb{W} = \bigcup_{i=1}^{d} \mathbb{W}_i.$$

This shows that assumption (iii) of Theorem 3.4 always holds whenever assumptions (i) and (ii) of Theorem 3.3 hold. Unfortunately, this argument may be too crude in practice to be useful. Note that the set $\mathbb{W}$ constructed by this argument satisfies $|\mathbb{W}| \leq 4nd$, which leads to a term of the form $\log(8nd)$ in the various bounds in Theorem 3.4. However, in the classical compressed sensing case §B.1, for instance, we have $\log(8nd) \lesssim s\log(\mathrm{e}N/s)$ due to (B.1). Thus, using this estimate would lead to a quadratic scaling in s in the overall measurement bound. Fortunately, in this specific case, we can derive a smaller set $\mathbb{W}$ (see (B.2)) which leads to a linear scaling in s.

C. Application to active learning: additional discussion and examples

In this section, we provide some additional discussion on the active learning for regression application considered in §4.

C.1. Improvement over inactive learning

Consider Example 1.1 and suppose that the underlying measure ρ is a probability measure. Inactive learning corresponds to the scenario where $\mu = \rho$, i.e., the sampling distribution is taken equal to the underlying measure. This is sometimes termed Monte Carlo sampling. In this case, we have $\nu = \mathrm{d}\mu/\mathrm{d}\rho = 1$, and therefore for any set $\mathbb{W} \subseteq \mathbb{X}_0$,

$$\Phi(S(\mathbb{W}); \mathcal{A}) = \operatorname*{ess\,sup}_{z \sim \rho} K(\mathbb{W})(z). \tag{C.1}$$

Hence, a version of Corollary 4.2 hold for inactive learning, where each integral over the Christoffel function is replaced by an essential supremum. In other words, the improvement of Christoffel sampling over inactive learning is equivalent to the difference between the mean value of $K(\mathbb{W})$, i.e., (4.5), and its maximal value, i.e., (C.1), where $\mathbb{W}$ equal to either $\mathbb{U}'$ (part (a) of Corollary 4.2), $\mathbb{U}' - \mathbb{U}'$ or $\mathbb{V}$. Since

$$\operatorname*{ess\,sup}_{z \sim \rho} K(\mathbb{W})(z) \geq \int_D K(\mathbb{W})(z)\,\mathrm{d}\rho(z),$$

Christoffel sampling always reduces the sample complexity bound. Practically, the extent of the improvement corresponds to how 'spiky' $K(\mathbb{W})(z)$ is over $z \in D$. If $K(\mathbb{W})(z)$ is fairly flat, then one expects very little benefit. On the other hand, if $K(\mathbb{W})(z)$ has sharp peaks, then one expects a significant improvement. As discussed in (Cohen & Migliorati, 2017; Avron et al., 2019; Adcock et al., 2022d; 2023; Erdelyi et al., 2020; Gajjar et al., 2023), there are many case where significant improvements are possible because of this phenomenon. However, this is not always the case (Adcock & Brugiapaglia, 2022).

To gain further insight, we now consider several examples.

C.2. Near-optimal sampling for linear subspaces

Let $\mathbb{U}$ be a subspace of dimension n with an orthonormal basis $\{u_i\}_{i=1}^n$ in the L^2_ρ-norm. Notice that $\mathbb{U}' = \mathbb{U} - \mathbb{U} = \mathbb{U}$ in this case. Moreover, it is a short exercise (see, e.g., (Adcock et al., 2023, Lem. E.1)) to show that

$$K(\mathbb{U})(z) = \sum_{i=1}^n |u_i(z)|^2.$$

In particular, $\int_D K(\mathbb{U})(z)\,\mathrm{d}\rho(z) = n$ due to orthonormality. Hence, the Christoffel sampling measure (4.4) becomes

$$\mathrm{d}\mu^\star(z) = \frac{\sum_{i=1}^n |u_i(z)|^2}{n}\,\mathrm{d}\rho(z),$$

and, using condition (c) of Corollary 4.2 with $d = 1$, we deduce the near-optimal measurement condition

$$m \gtrsim n \cdot \log(2n/\epsilon).$$

This result is well known (see, e.g., (Cohen & Migliorati, 2017)). Note that Example 2.1 (matrix sketching) can also be interpreted as a special case of this result, after interpreting vectors as functions defined on the discrete domain $\{1,\dots,N\}$. We omit the details.

The example considered in this subsection is highly informative, as it gives a class of problems where the active learning strategy of Christoffel sampling is provably optimal up to a log factor. However, many practical problems of interest involve nonlinear model classes. We now consider the nonlinear case in more detail.

C.3. Near-optimal sampling for a 'small' union of subspaces

We commence with the case of unions of subspaces. Suppose that $\mathbb{U}' = \mathbb{U} - \mathbb{U} = \mathbb{V}_1 \cup \cdots \cup \mathbb{V}_d$ is a union of d subspaces of dimension at most n. By definition,

$$K(\mathbb{U}')(z) = \max_{i=1,\dots,d} K(\mathbb{V}_d)(z).$$

Therefore, by a crude bound and the fact that $\int_D K(\mathbb{V}_i)\,\mathrm{d}\rho(z) = n$ (recall the previous example), we obtain

$$\int K(\mathbb{U}')(z)\,\mathrm{d}\rho(z) \leq \sum_{i=1}^d \int_D K(\mathbb{V}_i)\,\mathrm{d}\rho(z) = nd.$$

Hence, applying condition (c) of Corollary 4.2, we obtain the measurement condition

$$m \gtrsim n \cdot d \cdot \log(2nd/\epsilon). \tag{C.2}$$

The main purpose of this example is to illustrate a class of problems involving nonlinear model classes for which Christoffel sampling is probably optimal, up to the log term. Indeed, if d is small, then the sample complexity bound scales like $n\log(n/\epsilon)$.

Unfortunately, as we see in the next example (and also in the next section when considering generative models), in problems of interest the number d is often not small.

C.4. The case of sparse regression

Let $s, N \in \mathbb{N}$ and $\Psi = \{\psi_i\}_{i=1}^N \subseteq L^2_\rho(D)$ be an orthonormal system (one can also consider Riesz systems with few additional difficulties – see (Adcock et al., 2023, § A.4)). Define the model class

$$\mathbb{U} = \mathbb{U}_s = \left\{ \sum_{i\in S} c_i\psi_i : c_i \in \mathbb{C},\ \forall i,\ S \subseteq \{1,\dots,N\},\ |S| = s \right\}. \tag{C.3}$$

We refer to the corresponding regression problem as a *sparse regression* problem. Notice here that, as in Appendix B, $\mathbb{U}' = \mathbb{U}_s - \mathbb{U}_s = \mathbb{U}_{2s}$ is a union of $d = \binom{N}{2s}$ subspaces. Hence, in this case, the bound (C.2) becomes meaningless.

Fortunately, following ideas presented previously in (Adcock et al., 2022d), this issue is resolved by using Theorem 3.4. As in Appendix B, notice that $\mathbb{U}' = \mathbb{U}_{2s}$ is contained in $\mathrm{conv}(\mathbb{W})$, where $\mathbb{W}$ is the set

$$\mathbb{W} = \left\{ \pm\sqrt{2s}\psi_i, \pm\sqrt{2s}\mathrm{i}\psi_i,\ i = 1,\dots,N \right\}. \tag{C.4}$$

We now define the function

$$\widetilde{K}(\Psi)(z) = \max_{i=1,\dots,N} |\psi_i(z)|^2. \tag{C.5}$$

Corollary C.1 (Active learning for sparse regression)**.** *Consider the setup of Example 1.1, where $\mathbb{U} = \mathbb{U}_s$ is as in (C.3) and $\mathbb{W}$ is as in (C.4). Then $K(\mathbb{U}' \cup \mathbb{W}) = 2s\widetilde{K}(\Psi)$, where $\widetilde{K}(\Psi)$ is as in (C.5). Now let μ be given by*

$$\mathrm{d}\mu(z) = \frac{\widetilde{K}(\Psi)(z)}{\theta(\Psi)}\,\mathrm{d}\rho(z), \quad \textit{where } \theta(\Psi) = \int_D \widetilde{K}(\Psi)(z)\,\mathrm{d}\rho(z) \tag{C.6}$$

and suppose that

$$m \gtrsim \theta(\Psi)\cdot s \cdot \left[\log(2s\theta(\Psi))\cdot \log^2(s\log(N/s))\log(N) + \log(\epsilon^{-1})\right].$$

Let $f \in C(\overline{D})$, $\theta \geq \|f\|_{L^2_\rho(D)}$ and $\zeta \geq 0$. Then

$$\mathbb{E}\|f - \check{f}\|^2_{L^2_\rho(D)} \lesssim \inf_{u\in\mathbb{U}} \|f-u\|^2_{L^2_\rho(D)} + \theta^2\epsilon + \zeta^2 + \frac{1}{m}\|e\|_2^2, \quad \textit{where } \check{f} = \min\{1, \theta/\|f\|_{L^2_\rho(D)}\}\hat{f}$$

for any ζ-minimizer $\hat{f}$ of (1.4), where $e = (e_i)_{i=1}^m$.

Proof. For the first result, observe that

$$K(\mathbb{U}'\cup\mathbb{W})(z) = \max\{K(\mathbb{U}')(z), K(\mathbb{W})(z)\}.$$

Consider the first term. Let $u = \sum_{i\in S} c_i\psi_i \in \mathbb{U}_{2s} = \mathbb{U}'$, where $|S| \leq 2s$. Then, by the Cauchy–Schwarz inequality and Parseval's identity,

$$|u(z)|^2 \leq 2s \max_{i=1,\dots,N} |\psi_i(z)|^2 \|u\|^2_{L^2_\rho(D)}.$$

Since $u \in \mathbb{U}'$ was arbitrary, this implies that $K(\mathbb{U}')(z) \leq 2s\widetilde{K}(\Psi)(z)$. Moreover, by definition of $\mathbb{W}$, we clearly have

$$K(\mathbb{W})(z) = 2s \max_{i=1,\dots,N} |\psi_i(z)|^2 = 2s\widetilde{K}(\Psi)(z).$$

The first result now follows immediately.

For the second result, we shall apply Theorem 3.4. First, notice that

$$\Phi(S(\mathbb{U}')\cup\mathbb{W};\bar{\mathcal{A}}) = \max\{\Phi(S(\mathbb{U}');\bar{\mathcal{A}}), \Phi(\mathbb{W};\bar{\mathcal{A}})\}.$$

Consider the second term. We have

$$\Phi(\mathbb{W};\bar{\mathcal{A}}) = 2s \operatorname*{ess\,sup}_{z\sim\rho} \max\left\{ \frac{|\psi_i(z)|^2}{\nu(z)} : i = 1,\dots,N \right\} = 2s \operatorname*{ess\,sup}_{z\sim\rho} \left\{ \widetilde{K}(\Psi)(z)/\nu(z) \right\}.$$

We also know from (4.3) and the first result that

$$\Phi(S(\mathbb{U}');\bar{\mathcal{A}}) \leq 2s \operatorname*{ess\,sup}_{z\sim\rho} \left\{ \widetilde{K}(\Psi)(z)/\nu(z) \right\}.$$

Therefore, using the fact that $\nu = \mathrm{d}\mu/\mathrm{d}\rho$ and the definition of μ, we get

$$\Phi(S(\mathbb{U}')\cup\mathbb{W};\bar{\mathcal{A}}) = 2s \operatorname*{ess\,sup}_{z\sim\rho} \left\{ \widetilde{K}(\Psi)(z)/\nu(z) \right\} = 2s\theta(\Psi).$$

To apply Theorem 3.4 we also need to estimate the term $\gamma(\mathbb{U}';\bar{\mathcal{A}})$. However, notice from the above derivation that

$$\Phi(S(\mathbb{U}'-\mathbb{U}');\bar{\mathcal{A}}) = \Phi(S(\mathbb{U}_{4s});\bar{\mathcal{A}}) \leq 4s\theta(\Psi).$$

Therefore, recalling that our main results hold when the variation constant is replaced by an upper bound, we deduce that we may take $\gamma(\mathbb{U}';\bar{\mathcal{A}}) \leq 2$ in this instance. We now apply Theorem 3.4, recalling that $n = s$, $d = \binom{N}{s}$ and $M = 4N$ in this case (see Appendix B). □

This result describes an active learning strategy for sparse regression (previously derived in (Adcock et al., 2022d)) that is optimal up to the term $\theta(\Psi)$, since the measurement condition scales linearly in s up to log factors. As discussed in (Adcock et al., 2022d), one can numerically estimate this factor. In cases such as sparse polynomial regression, it is typically small (see the next subsection). However, whether it is possible to derive an active learning strategy that avoids this factor is an open problem.

C.5. The numerical example in Fig. 1

In Fig. 1 we consider a sparse regression problem where $D = \mathbb{R}^k$ and ρ is the Gaussian measure, i.e., $\mathrm{d}\rho(z) = (2\pi)^{-k/2}\mathrm{e}^{-\|z\|_2^2/2}\,\mathrm{d}x$. We now describe this experiment in detail.

First, we discuss the construction of the orthonormal system Ψ. Let $\{H_n\}_{n\in\mathbb{N}_0}$ denote the standard Hermite polynomials, defined by the recurrence relation $H_0(z) = 1$, $H_1(z) = 2z$ and

$$H_{n+1}(z) = 2zH_n(z) - 2nH_{n-1}(z), \quad n = 1, 2, \ldots.$$

We define the orthonormal polynomials

$$\phi_n(z) = \frac{1}{\sqrt{2^n n!}} H_n(z/\sqrt{2}), \quad n = 0, 1, \ldots.$$

When $k \geq 2$, we construct an orthonormal basis via tensorization. Let

$$\varphi_\nu(z) = \phi_{\nu_1}(x_1) \times \cdots \times \phi_{\nu_k}(x_k), \quad x = (x_i)_{i=1}^k \in \mathbb{R}^k,\ \nu = (\nu_i)_{i=1}^k \in \mathbb{N}_0^k.$$

The set $\{\varphi_\nu\}_{\nu\in\mathbb{N}_0^k}$ forms an orthonormal basis of $L^2_\rho(\mathbb{R}^k)$. We truncate this basis as follows. Given $n \in \mathbb{N}_0$, we consider the total degree index set

$$\Lambda = \{\nu = (\nu_i)_{i=1}^k : \nu_1 + \cdots + \nu_k \leq n\}.$$

Let $N = |\Lambda|$, $\{\nu_1, \ldots, \nu_N\}$ be an enumeration of the multi-indices in Λ and define the orthonormal system $\Psi = \{\psi_i\}_{i=1}^N$ by $\psi_i = \varphi_{\nu_i}$, $i = 1, \ldots, N$. In our experiments, we consider $k = 2$ and $n = 20$, which gives $N = 231$.

To facilitate computations, we use a finite grid, described as follows. For $K \in \mathbb{N}$, we draw $z_1, \ldots, z_K \sim_{\text{i.i.d.}} \rho$. Note that this grid is drawn once, prior to any other subsequent computations. In our experiments, we use $K = 100,000$. Given these points, we replace ρ by the discrete measure $\bar{\rho} = \frac{1}{K}\sum_{i=1}^K \delta_{z_i}$. This means that the Christoffel sampling measure $\bar{\mu}$ defined in (C.6) also becomes a discrete measure, given by $\bar{\mu} = \frac{1}{K}\sum_{i=1}^K \widetilde{K}(\Psi)(z_i)\delta_{z_i}/\theta(\Psi)$, where $\theta(\Psi) = \frac{1}{K}\sum_{i=1}^K \widetilde{K}(\Psi)(z_i)$.

The middle and right panels of Fig. 1 show *phase transition* portraits for this problem. Phase transition portraits are standard tools in compressed sensing to empirically investigate the performance of different sampling strategies (Monajemi et al., 2013; Donoho & Tanner, 2009). We construct these portraits as follows. For each $1 \leq m \leq N$, we first generate a set of m sample points according to the inactive learning measure $\bar{\rho}$ or the Christoffel sampling measure $\bar{\mu}$ defined above. Then, for each $1 \leq s \leq m$, we proceed as follows. For each trial $t = 1, \ldots, T$, we generate a random s-sparse vector $c = (c_i)_{i=1}^N$. Here, the locations of s nonzero entries are chosen uniformly and randomly and the nonzero coefficients are drawn independently from the standard normal distribution. We define the unknown function $f = \sum_{i=1}^N c_i\psi_i$ and sample it at the z_i, giving values $(f(z_i))_{i=1}^m$. We then compute its reconstruction $\hat{f}$ and calculate the relative error $\|f - \hat{f}\|_{L^2_{\bar{\rho}}(D)}/\|f\|_{L^2_{\bar{\rho}}(D)}$. If this error is below a tolerance ϵ_{tol} we declare the recovery successful. Otherwise it is unsuccessful. We then repeat this process for each of the T trials. Next, compute the proportion of successful trials, i.e., the empirical probability of successful recovery for the given values of m and s. In our experiments, we use $T = 50$ trials and set $\epsilon_{\mathsf{tol}} = 10^{-2}$. Moreover, since solving (1.3) is NP-hard for sparse regression problems, we follow a standard approach in compressed sensing and solve a convex ℓ^1-minimization problem instead. Specifically, we solve the basis pursuit problem (Adcock & Hansen, 2021, §5.4.2) using the software package SPGL1 (van den Berg & Friedlander, 2008). The code for this experiment and the experiment for Fig. 2 can be found at `https://github.com/JMcardenas/CS4ML`.

Returning to the discussion in the previous subsection, we can compute the constant $\theta(\Psi)$ for this experiment. In this case, it is $\theta(\Psi) \approx 9.7323$. In comparison, the corresponding constant for inactive learning, which is given by

$$\vartheta(\Psi) = \operatorname*{ess\,sup}_{z\sim\bar{\rho}} \widetilde{K}(\Psi)(z)$$

is over a thousand times large, specifically, $\vartheta(\Psi) \approx 11,613$. This large difference between $\theta(\Psi)$ and $\vartheta(\Psi)$ provides further theoretical credence to the significant performance gain witnessed in Fig. 1.

D. Application to compressed sensing with generative models: additional discussion and examples

In this section, we provide some additional discussion on the application to compressed sensing with generative models considered in §5.

D.1. Proof of Corollary 5.1

Proof of Corollary 5.1. The set $\mathbb{U}' = \mathbb{U} - \mathbb{U} = \mathrm{Ran}(G) - \mathrm{Ran}(G)$ is a cone by construction and satisfies $\mathbb{U}' \subseteq \Delta(\mathbb{U}')$, where the latter is a union of d subspaces of dimension at most $2n$. We now apply Theorem 3.3, and specifically condition (a), with $\mathbb{V} = \Delta(\mathbb{U}')$. □

D.2. Recovery guarantees for subsampled unitary matrices

We now consider the application of Corollary 5.1 to the case of subsampled unitary matrices considered in Example 2.4, the latter being a special case of our general framework. This case was previously considered in (Berk et al., 2023a;b) and we adopt similar notation and terminology. We now show how Corollary 5.1 implies the results of (Berk et al., 2023a;b). However, we also extend and improve this work in several ways:

- We improve over (Berk et al., 2023a, Thm. 1) and (Berk et al., 2023b, Thm. 2.1) by requiring the 'local coherences' to be evaluated over a smaller set.
- We provide error bounds in expectation instead of probability.

Following (Berk et al., 2023b), we first introduce the concept of *local coherences*. As in (Berk et al., 2023b, Defn. 1.4), we say that the *local coherences* of U with respect a set $\mathbb{V} \subseteq \mathbb{C}^N$ are the entries of the vector $\sigma = (\sigma_i)_{i=1}^N$, where

$$\sigma_i = \sup_{\substack{v \in \mathbb{V} \\ \|v\|_{\ell^2}=1}} |u_i^* v|, \quad i = 1, \dots, N. \tag{D.1}$$

(Note that (Berk et al., 2023a, Defn. 3) uses the notation α. We use σ as α is already used in the definition of nondegeneracy.) We now show how the local coherence relates to a special case of the variation (Definition 3.1). Let $\bar{\mathcal{A}}$ be the isotropic family of the subsampled unitary matrix model of Example 2.4. Then $\Phi(S(\mathbb{V}); \bar{\mathcal{A}})$ is readily seen to be

$$\Phi(S(\mathbb{V}); \bar{\mathcal{A}}) = \max_{i=1,\dots,N} \sup \left\{ \frac{|u_i^* v|^2}{\pi_i \|v\|_{\ell^2}^2} : v \in \mathbb{V},\ v \neq 0 \right\} = \max_{i=1,\dots,N} \left(\frac{\sigma_i^2}{\pi_i} \right). \tag{D.2}$$

Using this and Corollary 5.1, we deduce the following.

Corollary D.1 (Generative models with subsampled unitary matrices)**.** *Consider the setup of Corollary 5.1 with the randomly subsampled unitary matrix model of Example 2.4. Let $\sigma = (\sigma_i)_{i=1}^N$ be the local coherences of U with respect to $\mathbb{U} - \mathbb{U}$ and $\tilde{\sigma} = (\tilde{\sigma}_i)_{i=1}^N$ be the local coherences of U with respect to $\Delta(\mathbb{U} - \mathbb{U})$ (see (5.2)). Then the measurement condition (5.3) is implied by*

$$m \gtrsim C \cdot n \cdot \left(\sum_{i=1}^{\ell-1} \log(2\mathrm{e}p_i/n) + \log(2/\epsilon) + \log(\widetilde{C}/C) \right), \tag{D.3}$$

where $C = \max_{i=1,\dots,N} \left(\frac{\sigma_i^2}{\pi_i} \right)$ and $\widetilde{C} = \max_{i=1,\dots,N} \left(\frac{\tilde{\sigma}_i^2}{\pi_i} \right)$. It also implied by the condition

$$m \gtrsim \widetilde{C} \cdot n \cdot \left(\sum_{i=1}^{\ell-1} \log(2\mathrm{e}p_i/n) + \log(2/\epsilon) \right). \tag{D.4}$$

Proof. From (D.2), we see that

$$\Phi(S(\mathbb{U} - \mathbb{U}); \bar{\mathcal{A}}) = \max_{i=1,\dots,N} \left(\frac{\sigma_i^2}{\pi_i} \right) = C, \quad \Phi(\Delta(\mathbb{U} - \mathbb{U}); \bar{\mathcal{A}}) = \max_{i=1,\dots,N} \left(\frac{\tilde{\sigma}_i^2}{\pi_i} \right) = \widetilde{C}.$$

That (D.3) implies (5.3) now follows immediately. Next, we use the fact that $\widetilde{C} \geq C$ (since $\mathbb{U} - \mathbb{U} \subseteq \Delta(\mathbb{U} - \mathbb{U})$) and the inequality $\log(x) \leq x$ to show that (D.4) implies (D.3), and therefore (5.3), as required. □

Using this result, we can now derive optimized variable-density sampling strategies based on the local coherences. We do this by choosing the probabilities π_i to minimize the factors appearing in the measurement conditions in the previous corollary. This strategy was first proposed in (Adcock et al., 2023) without theoretical guarantees, then developed in (Berk et al., 2023b) with theoretical guarantees.

Corollary D.2 (Optimal variable-density sampling for generative models)**.** *Consider the setup of the previous corollary. If*

$$\pi_i = \frac{\sigma_i^2}{\|\sigma\|_{\ell^2}^2}, \quad i = 1, \dots, N,$$

then the measurement condition (5.3) *is implied by*

$$m \gtrsim \|\sigma\|_{\ell^2}^2 \cdot n \cdot \left(\sum_{i=1}^{\ell-1} \log(2\mathrm{e}p_i/n) + \log(2/\epsilon) + \max_{i=1,\dots,N} \log(\tilde{\sigma}_i/\sigma_i) \right) \tag{D.5}$$

and if

$$\pi_i = \frac{\tilde{\sigma}_i^2}{\|\tilde{\sigma}\|_{\ell^2}^2}, \quad i = 1, \dots, N,$$

then the measurement condition (5.3) *is implied by*

$$m \gtrsim \|\tilde{\sigma}\|_{\ell^2}^2 \cdot n \cdot \left(\sum_{i=1}^{\ell-1} \log(2\mathrm{e}p_i/n) + \log(2/\epsilon) \right). \tag{D.6}$$

These two results are very similar to those in (Berk et al., 2023b). The second conditions (D.4) and (D.6) are identical to (Berk et al., 2023b, Thms. A2.1 & 2.1), respectively. In the first conditions (D.3) and (D.5) we obtain a small improvement over (Berk et al., 2023b, Thms. A2.1 & 2.1). Specifically, the measurement condition is primarily influenced by the local coherences σ_i taken over $\mathbb{U} - \mathbb{U} = \mathrm{Ran}(G) - \mathrm{Ran}(G)$, with the local coherences $\tilde{\sigma}_i$, which are taken over the piecewise linear expansion set $\Delta(\mathbb{U} - \mathbb{U})$ only entering logarithmically into the condition. This has relevance to practice, since the σ_i can be empirically estimated more efficiently than the $\tilde{\sigma}_i$. This was done in (Adcock et al., 2023) for MRI reconstruction using generative models and later in (Berk et al., 2023b). As shown in these works, this can lead to substantial benefits over uniform random subsampling (this is analogous to the discussion of active versus inactive learning in the previous section).

E. Proofs of the main theorems

In this section we prove Theorems 3.3 and 3.4. The proofs adopt similar ideas to those used in classical compressed sensing (see, e.g., (Foucart & Rauhut, 2013, Chpt. 12) or (Adcock & Hansen, 2021, Chpt. 13)). In particular, they rely on Dudley's inequality, Maurey's lemma and a version of Talagrand's theorem. Our main innovations involve the significant generalization of these arguments to, firstly, much broader classes of sampling problems (i.e., not just linear functionals of finite vectors), and secondly, to arbitrary model classes, rather than classes of (structured) sparse vectors. Our results broaden and strengthen recent results in the active learning context found in (Adcock et al., 2023) and (Eigel et al., 2022). In particular, (Adcock et al., 2023) assumes a union-of-subspaces model and then uses matrix Chernoff-type estimates. This is similar (although less general in terms of the type of measurements allowed) to our condition (b) in Theorems 3.3. Besides the generalization in terms of the measurements, our main effort is to derive the stronger condition (a) in this theorem. The work (Eigel et al., 2022) makes very few assumptions on $\mathbb{U}$, then uses Hoeffding's inequality and covering number arguments. As noted in (Adcock et al., 2023, §A.3) the trade-off for this high level of generality is weaker theoretical guarantees.

E.1. Additional notation

We first require additional notation. Let $(\overline{\mathbb{Y}}, \langle \cdot, \cdot \rangle_{\overline{\mathbb{Y}}})$ be the Hilbert space defined as the direct sum of the $\mathbb{Y}_i$, i.e.,

$$\overline{\mathbb{Y}} = \mathbb{Y}_1 \oplus \cdots \oplus \mathbb{Y}_m.$$

Next, let $\bar{\mathcal{A}}$ be the distribution of bounded linear operators in $\mathcal{B}(\mathbb{X}_0, \overline{\mathbb{Y}})$ induced by the family $\{\mathcal{A}_i\}_{i=1}^m$. In other words, $\bar{A} \sim \bar{\mathcal{A}}$ if

$$\bar{A}(x) = \frac{1}{\sqrt{m}}(A_1(x), \dots, A_m(x)),$$

where the A_i are independent with $A_i \sim \mathcal{A}_i$ for each i (we include the factor $1/\sqrt{m}$ for convenience). With this notation, observe that nondegeneracy (1.1) is equivalent to

$$\alpha\|x\|_{\mathbb{X}}^2 \leq \mathbb{E}_{\bar{A}\sim\bar{\mathcal{A}}}\|\bar{A}(x)\|_{\overline{\mathbb{Y}}}^2 \leq \beta\|x\|_{\mathbb{X}}^2, \quad \forall x \in \mathbb{X}_0 \tag{E.1}$$

and the least-squares problem (1.3) is equivalent to

$$\hat{x} \in \operatorname*{argmin}_{u\in\mathbb{U}} \|\bar{b} - \bar{A}(u)\|_{\overline{\mathbb{Y}}}^2, \tag{E.2}$$

where $\bar{b} = \frac{1}{\sqrt{m}}(b_1, \ldots, b_m) \in \overline{\mathbb{Y}}$. For convenience, we also write $\bar{e} = \frac{1}{\sqrt{m}}(e_1, \ldots, e_m) \in \overline{\mathbb{Y}}$.

E.2. Empirical nondegeneracy

Consider a realization of the A_i. We say that *empirical nondegeneracy* holds over a set $\mathbb{U} \subseteq \mathbb{X}_0$ with constants $0 < \alpha' \leq \beta' < \infty$ if

$$\alpha'\|u\|_{\mathbb{X}}^2 \leq \frac{1}{m}\sum_{i=1}^{m} \|A_i(u)\|_{\mathbb{Y}_i}^2 \leq \beta'\|u\|_{\mathbb{X}}^2, \quad \forall u \in \mathbb{U}. \tag{E.3}$$

Using the notation introduced above this can be equivalently written as

$$\alpha'\|v\|_{\mathbb{X}}^2 \leq \|\bar{A}(v)\|_{\overline{\mathbb{Y}}}^2 \leq \beta'\|v\|_{\mathbb{X}}^2, \quad \forall v \in \mathbb{V}.$$

Note that in the classical compressed sensing setup (see Example 2.3), this equivalent to the measurement matrix A satisfying the Restricted Isometry Property (RIP) (Foucart & Rauhut, 2013, Chpt. 6). As the following lemma shows, empirical nondegeneracy is crucial in establishing a recovery guarantee in the general case.

Lemma E.1. *Let $A_i \in \mathcal{B}(\mathbb{X}_0, \mathbb{Y}_i)$, $i = 1, \ldots, m$, be such that (E.3) holds over the set $\mathbb{U} - \mathbb{U}$ with constants $0 < \alpha' \leq \beta' < \infty$. Let $x \in \mathbb{X}_0$, $\zeta \geq 0$ and $\hat{x} \in \mathbb{U}$ be a ζ-minimizer of (1.3) based on noisy measurements (1.2). Then*

$$\|x - \hat{x}\|_{\mathbb{X}} \leq \inf_{u\in\mathbb{U}} \left\{ \frac{2}{\sqrt{\alpha'}}\|\bar{A}(x-u)\|_{\overline{\mathbb{Y}}} + \|x-u\|_{\mathbb{X}} \right\} + \frac{2}{\sqrt{\alpha'}}\|\bar{e}\|_{\overline{\mathbb{Y}}} + \frac{\zeta}{\sqrt{\alpha'}}.$$

Although the proof is straightforward, we include it for completeness.

Proof. Let $u \in \mathbb{U}$. Then

$$\begin{aligned}
\|\hat{x} - x\|_{\mathbb{X}} &\leq \|\hat{x} - u\|_{\mathbb{X}} + \|x - u\|_{\mathbb{X}} \\
&\leq 1/\sqrt{\alpha'}\|\bar{A}(\hat{x} - u)\|_{\mathbb{X}} + \|x - u\|_{\mathbb{X}} \\
&\leq 1/\sqrt{\alpha'}\|\bar{A}(\hat{x}) - \bar{b}\|_{\overline{\mathbb{Y}}} + 1/\sqrt{\alpha'}\|\bar{b} - \bar{A}(u)\|_{\overline{\mathbb{Y}}} + \|x - u\|_{\mathbb{X}} \\
&\leq \zeta/\sqrt{\alpha'} + 2/\sqrt{\alpha'}\|\bar{A}(u) - \bar{b}\|_{\overline{\mathbb{Y}}} + \|x - u\|_{\mathbb{X}} \\
&\leq \zeta/\sqrt{\alpha'} + 2/\sqrt{\alpha'}\|\bar{A}(u - x)\|_{\overline{\mathbb{Y}}} + 2/\sqrt{\alpha'}\|\bar{e}\|_{\overline{\mathbb{Y}}} + \|x - u\|_{\mathbb{X}},
\end{aligned}$$

as required. □

Notice that this result in fact only requires the lower inequality in (E.3). The upper inequality will be used later in the derivation of the error bound in expectation. Consequently, the remainder of the proofs are devoted to deriving measurement conditions under which (E.3) holds, then using this and the above lemma to derive error bounds in expectation.

E.3. Measurement conditions for empirical nondegeneracy

Theorem E.2. *Consider the setup of §1.1. Let $0 < \delta, \epsilon < 1$ and $\mathbb{U} \subseteq \mathbb{X}_0$ be such that*

(i) $\mathbb{U}$ is a cone, and

(ii) $\mathbb{U} \subseteq \mathbb{V}_1 \cup \cdots \cup \mathbb{V}_d =: \mathbb{V}$, where each $\mathbb{V}_i \subseteq \mathbb{X}_0$ is a subspace of dimension at most n.

Suppose that either

(a) $m \gtrsim \delta^{-2} \cdot \alpha^{-1} \cdot \Phi(S(\mathbb{U}); \bar{\mathcal{A}}) \cdot \big[\log(2d/\epsilon) + n \log(2\gamma(\mathbb{U}; \bar{\mathcal{A}}))\big]$, *where*

$$\gamma(\mathbb{U}; \bar{\mathcal{A}}) = \frac{\min\{\Phi(S(\mathbb{V}); \bar{\mathcal{A}}), \Phi(S(\mathbb{U} - \mathbb{U}); \bar{\mathcal{A}})\}}{\Phi(S(\mathbb{U}); \bar{\mathcal{A}})}, \tag{E.4}$$

(b) $m \gtrsim \delta^{-2} \cdot \alpha^{-1} \cdot \Phi(S(\mathbb{U} - \mathbb{U}); \bar{\mathcal{A}}) \cdot [\log(2d/\epsilon) + n]$,

(c) or $m \gtrsim \delta^{-2} \cdot \alpha^{-1} \cdot \Phi(S(\mathbb{V}); \bar{\mathcal{A}}) \cdot \log(2nd/\epsilon)$.

Then, with probability at least $1 - \epsilon$, (E.3) *holds for* $\mathbb{U}$ *with constants* $\alpha' = (1 - \delta)\alpha$ *and* $\beta' = (1 + \delta)\beta$.

Note that these first two theorems will be used to establish Theorem 3.3. We split them into two statements as the proof techniques are quite different. For Theorem 3.4, we will use the following result.

Theorem E.3. *Consider the setup of §1.1. Let* $0 < \delta, \epsilon < 1$ *and* $\mathbb{U} \subseteq \mathbb{X}_0$ *be such that assumptions (i) and (ii) of Theorem E.2 hold, and also that*

(iii) $\{u \in \mathbb{U} : \|u\|_{\mathbb{X}} \leq 1\} \subseteq \mathrm{conv}(\mathbb{W})$, *where* $\mathbb{W}$ *is a finite set of size* $|\mathbb{W}| = M$.

Suppose that either

(a) $m \gtrsim \delta^{-2} \cdot \alpha^{-1} \cdot \Phi(S(\mathbb{U}) \cup \mathbb{W}; \bar{\mathcal{A}}) \cdot L$, *where*

$$L = \log\big(2\Phi(S(\mathbb{U}) \cup \mathbb{W}; \bar{\mathcal{A}})/\alpha\big) \cdot \big[\log\big(2\gamma(\mathbb{U}; \bar{\mathcal{A}})\big) + \log(2M) \cdot \log^2(\log(2d) + n)\big] + \log(\epsilon^{-1})$$

and $\gamma(\mathbb{U}; \bar{\mathcal{A}})$ *is as in* (E.4)*;*

(b) $m \gtrsim \delta^{-2} \cdot \alpha^{-1} \cdot \Phi(S(\mathbb{U} - \mathbb{U}) \cup \mathbb{W}; \bar{\mathcal{A}}) \cdot L$, *where*

$$L = \log\big(2\Phi(S(\mathbb{U}) \cup \mathbb{W}; \bar{\mathcal{A}})/\alpha\big) \cdot \log(2M) \cdot \log^2(\log(2d) + n) + \log(\epsilon^{-1});$$

or

(c) $m \gtrsim \delta^{-2} \cdot \alpha^{-1} \cdot \Phi(S(\mathbb{V}) \cup \mathbb{W}; \bar{\mathcal{A}}) \cdot L$, *where*

$$L = \log\big(2\Phi(S(\mathbb{V}) \cup \mathbb{W}; \bar{\mathcal{A}})/\alpha\big) \cdot \log(2M) \cdot \log^2(\log(2d) + n) + \log(\epsilon^{-1}).$$

Then, with probability at least $1 - \epsilon$, (E.3) *holds for* $\mathbb{U}$ *with constants* $\alpha' = (1 - \delta)\alpha$ *and* $\beta' = (1 + \delta)\beta$.

We now prove these results.

E.3.1. Setup

Let $\mathbb{U} \subseteq \mathbb{X}$ be such that assumption (ii) of Theorem E.2 holds. As per the discussion §3.2, we will not assume that assumption (i) holds for the moment.

Nondegeneracy (1.1) implies that the quantity

$$|||x|||_{\mathbb{X}} = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{A_i \sim \mathcal{A}_i} \|A_i(x)\|_{\mathbb{Y}_i}^2} = \sqrt{\mathbb{E}_{\bar{A} \sim \bar{\mathcal{A}}} \|\bar{A}(x)\|_{\mathbb{Y}}^2}, \quad x \in \mathbb{X}_0,$$

defines a norm on $\mathbb{X}_0$ that is equivalent to $\|\cdot\|_{\mathbb{X}}$. Let

$$\widetilde{S}(\mathbb{U}) = \{u/|||u|||_{\mathbb{X}} : u \in \mathbb{U} \backslash \{0\}\}.$$

Then, for empirical degeneracy (E.3) to hold with constants $\alpha' = (1 - \delta)\alpha$ and $\beta' = (1 + \delta)\beta$, it suffices to show that the random variable

$$\delta_{\mathbb{U}} = \sup_{u \in \widetilde{S}(\mathbb{U})} \left| \frac{1}{m}\sum_{i=1}^{m} \|A_i(u)\|_{\mathbb{Y}_i}^2 - 1 \right| \tag{E.5}$$

satisfies $\delta_{\mathbb{U}} \leq \delta$ with probability at least $1 - \epsilon$. Following a standard route, we show this by first bounding the expectation of $\delta_{\mathbb{U}}$ and then by estimating the probability it exceeds δ.

E.3.2. EXPECTATION BOUNDS: SETUP

Recall that a Rademacher random variable is a random variable that takes the values $+1$ and -1 with probability $\frac{1}{2}$.

Lemma E.4. *Let $\{\epsilon_i\}_{i=1}^m$ be independent Rademacher random variables that are also independent of the random variables $\{A_i\}_{i=1}^m$. Then random variable (E.5) satisfies*

$$\mathbb{E}_{\bar{\mathcal{A}}}(\delta_{\mathbb{U}}) \leq 2m^{-1}\mathbb{E}_{\bar{\mathcal{A}}}\mathbb{E}_{\bar{\epsilon}} \sup_{u\in\widetilde{S}(\mathbb{U})} \left| \sum_{i=1}^m \epsilon_i \|A_i(u)\|_{\mathbb{Y}_i}^2 \right|.$$

Here $\mathbb{E}_{\bar{\mathcal{A}}}$ denotes expectation with respect to the variables $\{A_i\}_{i=1}^m$ and $\mathbb{E}_{\bar{\epsilon}}$ denotes expectation with respect to the $\{\epsilon_i\}_{i=1}^m$.

Proof. Since

$$1 = \|\!|u|\!\|_{\mathbb{X}}^2 = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{A_i\sim\mathcal{A}_i}\|A_i(u)\|_{\mathbb{Y}_i}^2, \quad \forall u \in \widetilde{S}(\mathbb{U}), \tag{E.6}$$

we have

$$\delta_{\mathbb{U}} = \sup_{u\in\widetilde{S}(\mathbb{U})} \left| \frac{1}{m}\sum_{i=1}^m \left(\|A_i(u)\|_{\mathbb{Y}_i}^2 - \mathbb{E}_{A_i\sim\mathcal{A}_i}\|A_i(u)\|_{\mathbb{Y}_i}^2 \right) \right| \quad \text{a.s.} \tag{E.7}$$

Assumption (ii) implies that $\mathbb{U} \subseteq \overline{\mathbb{V}}$, where $\overline{\mathbb{V}} \subseteq \mathbb{X}_0$ is the finite-dimensional vector space $\mathbb{V}_1 + \cdots + \mathbb{V}_d$. Since $\overline{\mathbb{V}}$ is a vector space, we also have $\widetilde{S}(\mathbb{U}) \subseteq \overline{\mathbb{V}}$. Consider $\overline{\mathbb{V}}$ as a Hilbert space with the inner product from $\mathbb{X}$. Then the map

$$B_i : (\overline{\mathbb{V}}, \|\cdot\|_{\mathbb{X}}) \to (\mathbb{Y}_i, \|\cdot\|_{\mathbb{Y}_i}),\ v \mapsto A_i(v),$$

is bounded. Indeed,

$$\|B_i(v)\|_{\mathbb{Y}_i} \leq \|A_i\|_{\mathbb{X}_0\to\mathbb{Y}_i}\|v\|_{\mathbb{X}_0} \leq c\|A_i\|_{\mathbb{X}_0\to\mathbb{Y}_i}\|v\|_{\mathbb{X}}, \quad \forall v \in \overline{\mathbb{V}}.$$

Here, in the first inequality we used the fact that the map $A_i \in \mathcal{B}(\mathbb{X}_0, \mathbb{Y}_i)$ is bounded by assumption and in the second step, we used the fact that $\overline{\mathbb{V}}$ is finite dimensional and all norms on finite-dimensional vector spaces are equivalent. Therefore, B_i has a unique bounded adjoint

$$B_i^* : (\mathbb{Y}_i, \|\cdot\|_{\mathbb{Y}_i}) \to (\overline{\mathbb{V}}, \|\cdot\|_{\mathbb{X}}).$$

In particular, we may write

$$\|A_i(v)\|_{\mathbb{Y}_i}^2 = \langle B_i^*B_i(v), v\rangle_{\mathbb{X}} = \langle U_i(v), v\rangle_{\mathbb{X}}, \quad \forall v \in \overline{\mathbb{V}},$$

where $U_i = B_i^*B_i$ is a random variable taking values in $\mathcal{B}(\overline{\mathbb{V}})$, the finite-dimensional vector space of bounded linear operators on $(\overline{\mathbb{V}}, \|\cdot\|_{\mathbb{X}})$. Using this we can write

$$\delta_{\mathbb{U}} = m^{-1} \sup_{u\in\widetilde{S}(\mathbb{U})} \left| \sum_{i=1}^m \langle U_i(u) - \mathbb{E}\left(U_i(u)\right), u\rangle_{\mathbb{X}} \right| = m^{-1} f\left(\sum_{i=1}^m (U_i - \mathbb{E}(U_i)) \right), \tag{E.8}$$

where f is the convex function

$$f : \mathcal{B}(\overline{\mathbb{V}}) \to \mathbb{R}, \quad U \mapsto \sup_{u\in\widetilde{S}(\mathbb{U})} |\langle U(u), u\rangle_{\mathbb{X}}|.$$

We now apply symmetrization (see, e.g., (Adcock & Hansen, 2021, Lem. 13.26)) to get

$$\mathbb{E}_{\bar{\mathcal{A}}}(\delta_{\mathbb{U}}) \leq m^{-1}\mathbb{E}_{\bar{\mathcal{A}}}\mathbb{E}_{\bar{\epsilon}} f\left(2\sum_{i=1}^m \epsilon_i U_i \right).$$

Now observe that

$$f\left(2\sum_{i=1}^m \epsilon_i U_i \right) = 2 \sup_{u\in\widetilde{S}(\mathbb{U})} \left| \left\langle \sum_{i=1}^m \epsilon_i U_i(u), u \right\rangle_{\mathbb{X}} \right| = 2 \sup_{u\in\widetilde{S}(\mathbb{U})} \left| \sum_{i=1}^m \epsilon_i \|A_i(u)\|_{\mathbb{Y}_i}^2 \right|.$$

This completes the proof. □

The next several result involves the covering number $\mathcal{N}$ (see, e.g., (Adcock & Hansen, 2021, Defn. 13.21)) and Dudley's inequality (see, e.g., (Adcock & Hansen, 2021, Thm. 13.25)).

Lemma E.5. *Fix a realization of the A_i, $i = 1, \dots, m$. Then*

$$\mathbb{E}_{\overline{\epsilon}} \sup_{u \in \widetilde{S}(\mathbb{U})} \left| \sum_{i=1}^{m} \epsilon_i \|A_i(u)\|_{\mathbb{Y}_i}^2 \right| \lesssim R_{p,\bar{\mathcal{A}}} \int_0^{\chi_q/2} \sqrt{\log(2\mathcal{N}(\widetilde{S}(\mathbb{U}), \|\cdot\|_{\bar{\mathcal{A}}}, t))}\, \mathrm{d}t \tag{E.9}$$

for all $1 \leq p < \infty$, where

$$R_{p,\bar{\mathcal{A}}} = \sup_{u \in \widetilde{S}(\mathbb{U})} \left(\sum_{i=1}^{m} \|A_i(u)\|_{\mathbb{Y}_i}^{2p} \right)^{\frac{1}{2p}}, \tag{E.10}$$

$1 < q \leq \infty$ is such that $1/p + 1/q = 1$,

$$\chi_q = \sqrt{\frac{m^{\frac{1}{q}} \Phi(S(\mathbb{U}), \bar{\mathcal{A}})}{\alpha}}, \tag{E.11}$$

and

$$\|x\|_{\bar{\mathcal{A}}} = \left(\sum_{i=1}^{m} \|A_i(x)\|_{\mathbb{Y}_i}^{2q} \right)^{\frac{1}{2q}}, \quad x \in \mathbb{X}_0. \tag{E.12}$$

Proof. The left-hand side of (E.9) is the expected value of the supremum of the absolute value of the Rademacher process $\{X_u : u \in \widetilde{S}(\mathbb{U})\}$, where

$$X_u = \sum_{i=1}^{m} \epsilon_i \|A_i(u)\|_{\mathbb{Y}_i}^2.$$

The corresponding pseudometric is

$$d(u, v) = \sqrt{\sum_{i=1}^{m} \left(\|A_i(u)\|_{\mathbb{Y}_i}^2 - \|A_i(v)\|_{\mathbb{Y}_i}^2 \right)^2}, \quad u, v \in \widetilde{S}(\mathbb{U}).$$

Dudley's inequality implies that

$$\mathbb{E}_{\overline{\epsilon}} \sup_{u \in \widetilde{S}(\mathbb{U})} \left| \sum_{i=1}^{m} \epsilon_i \|A_i(u)\|_{\mathbb{Y}_i}^2 \right| \lesssim \int_0^{\varpi/2} \sqrt{\log(2\mathcal{N}(\widetilde{S}(\mathbb{U}), d, z))}\, \mathrm{d}z, \tag{E.13}$$

where $\varpi = \sup_{u \in \widetilde{S}(\mathbb{U})} \sqrt{\mathbb{E}_{\overline{\epsilon}} |X_u|^2}$. The remainder of the proof involves upper bounding the pseudometric d and the constant ϖ.

First, for two sequences a_i, b_i, observe that

$$\begin{aligned} \sum_i (|a_i|^2 - |b_i|^2)^2 &= \sum_i (|a_i| + |b_i|)^2 (|a_i| - |b_i|)^2 \\ &\leq \left(\sum_i (|a_i| + |b_i|)^{2p} \right)^{\frac{1}{p}} \left(\sum_i (|a_i - b_i|)^{2q} \right)^{\frac{1}{q}} \end{aligned}$$

where q is such that $1/p + 1/q = 1$. Therefore

$$\sqrt{\sum_i (|a_i|^2 - |b_i|^2)^2} \leq 2 \max\{\|a\|_{2p}, \|b\|_{2p}\} \|a - b\|_{2q}.$$

We deduce that

$$d(u, v) \leq 2R_{p,\bar{\mathcal{A}}} \|u - v\|_{\bar{\mathcal{A}}}, \tag{E.14}$$

where R_p and $\|\cdot\|_{\bar{\mathcal{A}}}$ are as in (E.10) and (E.12), respectively.

Second, consider the term ϖ. Then, defining additionally $X_0 = 0$ (since $0 \notin \widetilde{S}(\mathbb{U})$), we see that

$$\varpi = \sup_{u \in \widetilde{S}(\mathbb{U})} d(u,0) \leq 2R_{p,\bar{\mathcal{A}}} \sup_{u \in \widetilde{S}(\mathbb{U})} \|u\|_{\bar{\mathcal{A}}}.$$

Now let $u = v/|||v|||_{\mathbb{X}} \in \widetilde{S}(\mathbb{U})$, where $v \in \mathbb{U}$. Then

$$\|A_i(v)\|_{\mathbb{Y}_i}^2 \leq \Phi(S(\mathbb{U}), \mathcal{A}_i)\|v\|_{\mathbb{X}}^2$$

Also, recall that $\sqrt{\alpha}\|x\|_{\mathbb{X}} \leq |||x|||_{\mathbb{X}} \leq \sqrt{\beta}\|x\|_{\mathbb{X}}$, $\forall x \in \mathbb{X}_0$. We deduce that

$$\|A_i(u)\|_{\mathbb{Y}_i}^2 \leq \Phi(S(\mathbb{U}); \mathcal{A}_i)/\alpha, \quad \forall u \in \widetilde{S}(\mathbb{U}). \tag{E.15}$$

This gives

$$\|u\|_{\bar{\mathcal{A}}} \leq \chi_q, \quad \forall u \in \widetilde{S}(\mathbb{U}).$$

Hence $\varpi \leq 2R_{p,\bar{\mathcal{A}}}\chi_q$.

We now combine this with (E.14) and standard properties of covering numbers to obtain

$$\mathbb{E}_{\bar{\epsilon}} \sup_{u \in \widetilde{S}(\mathbb{U})} \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \|A_i(u)\|_{\mathbb{Y}_i}^2 \right| \lesssim \int_0^{R_{p,\bar{\mathcal{A}}}\chi_q} \sqrt{\log(2\mathcal{N}(\widetilde{S}(\mathbb{U}), \|\cdot\|_{\bar{\mathcal{A}}}, z/2R_{p,\bar{\mathcal{A}}}))}\, \mathrm{d}z.$$

The result now follows from a change of variables. □

The next steps involves estimating the covering number. We do this in two ways via the following two lemmas. Their proofs rely on a number of standard properties of covering numbers. See, e.g., (Adcock & Hansen, 2021, §13.5.2).

Lemma E.6 (First covering number bound)**.** *Consider the setup of the previous lemma. Then*

$$\sqrt{\log(2\mathcal{N}(\widetilde{S}(\mathbb{U}), \|\cdot\|_{\bar{\mathcal{A}}}, t))} \leq \sqrt{\log(2d)} + \sqrt{n}\sqrt{\log\left(1 + 4\chi_q'/t\right)}, \tag{E.16}$$

where

(a) $\chi_q' = \sqrt{\frac{m^{\frac{1}{q}} \min\{\Phi(S(\mathbb{V});\bar{\mathcal{A}}), \Phi(S(\widetilde{S}(\mathbb{U}) - \widetilde{S}(\mathbb{U}));\bar{\mathcal{A}})\}}{\alpha}}$ *if* $\mathbb{U}$ *satisfies assumption (ii) of Theorem E.2; or*

(b) $\chi_q' = \sqrt{\frac{m^{\frac{1}{q}} \min\{\Phi(S(\mathbb{V});\bar{\mathcal{A}}), \Phi(S(\mathbb{U} - \mathbb{U});\bar{\mathcal{A}})\}}{\alpha}}$ *if* $\mathbb{U}$ *satisfies assumptions (i) and (ii) of Theorem E.2.*

Proof. Write $\chi_q' = \min\{\chi_{q,1}', \chi_{q,2}'\}$, where

$$\chi_{q,1}' = \begin{cases} \sqrt{\frac{m^{\frac{1}{q}} \Phi(S(\widetilde{S}(\mathbb{U}) - \widetilde{S}(\mathbb{U}));\bar{\mathcal{A}})}{\alpha}} & \text{case (a)} \\ \sqrt{\frac{m^{\frac{1}{q}} \Phi(S(\mathbb{U} - \mathbb{U}));\bar{\mathcal{A}})}{\alpha}} & \text{case (b)} \end{cases} \quad \text{and} \quad \chi_{q,2}' = \sqrt{\frac{m^{\frac{1}{q}} \Phi(S(\mathbb{V}));\bar{\mathcal{A}})}{\alpha}},$$

Then it suffices to prove the bound (E.16) first with $\chi_{q,1}'$ in place of χ_q' and then with $\chi_{q,2}'$ in place of χ_q'.

Step 1: showing (E.16) *with* $\chi_{q,1}'$ *in place of* χ_q'. By definition,

$$\|A_i(u - v)\|_{\mathbb{Y}_i}^2 \leq \Phi(S(\widetilde{S}(\mathbb{U}) - \widetilde{S}(\mathbb{U})); \mathcal{A}_i)\|u - v\|_{\mathbb{X}}^2, \quad \forall u, v \in \widetilde{S}(\mathbb{U}).$$

Therefore, the norm (E.12) satisfies

$$\|u - v\|_{\bar{\mathcal{A}}} \leq \sqrt{m^{\frac{1}{q}} \Phi(S(\widetilde{S}(\mathbb{U}) - \widetilde{S}(\mathbb{U})); \bar{\mathcal{A}})}\|u - v\|_{\mathbb{X}}, \quad \forall u, v \in \widetilde{S}(\mathbb{U}). \tag{E.17}$$

Notice that the right-hand side is equal to $\chi'_{q,1}\|u-v\|_{\mathbb{X}}$ in case (a). Now consider case (b). Note that $\widetilde{S}(\mathbb{U}) = \{u/|||u|||_{\mathbb{X}} : u \in \mathbb{U}\backslash\{0\}\} \subseteq \mathbb{U}$ whenever $\mathbb{U}$ is a cone. Therefore,

$$S(\widetilde{S}(\mathbb{U}) - \widetilde{S}(\mathbb{U})) \subseteq S(\mathbb{U}-\mathbb{U})$$

from which it follows that

$$\Phi(\widetilde{S}(\mathbb{U}) - \widetilde{S}(\mathbb{U}); \bar{\mathcal{A}}) \leq \Phi(S(\mathbb{U}-\mathbb{U}); \bar{\mathcal{A}}).$$

Therefore, the right-hand side of (E.17) can be bounded by $\chi'_{q,1}\|u-v\|_{\mathbb{X}}$ in case (b) as well. Using the equivalence between $\|\cdot\|_{\mathbb{X}}$ and $|||\cdot|||_{\mathbb{X}}$ once more, we deduce that

$$\|u-v\|_{\bar{\mathcal{A}}} \leq \chi'_{q,1}|||u-v|||_{\mathbb{X}}, \quad \forall u, v \in \widetilde{S}(\mathbb{U}), \tag{E.18}$$

in either case (a) or (b). Now, using standard properties of covering numbers, we get

$$\mathcal{N}(\widetilde{S}(\mathbb{U}), \|\cdot\|_{\bar{\mathcal{A}}}, t) \leq \mathcal{N}(\widetilde{S}(\mathbb{U}), \chi'_{q,1}|||\cdot|||_{\mathbb{X}}, t) = \mathcal{N}(\widetilde{S}(\mathbb{U}), |||\cdot|||_{\mathbb{X}}, t/\chi'_{q,1}).$$

Now, assumption (ii) of Theorem E.2 gives that $\widetilde{S}(\mathbb{U}) \subseteq \widetilde{B}(\mathbb{V}_1) \cup \cdots \cup \widetilde{B}(\mathbb{V}_d)$, where $\widetilde{B}(\mathbb{V}_i) = \{v_i \in \mathbb{V}_i : |||v_i|||_{\mathbb{X}} \leq 1\}$ is the unit ball of $(\mathbb{V}_i, |||\cdot|||_{\mathbb{X}})$. Using further properties of covering numbers and (E.18), we get

$$\begin{aligned}
\mathcal{N}(\widetilde{S}(\mathbb{U}), \|\cdot\|_{\bar{\mathcal{A}}}, t) &\leq \mathcal{N}(\widetilde{B}(\mathbb{V}_1) \cup \cdots \cup \widetilde{B}(\mathbb{V}_d), |||\cdot|||_{\mathbb{X}}, t/(2\chi'_{q,1})) \\
&\leq \sum_{i=1}^{d} \mathcal{N}(\widetilde{B}(\mathbb{V}_i), |||\cdot|||_{\mathbb{X}}, t/(2\chi'_{q,1})) \\
&\leq d\left(1 + 4\chi'_{q,1}/t\right)^n.
\end{aligned}$$

We now take the logarithm and square root, and using the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, $a, b \geq 0$. This gives (E.16) with $\chi'_{q,1}$ in place of χ'_q.

Step 2: showing (E.16) *with* $\chi'_{q,2}$ *in place of* χ'_q. We now switch the order of the above arguments. First, we write

$$\mathcal{N}(\widetilde{S}(\mathbb{U}), \|\cdot\|_{\bar{\mathcal{A}}}, t) \leq \sum_{i=1}^{d} \mathcal{N}(\widetilde{B}(\mathbb{V}_i), \|\cdot\|_{\bar{\mathcal{A}}}, t/2).$$

Now, since $\mathbb{V}_j$ is a subspace, we have

$$\|A_i(u-v)\|_{\mathbb{Y}_i}^2 \leq \Phi(S(\mathbb{V}_j); \mathcal{A}_i)\|u-v\|_{\mathbb{X}}^2 \leq \Phi(S(\mathbb{V}); \mathcal{A}_i)\|u-v\|_{\mathbb{X}}^2, \quad \forall u, v \in \mathbb{V}_j,$$

and therefore

$$\|u-v\|_{\bar{\mathcal{A}}} \leq \chi'_{q,2}|||u-v|||_{\mathbb{X}}, \qquad \forall u, v \in \mathbb{V}_i.$$

Since $\widetilde{B}(\mathbb{V}_i)$ is the unit ball of $\mathbb{V}_i$ with respect to $|||\cdot|||_{\mathbb{X}}$, we deduce that

$$\mathcal{N}(\widetilde{S}(\mathbb{U}), \|\cdot\|_{\bar{\mathcal{A}}}, t) \leq \sum_{i=1}^{d} \mathcal{N}(\widetilde{B}(\mathbb{V}_i), |||\cdot|||_{\mathbb{X}}, t/(2\chi'_{q,2})) \leq d(1 + 4\chi'_{q,2}/t)^n.$$

We now take the logarithm and square root of both sides, as before. □

To derive our second bound for the covering number makes we first need to establish a version of Khintchine's inequality in Hilbert spaces. For this, we need the following.

Lemma E.7 (Hoeffding's inequality in a Hilbert space)**.** *Let $X_1, \ldots, X_n$ be independent mean-zero random variables taking values in a separable Hilbert space $\mathbb{H}$ such that $\|X_i\|_{\mathbb{H}} \leq c/2$ and let $v = nc^2/4$. Then, for all $t \geq \sqrt{v}$,*

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} X_i\right\|_{\mathbb{H}} > t\right) \leq \mathrm{e}^{-(t-\sqrt{v})^2/(2v)}.$$

Note that this lemma is can be proved directly from McDiarmid's inequality. Using this, we now obtain the following.

Lemma E.8 (Khintchine's inequality in a Hilbert space)**.** *Let $x_1, \dots, x_n \in \mathbb{H}$, where $\mathbb{H}$ is a separable Hilbert space and $\epsilon_1, \dots, \epsilon_n$ be independent Rademacher random variables. Then*

$$\left(\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i x_i\right\|_{\mathbb{H}}^p\right)^{\frac{1}{p}} \lesssim \sqrt{p}\sqrt{n} \max_{i=1,\dots,n} \|x_i\|_{\mathbb{H}}, \quad \forall p \geq 1.$$

Proof. Let $Z = \|\sum_{i=1}^n \epsilon_i x_i\|_{\mathbb{H}}$ and note that our aim is to bound $(\mathbb{E}(Z^p))^{\frac{1}{p}}$. Now let $\gamma = 2\max_{i=1,\dots,n} \|x_i\|_{\mathbb{H}}$. Then, by the previous lemma.

$$\mathbb{P}\left(Z \geq t\right) \leq \exp\left(-\frac{(t-\sqrt{n}\gamma/2)^2}{n\gamma^2/2}\right)$$

whenever $t \geq \sqrt{n}\gamma/2$. In particular, if $t \geq \sqrt{n}\gamma$, then

$$\mathbb{P}\left(Z \geq t\right) \leq \exp\left(-\frac{t^2}{2n\gamma^2}\right).$$

Now fix $\sqrt{n}\gamma/2 \leq \tau < \infty$. Then (see, e.g., (Vershynin, 2018, Ex. 1.2.3))

$$\begin{aligned}
\mathbb{E}(Z^p) &= \int_0^\infty pt^{p-1}\mathbb{P}(Z \geq t)\,\mathrm{d}t \\
&\leq p\tau^p + \int_\tau^\infty pt^{p-1}\exp\left(-\frac{t^2}{2n\gamma^2}\right)\,\mathrm{d}t \\
&\leq p\tau^p + \int_0^\infty pt^{p-1}\exp\left(-\frac{t^2}{2n\gamma^2}\right)\,\mathrm{d}t \\
&= p\tau^p + (\sqrt{n}\gamma)^p \int_0^\infty pt^{p-1}\exp\left(-\frac{t^2}{2}\right)\,\mathrm{d}t \\
&= p\tau^p + \sqrt{\pi/2}(\sqrt{n}\gamma)^p p \int_{-\infty}^\infty |t|^{p-1}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)\,\mathrm{d}t \\
&= p\tau^p + \sqrt{\pi/2}(\sqrt{n}\gamma)^p p\mathbb{E}|X|^{p-1},
\end{aligned}$$

where $X \sim \mathcal{N}(0,1)$. Now choose $\tau = \sqrt{n}\gamma$ and take the pth root to obtain

$$(\mathbb{E}(Z^p))^{\frac{1}{p}} \lesssim p^{\frac{1}{p}}\sqrt{n}\gamma(1 + (\mathbb{E}|X|^{p-1})^{1/p}).$$

The moments $(\mathbb{E}|X|^q)^{1/q} \lesssim \sqrt{q}$ (see, e.g., (Vershynin, 2018, Ex. 2.5.1)). Using this and the fact that $p^{\frac{1}{p}} \lesssim 1$ for $p \geq 1$, we deduce the result. □

Lemma E.9 (Second covering number bound)**.** *Consider the setup of Lemma E.5 and suppose that assumption (iii) of Theorem E.3 holds. Then*

$$\sqrt{\log(\mathcal{N}(\widetilde{S}(\mathbb{U}), \|\cdot\|_{\bar{\mathcal{A}}}, t))} \lesssim \sqrt{\frac{qm^{\frac{1}{q}}\Phi(\mathbb{W};\bar{\mathcal{A}})\log(M)}{\alpha}} \cdot t^{-1}.$$

Proof. Let $v = u/\|u\|_{\mathbb{X}} \in \widetilde{S}(\mathbb{U})$. Then, by norm equivalence, $\|v\|_{\mathbb{X}} \leq 1/\sqrt{\alpha}$. Hence $v \in B(\mathbb{U})/\sqrt{\alpha}$, where $B(\mathbb{U}) = \{u \in \mathbb{U} : \|u\|_{\mathbb{X}} \leq 1\}$. We deduce that

$$\widetilde{S}(\mathbb{U}) \subseteq B(\mathbb{U})/\sqrt{\alpha} \subseteq \mathrm{conv}(\mathbb{W})/\sqrt{\alpha} = \mathrm{conv}(\mathbb{W}/\sqrt{\alpha}).$$

Therefore

$$\mathcal{N}(\widetilde{S}(\mathbb{U}), \|\cdot\|_{\bar{\mathcal{A}}}, t) \leq \mathcal{N}(\mathrm{conv}(\mathbb{W}/\sqrt{\alpha}), \|\cdot\|_{\bar{\mathcal{A}}}, t/2).$$

Maurey's lemma (see, e.g., (Adcock & Hansen, 2021, Lem. 13.30)) implies that

$$\log(\mathcal{N}(\widetilde{S}(\mathbb{U}), \|\cdot\|_{\bar{\mathcal{A}}}, t)) \lesssim (C/t)^2\log(M),$$

where C is such that

$$\mathbb{E}_{\bar{\epsilon}}\left\|\sum_{l=1}^{L}\epsilon_l\frac{w_l}{\sqrt{\alpha}}\right\|_{\bar{\mathcal{A}}}\leq C\sqrt{L},\quad \forall L\in\mathbb{N},\ w_1,\dots,w_L\in\mathbb{W}.$$

We now estimate C. Let $L\in\mathbb{N}$ and $w_1,\dots,w_L\in\mathbb{W}$. By Hölder's inequality and linearity,

$$\begin{aligned}\mathbb{E}_{\bar{\epsilon}}\left\|\sum_{l=1}^{L}\epsilon_l\frac{w_l}{\sqrt{\alpha}}\right\|_{\bar{\mathcal{A}}}&\leq\left(\mathbb{E}_{\bar{\epsilon}}\left\|\sum_{l=1}^{L}\epsilon_l\frac{w_l}{\sqrt{\alpha}}\right\|_{\bar{\mathcal{A}}}^{2q}\right)^{\frac{1}{2q}}\\&=\frac{1}{\sqrt{\alpha}}\left(\sum_{i=1}^{m}\mathbb{E}_{\bar{\epsilon}}\left\|\sum_{l=1}^{L}\epsilon_l A_i(w_l)\right\|_{\mathbb{Y}_i}^{2q}\right)^{\frac{1}{2q}}\\&\leq\frac{m^{\frac{1}{2q}}}{\sqrt{\alpha}}\max_{i=1}^{m}\left(\mathbb{E}_{\bar{\epsilon}}\left\|\sum_{l=1}^{L}\epsilon_l A_i(w_l)\right\|_{\mathbb{Y}_i}^{2q}\right)^{\frac{1}{2q}}.\end{aligned}$$

By Lemma E.8,

$$\left(\mathbb{E}_{\bar{\epsilon}}\left\|\sum_{l=1}^{L}\epsilon_l A_i(w_l)\right\|_{\mathbb{Y}_i}^{2q}\right)^{\frac{1}{2q}}\lesssim\sqrt{qL}\max_{l=1,\dots,L}\|A_i(w_l)\|_{\mathbb{Y}_i}\leq\sqrt{qL\Phi(\mathbb{W};\bar{\mathcal{A}})}.$$

We deduce that

$$\mathbb{E}_{\bar{\epsilon}}\left\|\sum_{l=1}^{L}\epsilon_l\frac{w_l}{\sqrt{\alpha}}\right\|_{\bar{\mathcal{A}}}\lesssim\sqrt{\frac{qm^{\frac{1}{q}}\Phi(\mathbb{W};\bar{\mathcal{A}})}{\alpha}}\sqrt{L}.$$

The result now follows from Maurey's lemma. □

E.3.3. Expectation bounds

We are now in a position to establish our first expectation bound.

Theorem E.10 (First expectation bound)**.** *Consider the setup of Section 1.1. Let $0<\delta<1$ and $\mathbb{U}\subseteq\mathbb{X}_0$ and suppose that*

$$m\gtrsim\delta^{-2}\cdot\alpha^{-1}\cdot\Phi(S(\mathbb{U});\bar{\mathcal{A}})\cdot\left[\log(2d)+n\log\left(2\left(1+\gamma(\mathbb{U};\bar{\mathcal{A}})\right)\right)\right],\tag{E.19}$$

where

(a) $\gamma(\mathbb{U};\bar{\mathcal{A}})=\frac{\min\{\Phi(S(\mathbb{V});\bar{\mathcal{A}}),\Phi(S(\widetilde{S}(\mathbb{U})-\widetilde{S}(\mathbb{U}));\bar{\mathcal{A}})\}}{\Phi(S(\mathbb{U});\bar{\mathcal{A}})}$ *if $\mathbb{U}$ satisfies assumption (ii) of Theorem E.2; or*

(b) $\gamma(\mathbb{U};\bar{\mathcal{A}})=\frac{\min\{\Phi(S(\mathbb{V});\bar{\mathcal{A}}),\Phi(S(\mathbb{U}-\mathbb{U});\bar{\mathcal{A}})\}}{\Phi(S(\mathbb{U});\bar{\mathcal{A}})}$ *if $\mathbb{U}$ satisfies assumptions (i) and (ii) of Theorem E.2.*

Then the random variable (E.5) satisfies $\mathbb{E}(\delta_{\mathbb{U}})\leq\delta$.

Proof. Lemmas E.4, E.5 and E.6 imply that

$$\begin{aligned}\mathbb{E}_{\bar{\mathcal{A}}}(\delta_{\mathbb{U}})&\lesssim m^{-1}\mathbb{E}_{\bar{\mathcal{A}}}(R_{p,\bar{\mathcal{A}}})\left(\chi_q\sqrt{\log(2d)}+\sqrt{n}\int_0^{\chi_q/2}\sqrt{\log\left(1+\frac{4\chi_q'}{t}\right)}\,\mathrm{d}t\right)\\&=m^{-1}\mathbb{E}_{\bar{\mathcal{A}}}(R_{p,\bar{\mathcal{A}}})\left(\chi_q\sqrt{\log(2d)}+\sqrt{n}\chi_q'\int_0^{\chi_q/(8\chi_q')}\sqrt{\log\left(1+1/s\right)}\,\mathrm{d}s\right),\end{aligned}$$

where χ_q is as in (E.11) and χ_q' is as in Lemma E.6. We now use (Foucart & Rauhut, 2013, Lem. C.9), to obtain

$$\mathbb{E}_{\bar{\mathcal{A}}}(\delta_{\mathbb{U}})\lesssim m^{-1}\mathbb{E}_{\bar{\mathcal{A}}}(R_{p,\bar{\mathcal{A}}})\chi_q\left(\sqrt{\log(2d)}+\sqrt{n}\sqrt{\log(\mathrm{e}(1+8\chi_q'/\chi_q))}\right).$$

Using (E.15) and the definition of $R_{p,\bar{\mathcal{A}}}$, we see that

$$R_{p,\bar{\mathcal{A}}} \leq m^{\frac{1}{2p}} \left(\frac{\Phi(S(\mathbb{U}); \bar{\mathcal{A}})}{\alpha} \right)^{\frac{1}{2}-\frac{1}{2p}} \sup_{u \in \widetilde{S}(\mathbb{U})} \left(\frac{1}{m} \sum_{i=1}^{m} \|A_i(u)\|_{\mathbb{Y}_i}^2 \right)^{\frac{1}{2p}}$$

Now consider the sum. Let $u \in \widetilde{S}(\mathbb{U})$. Then

$$\frac{1}{m} \sum_{i=1}^{m} \|A_i(u)\|_{\mathbb{Y}_i}^2 = \frac{1}{m} \sum_{i=1}^{m} \left(\|A_i(u)\|_{\mathbb{Y}_i}^2 - \mathbb{E}_{A_i \sim \mathcal{A}_i} \|A_i(u)\|_{\mathbb{Y}_i}^2 \right) + \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{A_i \sim \mathcal{A}_i} \|A_i(u)\|_{\mathbb{Y}_i}^2 \leq \delta_{\mathbb{U}} + 1,$$

where in the second step we used (E.7) and (E.6). Hence

$$R_{p,\bar{\mathcal{A}}} \leq m^{\frac{1}{2p}} \left(\frac{\Phi(S(\mathbb{U}); \bar{\mathcal{A}})}{\alpha} \right)^{\frac{1}{2}-\frac{1}{2p}} (\delta_{\mathbb{U}} + 1)^{\frac{1}{2p}}$$

Using the fact that $p \geq 1$ and the Cauchy–Schwarz inequality, we deduce that

$$\mathbb{E}_{\bar{\mathcal{A}}}(R_{p,\bar{\mathcal{A}}}) \leq m^{\frac{1}{2p}} \left(\frac{\Phi(S(\mathbb{U}); \bar{\mathcal{A}})}{\alpha} \right)^{\frac{1}{2}-\frac{1}{2p}} \sqrt{\mathbb{E}_{\bar{\mathcal{A}}}(\delta_{\mathbb{U}}) + 1}. \tag{E.20}$$

Combining this with the previous expression and using the definition of χ_q, we deduce that

$$\mathbb{E}_{\bar{\mathcal{A}}}(\delta_{\mathbb{U}}) \lesssim m^{\frac{1}{2p}-1} m^{\frac{1}{2q}} \left(\frac{\Phi(S(\mathbb{U}); \bar{\mathcal{A}})}{\alpha} \right)^{1-\frac{1}{2p}} \left(\sqrt{\log(2d)} + \sqrt{n} \sqrt{\log(\mathrm{e}(1 + 8\chi_q'/\chi_q))} \right) \sqrt{\mathbb{E}_{\bar{\mathcal{A}}}(\delta_{\mathbb{U}}) + 1}.$$

Now set $p = 1 + 1/\log(\lambda)$, where $\lambda = 1 + \Phi(S(\mathbb{U}); \bar{\mathcal{A}})/\alpha$ and notice that $q = 1 + \log(\lambda)$ in this case. Observe that

$$\left(\frac{\Phi(S(\mathbb{U}); \bar{\mathcal{A}})}{\alpha} \right)^{1-\frac{1}{2p}} \leq \left(\frac{\Phi(S(\mathbb{U}); \bar{\mathcal{A}})}{\alpha} \right)^{\frac{1}{2}} \lambda^{\frac{1}{2}-\frac{1}{2p}} \leq \left(\frac{\Phi(S(\mathbb{U}); \bar{\mathcal{A}})}{\alpha} \right)^{\frac{1}{2}} \sqrt{\mathrm{e}}.$$

Using this and the fact that $m^{\frac{1}{2p}-1} m^{\frac{1}{2q}} = m^{-1/2} (m/m)^{1/(2q)} \leq m^{-1/2}$ (since $m \leq m$), we deduce that

$$\mathbb{E}_{\bar{\mathcal{A}}}(\delta_{\mathbb{U}}) \lesssim m^{-1/2} \left(\frac{\Phi(S(\mathbb{U}); \bar{\mathcal{A}})}{\alpha} \right)^{\frac{1}{2}} \left(\sqrt{\log(2d)} + \sqrt{n} \sqrt{\log(\mathrm{e}(1 + 8\chi_q'/\chi_q))} \right) \sqrt{\mathbb{E}_{\bar{\mathcal{A}}}(\delta_{\mathbb{U}}) + 1}.$$

Hence, we see that $\mathbb{E}_{\bar{\mathcal{A}}}(\delta_{\mathbb{U}}) \leq \delta$, provided

$$m^{-1/2} \left(\frac{\Phi(S(\mathbb{U}); \bar{\mathcal{A}})}{\alpha} \right)^{\frac{1}{2}} \left(\sqrt{\log(2d)} + \sqrt{n} \sqrt{\log(2(1 + \chi_q'/\chi_q))} \right) \leq c\delta,$$

for suitably small constant $c > 0$. Rearranging and noticing that

$$\chi_q'/\chi_q = \sqrt{\gamma(\mathbb{U}; \bar{\mathcal{A}})}$$

now gives the result. □

Theorem E.11 (Second expectation bound)**.** *Consider the setup of Section 1.1. Let $0 < \delta < 1$ and $\mathbb{U} \subseteq \mathbb{X}_0$ satisfy assumption (iii) of Theorem E.3 and suppose that*

$$m \gtrsim \delta^{-2} \cdot \alpha^{-1} \cdot \Phi(S(\mathbb{U}) \cup \mathbb{W}; \bar{\mathcal{A}}) \cdot L_0 \cdot (L_1 + L_2) \cdot,$$

where

$$\begin{aligned} L_0 &= \log(2\Phi(S(\mathbb{U}) \cup \mathbb{W}; \bar{\mathcal{A}})/\alpha), \\ L_1 &= 1 + \log\left(1 + \gamma(\mathbb{U}; \bar{\mathcal{A}})(\log(2d) + n)\right), \\ L_2 &= \log(M) \log^2(\log(2d) + n), \end{aligned}$$

and

(a) $\gamma(\mathbb{U};\bar{\mathcal{A}}) = \frac{\min\{\Phi(S(\mathbb{V});\bar{\mathcal{A}}),\Phi(S(\widetilde{S}(\mathbb{U})-\widetilde{S}(\mathbb{U}));\bar{\mathcal{A}})\}}{\Phi(S(\mathbb{U});\bar{\mathcal{A}})}$ *if* $\mathbb{U}$ *satisfies assumption (ii) of Theorem E.2; or*

(b) $\gamma(\mathbb{U};\bar{\mathcal{A}}) = \frac{\min\{\Phi(S(\mathbb{V});\bar{\mathcal{A}}),\Phi(S(\mathbb{U}-\mathbb{U});\bar{\mathcal{A}})\}}{\Phi(S(\mathbb{U});\bar{\mathcal{A}})}$ *if* $\mathbb{U}$ *satisfies assumptions (i) and (ii) of Theorem E.2.*

Then the random variable (E.5) *satisfies* $\mathbb{E}(\delta_{\mathbb{U}}) \leq \delta$.

Proof. Lemmas E.4 and E.5 imply that

$$\mathbb{E}_{\bar{\mathcal{A}}}(\delta_{\mathbb{U}}) \lesssim m^{-1}\mathbb{E}_{\bar{\mathcal{A}}}(R_{p,\bar{\mathcal{A}}}) \int_0^{\chi_q/2} \sqrt{\log(2\mathcal{N}(\widetilde{S}(\mathbb{U}), \|\cdot\|_{\bar{\mathcal{A}}}, t))}\,\mathrm{d}t.$$

Let $0 < \tau < \chi_q/2$ and then use Lemmas E.6 and E.9 to obtain

$$\begin{aligned}\int_0^{\chi_q/2} \sqrt{\log(2\mathcal{N}(\widetilde{S}(\mathbb{U}), \|\cdot\|_{\bar{\mathcal{A}}}, t))}\,\mathrm{d}t \lesssim& \sqrt{\log(2d)}\tau + \sqrt{n}\int_0^{\tau} \sqrt{\log(1+2\chi_q'/t)}\,\mathrm{d}t \\ &+ \sqrt{\frac{qm^{\frac{1}{q}}\Phi(\mathbb{W};\bar{\mathcal{A}})\log(M)}{\alpha}} \int_{\tau}^{\chi_q/2} 1/t\,\mathrm{d}t.\end{aligned}$$

A short exercise gives that

$$\begin{aligned}\int_0^{\chi_q/2} \sqrt{\log(2\mathcal{N}(\widetilde{S}(\mathbb{U}), \|\cdot\|_{\bar{\mathcal{A}}}, t))}\,\mathrm{d}t \lesssim& \sqrt{\log(2d)}\tau + \sqrt{n}\tau\sqrt{\log(\mathrm{e}(1+2\chi_q'/\tau))} \\ &+ \sqrt{\frac{qm^{\frac{1}{q}}\Phi(\mathbb{W};\bar{\mathcal{A}})\log(M)}{\alpha}}\log(\chi_q/(2\tau)).\end{aligned}$$

Now set $\tau = \chi_q/(\sqrt{\log(2d)} + \sqrt{n})$ to obtain

$$\begin{aligned}\int_0^{\chi_q/2} \sqrt{\log(2\mathcal{N}(\widetilde{S}(\mathbb{U}), \|\cdot\|_{\bar{\mathcal{A}}}, t))}\,\mathrm{d}t \lesssim& \chi_q\sqrt{\log(\mathrm{e}(1+2\chi_q'/\chi_q(\sqrt{\log(2d)}+\sqrt{n})))} \\ &+ \sqrt{\frac{qm^{\frac{1}{q}}\Phi(\mathbb{W};\bar{\mathcal{A}})\log(M)}{\alpha}}\log((\sqrt{\log(2d)}+\sqrt{n})). \\ \lesssim& \chi_q\sqrt{L_1} + \sqrt{\frac{qm^{\frac{1}{q}}\Phi(\mathbb{W};\bar{\mathcal{A}})}{\alpha}}\sqrt{L_2}.\end{aligned}$$

Now (E.11) and the fact that

$$\max\{\Phi(S(\mathbb{U});\bar{\mathcal{A}}), \Phi(\mathbb{W};\bar{\mathcal{A}})\} = \Phi(S(\mathbb{U})\cup\mathbb{W};\bar{\mathcal{A}})$$

yield

$$\int_0^{\chi_q/2} \sqrt{\log(2\mathcal{N}(\widetilde{S}(\mathbb{U}), \|\cdot\|_{\bar{\mathcal{A}}}, t))}\,\mathrm{d}t \lesssim \sqrt{\frac{qm^{\frac{1}{q}}\Phi(S(\mathbb{U})\cup\mathbb{W};\bar{\mathcal{A}})}{\alpha}}\sqrt{L_1+L_2}$$

We now use (E.20) to obtain

$$\mathbb{E}_{\bar{\mathcal{A}}}(\delta_{\mathbb{U}}) \lesssim m^{-1/2}\left(\frac{\Phi(S(\mathbb{U})\cup\mathbb{W};\bar{\mathcal{A}})}{\alpha}\right)^{1-\frac{1}{2p}} \sqrt{q}\sqrt{L_1+L_2}\sqrt{\mathbb{E}_{\bar{\mathcal{A}}}(\delta_{\mathbb{U}})+1}.$$

We now choose $p = 1 + 1/\log(\lambda)$, where $\lambda = 1 + \Phi(S(\mathbb{U})\cup\mathbb{W};\bar{\mathcal{A}})/\alpha$ and proceed as before. This gives

$$\mathbb{E}_{\bar{\mathcal{A}}}(\delta_{\mathbb{U}}) \lesssim m^{-1/2}\left(\frac{\Phi(S(\mathbb{U})\cup\mathbb{W};\bar{\mathcal{A}})}{\alpha}\right)^{\frac{1}{2}} \sqrt{\log(\mathrm{e}\Phi(S(\mathbb{U})\cup\mathbb{W};\bar{\mathcal{A}})/\alpha)}\sqrt{L_1+L_2}\sqrt{\mathbb{E}_{\bar{\mathcal{A}}}(\delta_{\mathbb{U}})+1}.$$

This completes the proof. □

E.3.4. PROBABILITY BOUND

We now bound $\delta_{\mathbb{U}}$ in probability.

Theorem E.12. *Let $0 < \delta, \epsilon < 1$ and suppose that $\mathbb{U}$ satisfies assumption (ii) of Theorem E.2. Suppose also that $\mathbb{E}(\delta_{\mathbb{U}}) \leq \delta/2$. Then $\delta_{\mathbb{U}} \leq \delta$ with probability at least $1-\epsilon$, provided*

$$m \gtrsim \delta^{-2} \cdot \alpha^{-1} \cdot \Phi(S(\mathbb{U}); \bar{\mathcal{A}}) \cdot \log(2/\epsilon).$$

Proof. Our analysis is based on a version of Talagrand's theorem (see, e.g., (Adcock & Hansen, 2021, Thm. 13.33)). First, let B^* be a countable dense subset of $\widetilde{S}(\mathbb{U})$. This exists since $\mathbb{X}$ is separable. Then from (E.8) we have

$$\begin{aligned} Z = \delta_{\mathbb{U}} &= m^{-1} \sup_{u \in B^*} \left| \sum_{i=1}^m \left\langle U_i(u) - \mathbb{E}\left(U_i(u)\right), u \right\rangle_{\mathbb{X}} \right| \\ &= \sup_{u \in B^*} \max \left\{ +\sum_{i=1}^m \left\langle V_i(u), u \right\rangle_{\mathbb{X}}, -\sum_{i=1}^m \left\langle V_i(u), u \right\rangle_{\mathbb{X}} \right\}, \end{aligned}$$

where V_i is the $\mathcal{B}(\mathbb{V})$-valued random variable $V_i = m^{-1}(U_i - \mathbb{E}(U_i))$. Now consider the measurable space

$$\Omega = \{V \in \mathcal{B}(\mathbb{V}) : V = V^*, \ \sup_{u \in B^*} |\langle V(u), u \rangle_{\mathbb{X}}| \leq K\}, \qquad K := 2\Phi(S(\mathbb{U}); \bar{\mathcal{A}})/(\alpha m)$$

and let $f_u^{\pm} : \Omega \to \mathbb{R}$ be defined by $f_u^{\pm}(V) = \pm \langle V(u), u \rangle_{\mathbb{X}}$. Observe that we can write

$$\delta_{\mathbb{U}} = \sup_{u \in B^*} \max \left\{ \sum_{i=1}^m f_u^+(V_i), \sum_{i=1}^m f_u^-(V_i) \right\}.$$

Notice also that $\mathbb{E}(f^{\pm}(V_i)) = 0$. Next, notice that, for all $u \in B^*$,

$$|\langle U_i(u), u \rangle_{\mathbb{X}}| = \|A_i(u)\|_{\mathbb{Y}_i}^2 \leq \Phi(S(\mathbb{U}); \bar{\mathcal{A}})/\alpha.$$

Therefore $V_i \in \Omega$, $i = 1, \ldots, m$.

Now observe that

$$\begin{aligned} \mathbb{E}\left(\sum_{i=1}^m (f_u^{\pm}(V_i))^2 \right) &\leq m^{-2} \sum_{i=1}^m \mathbb{E} |\langle U_i(u), u \rangle_{\mathbb{X}}|^2 \\ &\leq \frac{2\Phi(S(\mathbb{U}); \bar{\mathcal{A}})}{\alpha m^2} \sum_{i=1}^m \mathbb{E} \|A_i(u)\|_{\mathbb{Y}_i}^2 \\ &= \frac{2\Phi(S(\mathbb{U}); \bar{\mathcal{A}})}{\alpha m} |||u|||_{\mathbb{X}}^2 \\ &= \frac{2\Phi(S(\mathbb{U}); \bar{\mathcal{A}})}{\alpha m} =: \sigma^2. \end{aligned}$$

Finally, observe that

$$\overline{Z} = \sup_{u \in B^*} \max \left\{ \left| \sum_{i=1}^m f_u^+(V_i) \right|, \left| \sum_{i=1}^m f_u^-(V_i) \right| \right\} = Z,$$

in this case. In particular, $\mathbb{E}(\overline{Z}) = \mathbb{E}(Z) = \mathbb{E}(\delta_{\mathbb{U}})$.

Now suppose that $\mathbb{E}(\delta_{\mathbb{U}}) \leq \delta/2$. Then, using Talagrand's theorem,

$$\begin{aligned} \mathbb{P}(\delta_{\mathbb{U}} \geq \delta) &\leq \mathbb{P}\left(|\delta_{\mathbb{U}} - \mathbb{E}(\delta_{\mathbb{U}})| \geq \delta/2 \right) \\ &\leq 3 \exp\left(-\frac{\delta \alpha m}{4c\Phi(S(\mathbb{U}); \bar{\mathcal{A}})} \log\left(1 + \frac{\delta}{2+\delta} \right) \right) \\ &\leq 3 \exp\left(-\frac{\log(4/3)\delta^2 \alpha m}{4c\Phi(S(\mathbb{U}); \bar{\mathcal{A}})} \right). \end{aligned}$$

Due to the condition on m, we deduce that $\mathbb{P}(\delta_{\mathbb{U}} \geq \delta) \leq \epsilon$, as required. □

E.3.5. PROOFS OF THEOREMS E.2 AND E.3

For the proof of Theorem E.2, we divide into two parts. The first deals with conditions (a) and (b), and the second deals with condition (c), which involves a different approach.

Proof of Theorem E.2; conditions (a) and (b). Consider condition (a). Recall that $\mathbb{U} \subseteq \mathbb{U} - \mathbb{U}$ whenever $\mathbb{U}$ is a cone. Therefore, since $\mathbb{U} \subseteq \mathbb{V}$ as well, we see that

$$\Phi(S(\mathbb{U}-\mathbb{U});\bar{\mathcal{A}}) \geq \Phi(S(\mathbb{U});\bar{\mathcal{A}}), \quad \Phi(S(\mathbb{V});\bar{\mathcal{A}}) \geq \Phi(S(\mathbb{U});\bar{\mathcal{A}}). \tag{E.21}$$

Hence $\gamma(\mathbb{U};\bar{\mathcal{A}}) \geq 1$ which implies that

$$\log(2(1+\gamma(\mathbb{U};\bar{\mathcal{A}}))) \lesssim \log(2\gamma(\mathbb{U};\bar{\mathcal{A}})),$$

The result subject to condition (a) now follows immediately from Theorems E.10 and E.12.

Moreover, using (E.21) and the inequality $\log(x) \leq x$, we see that

$$\Phi(S(\mathbb{U});\bar{\mathcal{A}})\log(\gamma(\mathbb{U};\bar{\mathcal{A}})) \leq \Phi(S(\mathbb{U});\bar{\mathcal{A}})\log\left[\frac{\Phi(S(\mathbb{U}-\mathbb{U});\bar{\mathcal{A}})}{\Phi(S(\mathbb{U});\bar{\mathcal{A}})}\right] \leq \Phi(S(\mathbb{U}-\mathbb{U});\bar{\mathcal{A}}). \tag{E.22}$$

Hence condition (b) implies condition (a), which implies the result. □

Remark E.13 (On assumption (ii) of Theorem E.2). Notice that Theorem E.12 does not require assumption (i) of Theorem E.2. Moreover, in Theorem E.10 we gave a bound for the expectation when only assumption (ii) holds. Therefore, it is straightforward to derive an extension of Theorem E.2 in which only assumption (ii) holds. The only difference is the definition of $\gamma(\mathbb{U};\bar{\mathcal{A}})$ in (E.4), which would now be given by the expression in Theorem E.10, part (a). Note that the same considerations also apply to Theorem 3.3 since, as we show in its proof below, it follows from a direct application of Theorem E.2 to the difference set $\mathbb{U}' = \mathbb{U} - \mathbb{U}$.

Next, for condition (c), we follow a different and simpler approach based on the matrix Chernoff bound.

Proof of Theorem E.2; conditions (c). Since $\mathbb{U} \subseteq \mathbb{V} = \mathbb{V}_1 \cup \cdots \cup \mathbb{V}_d$ is a union of d subspaces of dimension n it suffices, by the union bound, to show that (E.3) holds for each subspace $\mathbb{V}_i$ with probability at least $1 - \epsilon/d$.

Therefore, without loss of generality, we may now assume that $\mathbb{V}$ is a single subspace of dimension n and show (E.3) for this subspace. To do this, we follow a standard approach based on the matrix Chernoff bound. Let $\{v_i\}_{i=1}^n$ be an orthonormal basis for $\mathbb{V}$ and write $v = \sum_{i=1}^n c_i v_i \in \mathbb{V}$. Then

$$\frac{1}{m}\sum_{i=1}^m \|A_i(v)\|_{\mathbb{Y}_i}^2 = \sum_{i=1}^m c^* X_i c,$$

where X_i is the self-adjoint random matrix

$$X_i = \frac{1}{m}\left(\langle A_i(v_j), A_i(v_k)\rangle_{\mathbb{Y}_i}\right)_{j,k=1}^n \in \mathbb{C}^{n\times n}.$$

Therefore, (E.3) is equivalent to

$$(1-\delta)\alpha \leq \lambda_{\min}\left(\sum_{i=1}^m X_i\right) \leq \lambda_{\max}\left(\sum_{i=1}^m X_i\right) \leq (1+\delta)\beta.$$

Noticethat $c^* X_i c = m^{-1}\|A_i(v)\|_{\mathbb{Y}_i}^2$ and therefore X_i is nonnegative definite. Also, we have

$$\sum_{i=1}^m c^* \mathbb{E}_{A_i\sim\mathcal{A}_i} X_i c = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{A_i\sim\mathcal{A}_i}\|A_i(v)\|_{\mathbb{Y}_i}^2.$$

Hence, by (1.1), we have

$$\alpha \le \lambda_{\min}\left(\sum_{i=1}^m \mathbb{E}X_i\right) \le \lambda_{\max}\left(\sum_{i=1}^m \mathbb{E}X_i\right) \le \beta.$$

Therefore, it suffices to show that

$$(1-\delta)\lambda_{\min}\left(\sum_{i=1}^m \mathbb{E}X_i\right) \le \lambda_{\min}\left(\sum_{i=1}^m X_i\right) \le \lambda_{\max}\left(\sum_{i=1}^m X_i\right) \le (1+\delta)\lambda_{\min}\left(\sum_{i=1}^m \mathbb{E}X_i\right)$$

with high probability. Observe that

$$c^* X_i c = \frac{1}{m}\|A_i(v)\|_{\mathbb{Y}_i}^2 \le \frac{1}{m}\Phi(S(\mathbb{V});\bar{\mathcal{A}})\|v\|_{\mathbb{X}}^2 = \frac{1}{m}\Phi(S(\mathbb{V});\bar{\mathcal{A}})\|c\|_2^2,$$

almost surely, and therefore

$$\lambda_{\max}(X_i) \le \frac{1}{m}\Phi(S(\mathbb{V});\bar{\mathcal{A}}).$$

almost surely, for all $i = 1,\dots,m$. Thus, by the matrix Chernoff bound and the union bound, we have

$$\mathbb{P}\left(\text{(E.3) holds with } \alpha' = \alpha(1-\delta) \text{ and } \beta' = (1+\delta)\beta\right) \ge 1 - 2n\exp\left(-\frac{\alpha m((1+\delta)\log(1+\delta)-\delta)}{\Phi(S(\mathbb{V});\bar{\mathcal{A}})}\right).$$

Notice that $(1+\delta)\log(1+\delta) - \delta \ge \delta^2/3$. Hence the right-hand side is bounded below by

$$1 - 2n\exp\left(-\frac{\alpha m\delta^2/3}{\Phi(S(\mathbb{V});\bar{\mathcal{A}})}\right).$$

We deduce that if

$$m \gtrsim \delta^{-2}\cdot\alpha^{-1}\cdot\Phi(S(\mathbb{V});\bar{\mathcal{A}})\cdot\log(2n/\epsilon),$$

then, with probability at least $1-\epsilon$, (E.3) holds with $\alpha' = \alpha(1-\delta)$ and $\beta' = (1+\delta)\beta$ for the subspace $\mathbb{V}$. Replacing ϵ by ϵ/d now gives the result. □

Proof of Theorem E.3. The result follows from Theorems E.11 and E.12. Recall that $\gamma(\mathbb{U};\bar{\mathcal{A}}) \ge 1$ since assumption (i) holds. Therefore, we may bound the logarithmic term in Theorem E.11 as

$$\begin{aligned}
&L_0\cdot(L_1+L_2)\\
&\quad\lesssim \log(2\Phi(S(\mathbb{U})\cup\mathbb{W};\bar{\mathcal{A}})/\alpha)\cdot\left[\log(2\gamma(\mathbb{U};\bar{\mathcal{A}})(\log(2d)+n)) + \log^2(M)\log(\log(2d)+n)\right]\\
&\quad\lesssim \log(2\Phi(S(\mathbb{U})\cup\mathbb{W};\bar{\mathcal{A}})/\alpha)\cdot\left[\log(2\gamma(\mathbb{U};\bar{\mathcal{A}})) + \log(2M)\log^2(\log(2d)+n)\right].
\end{aligned}$$

Hence, we immediately deduce the result under the condition (a). Now recall (E.21). Then, much as in (E.22), we deduce that

$$\Phi(S(\mathbb{U})\cup\mathbb{W};\bar{\mathcal{A}})\log(\gamma(\mathbb{U};\bar{\mathcal{A}})) \le \Phi(S(\mathbb{U}-\mathbb{U})\cup\mathbb{W};\bar{\mathcal{A}})$$

We conclude that condition (b) implies condition (a). This gives the result under condition (b). Finally, recalling (E.21) once more, we also have

$$\Phi(S(\mathbb{U})\cup\mathbb{W};\bar{\mathcal{A}})\log(\gamma(\mathbb{U};\bar{\mathcal{A}})) \le \Phi(S(\mathbb{V})\cup\mathbb{W};\bar{\mathcal{A}}).$$

Hence condition (c) also implies condition (a), and therefore we get the result under this condition as well. □

E.4. Proof of Theorems 3.3 and 3.4

Proof of Theorems 3.3 and 3.4. We follow essentially the same ideas as in (Adcock et al., 2023, Thm. 4.8). Let $\delta = 1/2$ (this value is arbitrary) and E be the event that (E.3) holds over $\mathbb{U}'$ with $\alpha' = (1-\delta)\alpha$ and $\beta' = (1+\delta)\beta$. Theorems E.2–E.3 (with $\mathbb{U}'$ in place of $\mathbb{U}$) and the various conditions on m imply that $\mathbb{P}(E^c) \le \epsilon$. Now write

$$\mathbb{E}\|x-\check{x}\|_{\mathbb{X}}^2 = \mathbb{E}(\|x-\check{x}\|_{\mathbb{X}}^2|E)\mathbb{P}(E) + \mathbb{E}(\|x-\check{x}\|_{\mathbb{X}}^2|E^c)\mathbb{P}(E^c). \tag{E.23}$$

Observe that the mapping $\mathcal{C} : \mathbb{X} \to \mathbb{X}, x \mapsto \min\{1, \theta/\|x\|_{\mathbb{X}}\}x$ is a contraction. We also have $\check{x} = \mathcal{C}(\hat{x})$ and $x = \mathcal{C}(x)$, where in the latter case, we used the fact that $\theta \geq \|x\|_{\mathbb{X}}$ by assumption.

Fix $u \in \mathbb{U}$ and consider the first term. If the event E occurs, then the properties of $\mathcal{C}$ and Lemma E.1 give that

$$\|x - \check{x}\|_{\mathbb{X}} = \|\mathcal{C}(x) - \mathcal{C}(\hat{x})\|_{\mathbb{X}} \leq \|x - \hat{x}\|_{\mathbb{X}} \leq \left\{ \frac{2\sqrt{2}}{\sqrt{\alpha}} \|\bar{A}(x-u)\|_{\overline{\mathbb{Y}}} + \|x-u\|_{\mathbb{X}} \right\} + \frac{2\sqrt{2}}{\sqrt{\alpha}} \|\bar{e}\|_{\overline{\mathbb{Y}}} + \frac{\sqrt{2}\zeta}{\sqrt{\alpha}}.$$

By the Cauchy–Schwarz inequality, we deduce that

$$\mathbb{E}(\|x - \check{x}\|_{\mathbb{X}}^2 | E) \lesssim \frac{1}{\alpha} \mathbb{E}\|\bar{A}(x-u)\|_{\overline{\mathbb{Y}}}^2 + \|x-u\|_{\mathbb{X}}^2 + \frac{1}{\alpha}\|\bar{e}\|_{\overline{\mathbb{Y}}}^2 + \frac{\zeta^2}{\alpha}.$$

We now use (E.1) to deduce that

$$\mathbb{E}(\|x - \check{x}\|_{\mathbb{X}}^2 | E) \lesssim \frac{\beta}{\alpha}\|x-u\|_{\mathbb{X}}^2 + \frac{1}{\alpha}\|\bar{e}\|_{\overline{\mathbb{Y}}}^2 + \frac{\zeta^2}{\alpha}. \tag{E.24}$$

We next consider the second term of (E.23). Using the properties of $\mathcal{C}$, we see that

$$\|x - \check{x}\|_{\mathbb{X}} \leq 2\theta.$$

Substituting this and (E.24) into (E.23) and recalling that $\mathbb{P}(E^c) \leq \epsilon$ now gives the result. □

Remark E.14 (On the error bounds in expectation). The above proof explains why the truncation term is needed: namely, it bounds the error in the event where empirical nondegeneracy (E.3) fails. Modifying the estimator appears unavoidable if one is to obtain an error bound in expectation. A downside of the estimator $\check{x}$ is that it requires an a priori bound on $\|x\|_{\mathbb{X}}$. In some limited scenarios, one can avoid this by using a different estimator (Dolbeault & Cohen, 2022). Indeed, let E be the event in the above proof. Then define the conditional estimator

$$\check{x} = \begin{cases} \hat{x} & \text{if } E \text{ occurs} \\ 0 & \text{otherwise} \end{cases}.$$

One readily deduces that this estimator obeys the same error bound (1.6) with θ replaced by $\|x\|_{\mathbb{X}}$. Unfortunately, computing this estimator involves computing the empirical nondegeneracy constants. This is possible in some limited scenarios, such as when $\mathbb{U}$ is a linear subspace, as the constants then correspond to the maximum and minimum singular values of a certain matrix. However, this is generally impossible in the nonlinear case. For instance, in the classical compressed problem, we previously noted that (E.3) is equivalent to the RIP. It is well known that computing RIP constants in NP-hard (Tillmann & Pfetsch, 2014).

Remark E.15. The observant reader may have noticed a small technical issue with Theorem 3.3 and its proof: the estimator $\hat{x}$ (and therefore $\check{x}$) is nonunique, and therefore $\|x - \check{x}\|_{\mathbb{X}}$ is not a well-defined random variable. To resolve this issue, one can replace $\|x - \check{x}\|_{\mathbb{X}} = \|x - \mathcal{C}(\hat{x})\|_{\mathbb{X}}$ by the maximum error between x and *any* ζ-minimizer $\hat{x}$. The error bound remains the same for this (well-defined) random variable.