Comprehensive Exam Course Notes

Craig Gross

Fall 2019

Contents

CHAPTER 1

Uncertainty Quantification

1.1. General Problem

We study problems from uncertainty quantification (UQ) under the model of stochastic partial differential equations (PDE). The general problem is described as in [30]. For a spatial domain $\Omega \subset \mathbb{R}^\ell$ (where usually $\ell = 1, 2, 3$), and a time domain $[0, T]$ with $T > 0$, we consider a system of PDE given by

$$\begin{cases} \mathcal{L}(x, t, \omega)[u(x, t, \omega)] = f(x, t, \omega), & \Omega \times (0, T] \times \Xi, \\ \mathcal{B}(u) = 0, & \partial\Omega \times [0, T] \times \Xi, \\ u = u_0, & \Omega \times \{t = 0\} \times \Xi, \end{cases} \tag{1.1}$$

where $\mathcal{L}$ is a differential operator, $\mathcal{B}$ is the boundary condition operator, u_0 is the initial condition, and $\omega \in \Xi$ denotes the random inputs of the system in a probability space $(\Xi, \mathcal{F}, \mathbb{P})$. The solution is denoted

$$u(x, t, \omega) : \overline{\Omega} \times [0, T] \times \Xi \to \mathbb{R}^{n_u} \tag{1.2}$$

where $n_u \geqslant 1$ is the dimension of u.

Example 1.1. *Consider a stochastic (stationary) diffusion equation in one spatial dimension,*

$$\begin{cases} -D \cdot (a(x, \omega) D u) = f(x, \omega), & x \in (-1, 1) \\ u(-1, \omega) = u_\ell(\omega), & u(1, \omega) = u_r(\omega) \end{cases} . \tag{1.3}$$

The diffusivity coefficient $a(x, \omega)$ *and source* $f(x, \omega)$ *are random fields indexed by the spatial variable, and the boundary data* u_ℓ, u_r *are random variables.*

1.2. Parameterization

1.2.1. Finite-Dimensional Noise Assumption. In practice, we make a finite dimensional noise assumption [18] that all dependence on a probability space is through a finite set of (usually) independent random variables. This converts the purely random problem (1.1) into the *parameterized* problem

$$\begin{cases} \mathcal{L}(x, t, Z)[u(x, t, Z)] = f(x, t, Z), & \Omega \times (0, T] \times \Gamma \\ \mathcal{B}(u) = 0, & \partial\Omega \times [0, T] \times \Gamma, \\ u = u_0 & \Omega \times \{t = 0\} \times \Gamma \end{cases} \tag{1.4}$$

where $Z = (Z_1, \ldots, Z_N) \in \Gamma \subset \mathbb{R}^N$ are (usually) independent random variables parameterizing the randomness. We can also view Z as simply a deterministic parameter sequence, making (1.4) simply a parameterized PDE. We now view the solution (1.2) as depending on this parameter sequence

$$u(x, t, Z) : \overline{\Omega} \times [0, T] \times \Gamma \to \mathbb{R}^{n_u}. \tag{1.5}$$

Note now that we can view our setup entirely deterministically. Any realization of Z simply corresponds to some deterministic value in Γ. If we can analyze our solution depending on the parameterization $u(Z)$ over the parameter space Γ, any law of Z allows us to perform an analogous analysis over Ξ. As an example, if we view u as a random field and Z as a random vector with law P_Z and density π_Z, then the mean solution $\mathbb{E}u$ can be calculated as

$$\mathbb{E}u = \int_\Xi u(\omega)\ dP = \int_\Gamma u(Z)\ dP_Z = \int_\Gamma u(z)\pi_Z(z)\ dz,$$

where for the latter two integrals, only information of the deterministic quantity $u(Z)$ is necessary.

1.2.2. Karhunen-Loève. Often, we are able to parameterize the random fields in a random PDE by a Karhunen-Loève expansion, which, at its core, is a spectral expansion of the covariance function.

THEOREM 1.1 (Karhunen-Loève). *Let $a(t, \omega)$, $t \in T$ be a random field such that $a(t, \cdot) \in L^2(\Xi)$ for all $t \in T$, with continuous covariance function*

$$\mathrm{cov}_a(s,t) = \mathbb{E}[(a(s,\cdot) - \mathbb{E}a(s,\cdot))\,(a(t,\cdot) - \mathbb{E}a(t,\cdot))].$$

Then for λ_i, and orthonormal ψ_i satisfying the eigenvalue problem

$$\int_T \mathrm{cov}_a(s,t)\psi_i(s)\ ds = \lambda_i \psi_i(t), \qquad \textit{for all } t \in T,$$

and mean-zero, uncorrelated random variables Z_i defined by

$$Z_i = \frac{1}{\sqrt{\lambda_i}} \int_T (a(t,\cdot) - \mathbb{E}a(t,\cdot))\,\psi_i(t)\ dt,$$

we have the following Karhunen-Loève (KL) expansion of the random field

$$a(t,\cdot) = \mathbb{E}a(t,\cdot) + \sum_{i=1}^\infty \sqrt{\lambda_i}\psi_i(t) Z_i. \tag{1.6}$$

In practice we can truncate the Karhunen-Loève expansion of the random field, parameterizing the randomness in terms of the uncorrelated Z_i. If the Z_i are Gaussian, they are therefore independent. If they are non-Gaussian, their independence is in practice assumed [30]. Thus, we can use a truncated KL expansion to satisfy the finite dimensional noise assumption.

The level N at which we truncate the KL expansion is dictated by the eigenvalues of the covariance function of a. By Mercer's theorem, and the analogous Parseval's identity,

$$\int_T \mathrm{cov}_a(t,t)\ dt = \sum_{i=1}^\infty \lambda_i.$$

Thus, we can choose N so that some tolerance percentage $1 - \epsilon$ of the total variance is captured by the eigenvalues, e.g.,

$$\frac{\sum_{i=1}^N \lambda_i}{\int_T \mathrm{cov}_a(t,t)\ dt} \geqslant 1 - \epsilon.$$

In general, a highly correlated process with smoother covariance function will have faster decaying eigenvalues and therefore N can be taken smaller for a desired tolerance [18].

EXAMPLE 1.2. *We now rewrite Example 1.1 as a parameterized diffusion equation. Assuming a, f, u_ℓ, u_r are independent, we represent a and f by truncated KL expansions as in*

(1.6)*:*

$$a(x,\cdot) \approx \tilde{a}(x, Z^a) = \mathbb{E}a(x,\cdot) + \sum_{i=1}^{N_a} \sqrt{\lambda_i^a}\psi_i^a Z_i^a,$$

$$f(x,\cdot) \approx \tilde{a}(x, Z^f) = \mathbb{E}a(x,\cdot) + \sum_{i=1}^{N_f} \sqrt{\lambda_i^f}\psi_i^f Z_i^f.$$

Taking

$$Z = (Z_1, \dots, Z_N) = (Z_1^a, \dots, Z_{N_a}^a, Z_1^f, \dots, Z_{N_f}^f, u_\ell, u_r)$$

with $N = N_a + N_f + 2$*, we have*

$$\begin{cases} -D\cdot(\tilde{a}(x,Z)Du) = \tilde{f}(x,Z), & x \in (-1,1) \\ u(-1,Z) = Z_{N-1}, \qquad u(1,Z) = Z_N \end{cases}. \tag{1.7}$$

1.3. Weak Formulations and Well-Posedness

1.3.1. Non-Parametric Problems. We provide an abstract definition for weak solutions to general PDE with specific examples and refer to [**14**] for details in specific classes of PDE. Suppose we have a deterministic (or non-parametric) problem in the form of (1.1) (or (1.4)),

$$\begin{cases} \mathcal{L}(x,t)[u(x,t)] = f(x,t), & \Omega \times (0,T], \\ \mathcal{B}(u) = 0, & \partial\Omega \times [0,T], \\ u = u_0, & \Omega \times \{t = 0\}, \end{cases} . \tag{1.8}$$

The weak formulation of the PDE (1.8) is a generalization of the equation which no longer holds pointwise in the domain of u, but rather pointwise in a function space. This allows for us to abstract the problem, and consider solutions in larger classes of spaces which can then, in certain circumstances, be shown to solve the pointwise problem as well.

We will now abstract the problem in the form of [**23**]. First, let $\mathcal{X}, \mathcal{Y}$ be separable, reflexive Banach spaces over coefficient field $\mathbb{R}$ with duals $\mathcal{X}^*, \mathcal{Y}^*$ respectively. We think of $\mathcal{X}$ as the space of solutions and $\mathcal{Y}$ as the space of test functions. Now, suppose that $\mathcal{L} : \mathcal{X} \to \mathcal{Y}^*$ is a bounded linear operator. We associate the bilinear form $\mathfrak{L} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ to $\mathcal{L}$ by

$$\mathfrak{L}(v, w) = {}_{\mathcal{Y}^*}\langle \mathcal{L}v, w\rangle_{\mathcal{Y}} \tag{1.9}$$

where the latter notation is applying the element of the dual $\mathcal{L}v$ to w. Finally, we also assume $f \in \mathcal{Y}^*$.

Definition 1.1. *We say that* $u \in \mathcal{X}$ *is* the weak solution *of* (1.8) *with the previously mentioned assumptions if*

$$\mathfrak{L}(u, w) = {}_{\mathcal{Y}^*}\langle f, w\rangle_{\mathcal{Y}} \qquad \textit{for all } w \in \mathcal{Y}. \tag{1.10}$$

Example 1.3. *We now provide a concrete example which motivates Definition 1.1. Suppose we consider the simple second order elliptic equation in divergence form,*

$$\begin{cases} -D\cdot(A(x)Du(x)) = f(x), & x \in \Omega \\ u(x) = 0, & x \in \partial\Omega \end{cases}. \tag{1.11}$$

for some domain $\Omega \subset \mathbb{R}^\ell$*,* $f \in L^2(\Omega)$*, and symmetric matrix* $A \in \mathbb{R}^{\ell,\ell}$ *with entries in* $L^\infty(\Omega)$ *satisfying the uniform ellipticity assumption*

$$v^T A(x) v \geqslant \alpha \qquad \textit{a.e. } x \in \Omega \textit{ and for all } v \in \mathbb{S}^{\ell-1}$$

for some $\alpha > 0$.

The weak formulation of (1.11) *attempts to make as little assumption on the regularity of* u *as possible. In order to reduce the number of derivatives on* u *we multiply by a test function* w *in some sufficient class of functions and integrate by parts:*

$$\begin{aligned} -\int_\Omega D \cdot (ADu)w &= \int_\Omega (Dw)^T ADu - \int_{\partial\Omega} \nu^T ADuw \, dS \\ &= \int_\Omega (Dw)^T ADu, \end{aligned}$$

if we assume that $w = 0$ *on* $\partial\Omega$. *So at this point, we only need that* u *is once differentiable. But for the differentiability under the integral, all that's truly necessary is weak differentiability. Taking into account the boundary conditions, it thus makes sense for* $u, w \in H^1_0(\Omega)$, *the Sobolev space of trace-zero, once weakly differentiable functions in* $L^2(\Omega)$ *with weak derivatives also in* $L^2(\Omega)$. *This gives us* $\mathcal{X} = \mathcal{Y} = H^1_0(\Omega)$.

Now, the bilinear form induced by $\mathcal{L} = -D \cdot (AD\cdot))$ *can be taken as*

$$\mathfrak{L}(u, w) = \int_\Omega (Dw)^T ADu.$$

Multiplying the right hand side and integrating, we identify $f \in L^2$ *with the element in* $H^{-1}(\Omega)$ *(the dual of* $H^1_0(\Omega)$*) given by*

$${}_{H^{-1}}\langle f, w\rangle_{H^1_0} = \int_\Omega fw + \int_\Omega 0 \cdot Dw.$$

Thus, a weak solution of (1.11) *is* $u \in H^1_0(\Omega)$ *such that for every* $w \in H^1_0(\Omega)$,

$$\int_\Omega (Dw)^T ADu = \int_\Omega fw.$$

Note that in this case, because $\mathfrak{L}$ *is symmetric, it induces an inner product on* $H^1_0(\Omega)$, *so by the Riesz-representation theorem, the weak formulation has a unique solution for all* $f \in H^{-1}(\Omega)$ *that satisfies the stability estimate*

$$\|u\|_{H^1_0(\Omega)} \leqslant \frac{1}{\alpha}\|f\|_{H^{-1}(\Omega)}$$

where we take $\|u\|_{H^1_0(\Omega)} = \|Du\|_{L^2(\Omega)}$ *in the energy sense and* α *is the constant from the uniform ellipticity assumption.*

For more general $\mathfrak{L}$ we may not be able to exploit symmetry and must therefore use the Lax-Milgram theorem (or for $\mathcal{X} \neq \mathcal{Y}$, more general existence, uniqueness, and stability theorems).

THEOREM 1.2 (Lax-Milgram). *Suppose the bilinear form* $\mathfrak{L} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is continuous:*

(1.12) *there exists some* $\gamma > 0$ *such that* $|\mathfrak{L}(u, w)| \leqslant \gamma\|u\|_{\mathcal{X}}\|w\|_{\mathcal{X}}$ *for all* $u, w \in \mathcal{X}$,

and strongly coercive:

(1.13) *there exists some* $\alpha > 0$ *such that* $\mathfrak{L}(u, u) \geqslant \alpha\|u\|^2_{\mathcal{X}}$ *for all* $u \in \mathcal{X}$.

Then for any $f \in \mathcal{X}^*$, *problem* (1.10) *has a unique solution* u *satisfying the stability estimate*

$$\|u\|_{\mathcal{X}} \leqslant \frac{1}{\alpha}\|f\|_{\mathcal{X}^*}.$$

1.3.2. Parameterized Problems. We consider two types of weak formulations of the parameterized PDE of the form (1.4). In both cases, we modify the bilinear form in (1.9) to depend on the parameter sequence $Z \in \Gamma$:

$$\mathfrak{L}(u, w; Z) = {}_{\mathcal{Y}^*}\langle \mathcal{L}(Z)[u], w\rangle_{\mathcal{Y}}.$$

Additionally, we also view the right hand side $f(Z) : \Gamma \to \mathcal{Y}^*$ as depending on the parameter sequence.

Definition 1.2. *A function* $u(Z) : \Gamma \to \mathcal{X}$ *solves the parameterized PDE* (1.4) pointwise in the weak sense *if*

$$\mathfrak{L}(u(Z), w; Z) = {}_{\mathcal{Y}^*}\langle f(Z), w\rangle_{\mathcal{Y}} \qquad \text{for all } w \in \mathcal{Y} \text{ and almost all } Z \in \Gamma. \tag{1.14}$$

Example 1.4. *Now suppose that in Example 1.3, we allow* $A = A(x, Z)$ *to depend on the parameter sequence* $Z \in \Gamma$*. The pointwise weak formulation of*

$$\begin{cases} -D \cdot (A(x, Z)Du(x, Z)) = f(x, Z), & x \in \Omega, \quad Z \in \Gamma \\ u(x, Z) = 0, & x \in \partial\Omega, \quad Z \in \Gamma \end{cases} \tag{1.15}$$

is to find $u : \Gamma \to H_0^1(\Omega)$ *such that for all* $w \in H_0^1(\Omega)$ *and almost surely for* $Z \in \Gamma$*,*

$$\int_\Omega (Dw(x))^T A(x, Z) D(u(Z))(x)\, dx = \int_\Omega f(x, Z) w(x)\, dx. \tag{1.16}$$

Note that if for all $Z \in \Gamma$*,* $A(x, Z)$ *satisfies (uniform) uniform ellipticity over* Z

$$v^T A(x, Z) v \geqslant \alpha \qquad a.e.\ x \in \Omega \text{ and for all } v \in S^{\ell-1},$$

by the Lax-Milgram theorem, we have the pointwise stability estimate

$$\|u(\cdot, Z)\|_{H_0^1(\Omega)} \leqslant \frac{1}{\alpha}\|f(\cdot, Z)\|_{H^{-1}(\Omega)} \tag{1.17}$$

for all $Z \in \Gamma$*.*

In order to define the second type of weak solution, we reduce our generality a bit and specify the previously mentioned Banach spaces to more specific function spaces.

Definition 1.3. *For a Banach space* $\mathcal{X}(\Omega)$ *of functions over the domain* $\Omega \subset \mathbb{R}^d$ *and* Γ *a measure space with measure defined by the density* $\pi(z)dz$*, we define the tensor product space*

$$\mathcal{X}(\Omega) \otimes L_\pi^q(\Gamma) = \left\{ u : \Omega \times \Gamma \to \mathbb{R} \mid \|u\|_{\mathcal{X}(\Omega) \otimes L_\pi^q(\Gamma)} < \infty \right\},$$

where

$$\|u\|_{\mathcal{X}(\Omega)\otimes L_\pi^q(\Gamma)} = \left\| \|u\|_{\mathcal{X}(\Omega)} \right\|_{L_\pi^q(\Gamma)}$$

Remark 1.1. *In these notes, we will always take* $\mathcal{X}(\Omega) = H^k(\Omega)$ *and* $q = 2$*. In this case, by Fubini's theorem, we may interchange the order of the norms in the definition of* $\|\cdot\|_{H^k(\Omega)\otimes L_\pi^2(\Gamma)}$*. Then for* $u \in H^k(\Omega) \otimes L_\pi^2(\Gamma)$*, this implies that both* $u(\cdot, Z) \in H^k(\Omega)$ *almost surely in* Γ *and* $u(x, \cdot) \in L_\pi^2(\Gamma)$ *almost everywhere in* Ω*. We can think of this as saying that* u *is jointly* H^k *and* L_π^2 *in its two domains.*

Additionally, for this type of weak solution we allow our differential operator $\mathcal{L}$ to induce a not-necessarily-bilinear form as in [**16**]. Specifically, suppose that after multiplying the PDE

$$\mathcal{L}(Z)[u] = f$$

by a test function $w \in \mathcal{X}(\Omega) \otimes L^q_\pi(\Gamma)$ and integrating by parts, we obtain on the left hand side the operator $\mathfrak{L} : (\mathcal{X}(\Omega) \otimes L^q_\pi(\Gamma))^2 \to \mathbb{R}$ defined by

$$\mathfrak{L}(u, w) = \sum_{k=1}^{K} \int_\Gamma \int_\Omega S_k(u; z) T_k(w) \pi(z)\ dx\ dz \tag{1.18}$$

where $S_k(\cdot; \cdot)$, $k \in [K]$ are in general nonlinear operators and $T_k(\cdot, \cdot)$, $k \in [K]$ are linear operators.

Definition 1.4. *A function* $u(x, Z) \in \mathcal{X}(\Omega) \otimes L^q_\pi(\Gamma)$ *solves the parameterized PDE* (1.4) jointly in the weak sense *if for* $\mathfrak{L}$ *defined by* (1.18),

$$\mathfrak{L}(u, w) = \int_\Gamma \int_\Omega w(x, z) f(x, z) \pi(z)\ dx\ dz \qquad \textit{for all } w \in \mathcal{X}(\Omega) \otimes L^q_\pi(\Gamma). \tag{1.19}$$

Example 1.5. *In the joint weak sense, the second order parameterized elliptic PDE* (1.15) *from Example 1.4 is to find* $u \in H^1_0(\Omega) \otimes L^2_\pi(Z)$ *such that*

$$\begin{aligned} &\int_\Gamma \int_\Omega (D_x w(x, z))^T A(x, z) D_x u(x, z) \pi(z)\ dx\ dz \\ &\qquad = \int_\Gamma \int_\Omega f(x, z) w(x, z) \pi(z)\ dx\ dz \quad \textit{for all } w \in H^1_0(\Omega) \otimes L^2_\pi(\Gamma) \end{aligned}, \tag{1.20}$$

where D_x *is the weak gradient with respect to* $x \in \Omega$.

Proposition 1.1. *A function* u *solves problem* (1.16) *with* $f \in H^1_0(\Omega) \otimes L^2_\pi(\Gamma)$ *if and only if it solves* (1.20).

Proof. Assume $u \in H^1_0(\Omega) \otimes L^2_\pi(\Gamma)$ satisfies (1.20). Now let $w \in H^1_0(\Omega)$, $\phi \in C^\infty_C(\Gamma)$, and denote

$$g(z) = \int_\Omega \left[(Dw(x))^T A(x, z) Du(x, z) - f(x, z) w(x) \right]\ dx.$$

Then $w(x)\phi(Z) \in H^1_0(\Omega) \otimes L^2_\pi(\Gamma)$, and by virtue of (1.20),

$$\int_\Gamma g(z) \phi(z) \pi(z)\ dz = 0.$$

Since $g(z) \in L^2_\pi(\Gamma)$ (as $u \in H^1_0(\Omega) \otimes L^2_\pi(\Gamma)$), the linear functional on $L^2_\pi(\Gamma)$ defined by

$$\phi \mapsto \int_\Gamma g(z) \phi(z) \pi(z)\ dz$$

is zero on $C^\infty_c(\Gamma)$ which is dense in $L^2_\pi(\Gamma)$. Thus, the aforementioned functional is zero, and so is $g(z)$ almost surely, satisfying (1.16).

On the other hand, if u solves (1.16) with $f \in H^1_0(\Omega) \otimes L^2_\pi(\Gamma)$, the stability estimate (1.17) guarantees that $u \in H^1_0(\Omega) \otimes L^2_\pi(\Gamma)$. Integrating both sides of (1.16) over Γ then gives (1.20). □

1.4. Monte-Carlo Methods

Assuming that we can solve a non-parameterized problem, our main goal is to solve the parameterized problem (1.4). In practice however, we usually desire some quantities of interest (QoI) of our solution (1.5). This usually takes the form of statistics of $u(Z)$, such as the expectation, variance, or higher moments. In cases where the QoI is statistical (that is, we desire some information of u in expectation), the most straightforward method of solution is the Monte-Carlo method.

The procedure is as follows. Draw M i.i.d. realizations of the parameter sequence Z as $\left\{Z^{(i)}\right\}_{i=1}^{M}$ and let $\left\{u\left(Z^{(i)}\right)\right\}_{i=1}^{M}$ be the ensemble of solutions to the deterministic problem for each parameter realization. If $\overline{u}$ is the expectation of $u(Z)$, then the law of large numbers ensures that

$$u_M := \frac{1}{M}\sum_{i=1}^{M} u\left(Z^{(i)}\right) \to \overline{u}.$$

We can actually quantify the rate of this convergence. If we denote $\|w\| = \|w\|_{H_0^1(\Omega)} = \|Dw\|_{L^2(\Omega)}$, we calculate

$$\begin{aligned}\mathbb{E}\left(\|\overline{u}-u_M\|^2\right) &= \frac{1}{M^2}\mathbb{E}\left(\left\|\sum_{i=1}^{M}(u\left(Z^{(i)}\right)-\overline{u})\right\|^2\right)\\ &= \frac{1}{M}\mathbb{E}\left(\|u(Z)-\overline{u}\|^2\right) \text{ since the } Z^{(i)} \text{ are i.i.d.}\\ &= \frac{1}{M}\mathbb{V}\left[u(Z)\right].\end{aligned}$$

Thus, $u_M \to \overline{u}$ in $H_0^1(\Gamma)\otimes(L_\pi^2(\Gamma)$ at a rate of $O(M^{-1/2})$ which is unfortunately rather slow. There are techniques for boosting this rate of convergence such as multi-level Monte-Carlo. However, these methods are outside the scope of these notes, as we will focus on stochastic Galerkin and stochastic collocation methods.

1.5. Polynomial Chaos

In order to account for the parameter domain in numerical approximations to solutions of (1.4), we will use spectral methods. The main idea of spectral methods is to solve for the generalized Fourier coefficients of the solution making use of the weak formulation of the problem. In our setting, the generalized Fourier basis will be orthogonal polynomials over the parameter domain.

Definition 1.5. *Let* $\{\phi_\nu\}_{\nu\in\mathbb{N}_0^N}$ *be a collection of orthonormal polynomials with respect to an orthogonalization measure* $\pi:\Gamma\to\mathbb{R}$*, which is to say*

$$\int_\Gamma \phi_\nu(z)\phi_\eta(z)\pi(z)\,dz = \delta_{\nu,\eta} \tag{1.21}$$

for all multiindices $\nu,\eta\in\mathbb{N}_0^N$*. The* generalized polynomial chaos (gPC) expansion *of a function* $u(Z)$ *is given by*

$$u(Z) = \sum_{\nu\in\mathbb{N}_0^N}\hat{u}_\nu\phi_\nu(Z),$$

where the gPC coefficients are calculated as

$$\hat{u}_\nu = \int_\Gamma u(z)\phi_\nu(z)\pi(z)\,dz.$$

In the multivariate case (that is, when $N > 1$), we usually construct the gPC basis as a tensor product of a copies of a univariate one. So if $\{\phi_k\}_{k=0}^{\infty}$ is a univariate gPC basis, we will assume that the multivariate polynomials are constructed as

$$\phi_\nu(Z) = \prod_{n=1}^{N}\phi_{\nu_n}(Z_n). \tag{1.22}$$

Note that it is not necessary to assume the use of the same univariate basis in each component. If we have N univariate bases indexed by n, $\left\{\phi_k^{(n)}\right\}_{k=1}^{\infty}$, we can take

$$\phi_\nu(Z) = \prod_{n=1}^{N} \phi_{\nu_n}^{(n)}(Z_n). \tag{1.23}$$

When we are dealing with a stochastic problem and the parameter sequence Z represents a sequence of random variables, it is natural then to let $\pi\, dz$ to be the distribution of Z, giving that the orthonormal polynomial basis ϕ_ν is then orthonormal with respect to the distribution of Z. In this case, the coefficients of u simplify to

$$\hat{u}_\nu = \mathbb{E}\left[u\phi_\nu\right].$$

Additionally when the components of Z are i.i.d. (or at least independent), the tensor product structure of the multivariate gPC basis (1.22) (or (1.23)) is required.

The benefit of using a gPC expansion corresponding to the distribution of Z is that statistics of the solution u are easily obtained from the gPC expansion coefficients. For example, assuming that $\phi_0 = 1$, the orthonormality relationship (1.21) gives that $\mathbb{E}\phi_\nu = 0$ for all $\nu \neq 0$. Thus,

$$\mathbb{E}u = \mathbb{E} \sum_{\nu \in \mathbb{N}_0^N} \hat{u}_\nu \phi_\nu = \hat{u}_0.$$

For the variance, a similar calculation to Parseval's identity gives

$$\begin{aligned}\mathbb{V}u &= \mathbb{E}\left(u - \mathbb{E}u\right)^2 \\ &= \mathbb{E}\left(\sum_{\nu \in \mathbb{N}_0^N \setminus \{0\}} \hat{u}_\nu \phi_\nu\right)^2 \\ &= \sum_{\nu \in \mathbb{N}^N} \hat{u}_\nu^2.\end{aligned}$$

Table 1.1 lists a few distributions with their corresponding polynomial chaos bases.

TABLE 1.1. Univariate Distributions and Generalized Polynomial Chaos Bases.

Distribution	Density/weight	gPC Basis	Support
Uniform	$\frac{1}{b-a}$	Legendre	$[a, b]$
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$	Hermite	$(-\infty, \infty)$
Gamma	$e^{-x}x^{\alpha}$	Laguerre	$[0, \infty)$
Wigner semicircle	$\frac{1}{\pi\sqrt{1-x^2}}$	Chebyshev	$[-1, 1]$

1.5.1. Spectral Convergence. When a gPC basis is obtained from eigenfunctions of a singular Sturm-Liouville problem (as are those listed in Table 1.1), we can derive convergence estimates of the gPC expansion to the original function based on eigenvalues of the Sturm-Liouville problem and more importantly, the smoothness of the original function. We make these notions precise as follows.

THEOREM 1.3. *Suppose that a univariate gPC basis* $\{\phi_k\}_{k\in\mathbb{N}_0}$ *satisfies the singular Sturm-Liouville eigenvalue problem over* (a,b) *(where* a *or* b *can equal* ∞*)*

$$\mathcal{Q}(x)\phi_k(x) = -\lambda_k\phi_k(x), \tag{1.24}$$

where

$$\mathcal{Q}(x) = \frac{1}{\pi(x)}\left[D\left(p(x)D\right) + q(x)\right],$$

and $p(x), \pi(x) > 0$, $q(x) \geqslant 0$ *over* (a,b), *and* $p(a) = p(b) = 0$. *Then if* $P_M : L^2_\pi \to \mathcal{P}_M$ *is the projector onto* $\mathcal{P}_M = \operatorname{span}\{\phi_k\}_{k=0}^M$ *taken by truncating the gPC expansion,*

$$\|u - P_M u\|_{L^2_\pi} \leqslant \frac{C}{\lambda_M^m}\|\mathcal{Q}^m u\|_{L^2_\pi},$$

when the right hand side exists.

PROOF. Letting $(\cdot,\cdot) = (\cdot,\cdot)_{L^2_\pi}$, twice applying integration by parts gives

$$\begin{aligned}(\mathcal{Q}v, w) &= \int_a^b wD\left(pDv\right) + \int_a^b qvw \\ &= \int_a^b vD\left(pDw\right) + \int_a^b qvw \\ &= \left(v, \mathcal{Q}w\right).\end{aligned}$$

Thus, $\mathcal{Q}$ is self-adjoint.

In particular

$$\begin{aligned}(u, \phi_k) &= \frac{(-1)^m}{\lambda_k^m}\left(u, \mathcal{Q}^m\phi_k\right) \\ &= \frac{(-1)^m}{\lambda_k^m}\left(\mathcal{Q}^m u, \phi_k\right),\end{aligned}$$

so long as $\mathcal{Q}^m u \in L^2_\pi$. Applying to the remainder,

$$\begin{aligned}\|u - P_M u\|^2_{L^2_\pi} &= \sum_{k=M+1}^{\infty}\left(u, \phi_k\right)^2 \\ &= \sum_{k=M+1}^{\infty}\frac{1}{\lambda_k^{2m}}(\mathcal{Q}^m u, \phi_k)^2 \\ &\leqslant \frac{1}{\lambda_M^{2m}}\sum_{k=0}^{\infty}(\mathcal{Q}^m u, \phi_k)^2 \\ &= \frac{1}{\lambda_M^{2m}}\|\mathcal{Q}^m u\|^2_{L^2_\pi}.\end{aligned}$$

Taking square roots gives the desired inequality. □

EXAMPLE 1.6. *Consider the example of Legendre polynomials which solve the Sturm-Liouville problem* (1.24) *on the interval* $[-1,1]$ *with* $p(x) = (1-x^2)$, $q(x) = 0$, $\pi(x) = \frac{1}{2}$, *and* $\lambda_k = O(k^2)$*. If* $u \in H^k_\pi$*, note first that*

$$\begin{aligned}\|\mathcal{Q}u\|_{L^2_\pi} &= \left\|(1-x^2)D^2u(x) - 2xDu(x)\right\|_{L^2_\pi} \\ &\leqslant \left\|(1-x^2)D^2u(x)\right\|_{L^2_\pi} + \|2xDu(x)\|_{L^2} \\ &\leqslant \left\|(1-x^2)\right\|_{L^\infty}\|D^2u\|_{L^2_\pi} + \|2x\|_{L^\infty}\|Du\|_{L^2_\pi} \\ &\leqslant C\|u\|_{H^2_\pi}\end{aligned} \tag{1.25}$$

Additionally, we can obtain estimates on higher derivatives of $\mathcal{Q}u$ *as*

$$\left\|D^\ell \mathcal{Q}u\right\|_{L^2_\pi} \leqslant C\|u\|_{H^{\ell+2}_\pi}.$$

Indeed, for $D\mathcal{Q}u$*, we have*

$$\begin{aligned}\|D\mathcal{Q}u\|_{L^2_\pi} &= \left\|D[(1-x^2)D^2u - 2xDu]\right\|_{L^2_\pi} \\ &= \left\|\mathcal{Q}Du - 2xD^2u - 2Du\right\|_{L^2_\pi} \\ &\leqslant \|\mathcal{Q}Du\|_{L^2_\pi} + C\|u\|_{H^2_\pi} \\ &\leqslant C\|u\|_{H^3_\pi}.\end{aligned} \tag{1.26}$$

Inductively, we see that

$$\begin{aligned}\left\|D^\ell Qu\right\|_{L^2_\pi} &= \left\|D^{\ell-1}\left[\mathcal{Q}Du - 2xD^2u - 2Du\right]\right\|_{L^2_\pi} \\ &\leqslant C\|Du\|_{H^{\ell+1}_\pi} + 2\left\|D^{\ell-1}xD^2u\right\|_{L^2_\pi} \\ &\leqslant C\|u\|_{H^{\ell+2}_\pi},\end{aligned}$$

where the last inequality is obtained through successive applications of the product rule giving $D^{\ell-1}xD^2u = (\ell-1)D^\ell u + xD^{\ell+1}u$*.*

Thus,

$$\|\mathcal{Q}u\|_{H^\ell_\pi} \leqslant C\|u\|_{H^{\ell+2}_\pi}.$$

Applying this relationship m *times, we obtain*

$$\|\mathcal{Q}^m u\|_{L^2_\pi} \leqslant C\|u\|_{H^{2m}_\pi}.$$

Taking $m = k/2$ *and using that* $\lambda_M = O(M^2)$*, Theorem 1.3 then gives us the spectral convergence result*

$$\|u - P_M u\|_{L^2_\pi} \leqslant C\frac{1}{M^k}\|u\|_{H^k_\pi}. \tag{1.27}$$

Thus, u *converges to its gPC expansion at rate* dependent on the smoothness of u*. The smoother the function, the faster the convergence.*

1.5.2. Finite-Dimensional Multivariate gPC Bases. In the case of univariate gPC expansions, it is straightforward to truncate the expansion and produce an approximation to the original function, as considered in the previous section. However, in the case of multivariate polynomials, the selection of a finite-dimensional basis becomes nontrivial.

We consider a general finite multi-index set $\Lambda(p) \subset \mathbb{N}_0^N$ indexed by a parameter p. This parameter is thought of as the polynomial order of the associated approximation. This indexing scheme will satisfy $\Lambda(0) = \{(0,0,\dots,0)\}$ the nesting property $\Lambda(p) \subseteq \Lambda(p+1)$. If $\{\phi_n\}$ We then define the corresponding basis set

$$\mathcal{P}_{\Lambda(p)} = \operatorname{span}\left\{\prod_{n=1}^N \phi_{\nu_n} \mid \nu \in \Lambda(p)\right\},$$

and let $M := \#\Lambda(p)$ be the dimension of the subspace.

We consider here three rules for constructing the index sets. The first is the simplest, the tensor product (TP) rule defined by choosing

$$\Lambda_{\mathrm{TP}}(p) = \left\{ \nu \in \mathbb{N}_0^N \mid \max_{1 \leqslant n \leqslant N} \nu_n \leqslant p \right\}, \tag{1.28}$$

and we obtain

$$M = \#\Lambda_{\mathrm{TP}} = (p+1)^N. \tag{1.29}$$

Next, we consider the total degree (TD) rule

$$\Lambda_{\mathrm{TD}}(p) = \left\{ \nu \in \mathbb{N}_0^N \mid \sum_{n=1}^{N} \nu_n \leqslant p \right\}, \tag{1.30}$$

with

$$M = \binom{N+p}{N} = \frac{(N+p)!}{N!p!}. \tag{1.31}$$

Next is the hyperbolic cross (HC) rule

$$\Lambda_{\mathrm{HC}}(p) = \left\{ \nu \in \mathbb{N}_0^N \mid \prod_{n=1}^{N} (\nu_n + 1) \leqslant p \right\}, \tag{1.32}$$

with size [**18**] bounded above by

$$M \leqslant p(1 + \log(p))^{N-1}. \tag{1.33}$$

1.5.3. Convergence via Stechkin Estimates. With varying choices of index sets in the multivariate case, quantifying the convergence of the truncated gPC expansion becomes more subtle. For a fixed M, we hope to find an index set Λ_M^{opt} which minimizes the error $u - u_{\Lambda_M^{\mathrm{opt}}}$ in some sense. This $u_{\Lambda_M^{\mathrm{opt}}}$ is then the corresponding best M-term approximation to u.

We take this point of view to provide an alternate take on the rate of convergence of gPC expansions which takes a more general perspective than the argument in Theorem 1.3. We no longer make use of structure in the gPC coefficients, but use only their summability in weighted ℓ_p norms.

As an aside, when using the standard Fourier basis, the class of functions summable in weighted ℓ_p norms with certain weights *coincide* with Sobolev spaces. So in a certain sense, the following notion of convergence can be considered as a generalization of spectral convergence.

We follow the procedure in [**24**, Section 3], and as a result introduce some notation standard in compressive sensing.

Definition 1.6. *For a sequence of weights $\omega = (\omega_\nu)_{\nu \in \Lambda}$ indexed over the same index set as the gPC basis, we define the* weighted ℓ_p space

$$\ell_{\omega,p} := \left\{ x = (x_\nu)_{\nu \in \Lambda} \mid \|x\|_{\omega,p} := \left(\sum_{\nu \in \Lambda} \omega_\nu^{2-p} |x_\nu|^p \right)^{1/p} < \infty \right\}, \quad 0 < p \leqslant 2,$$

with the weighted ℓ_0 norm as

$$\|x\|_{\omega,0} = \sum_{\nu \in \mathrm{supp}(x)} \omega_\nu^2.$$

Additionally, we define associated function quasi-normed function space

$$S_{\omega,p} := \left\{ u(x) = \sum_{\nu\in\Lambda} \hat{u}_\nu \phi_\nu(x) \mid \|u\|_{\omega,p} := \|\hat{u}\|_{\omega,p} < \infty \right\}, \quad 0 < p < 1.$$

Definition 1.7. *For a given basis of weighted cardinality* M*, we define the* error in the best weighted M-term approximation to a vector $x \in \ell_{\omega,p}$ *as*

$$\sigma_M(x)_{\omega,p} = \inf_{z:\|z\|_{\omega,0}\leqslant M} \|x - z\|_{\omega,p},$$

and associated error in the best weighted M-term approximation to a function $u \in S_{\omega,p}$ *as*

$$\sigma_M(u)_{\omega,p} = \sigma_M(\hat{u})_{\omega,p}.$$

We then take $u_{\Lambda^{\mathrm{opt}}_{M,p}}$ *as the minimizer (if it exists), and*

$$\Lambda^{\mathrm{opt}}_{M,p} = \operatorname{supp} u_{\Lambda^{\mathrm{opt}}_{M,p}}.$$

Theorem 1.4. *Suppose that the weight sequence* ω *satisfies* $\omega_\nu \geqslant \|\phi_\nu\|_\infty$ *on* $\mathbb{N}_0^N$ *and* $M > \|\omega\|_\infty^2$*. If* $u \in S_{\omega,p}$*,*

$$\left\| u - u_{\Lambda^{\mathrm{opt}}_{M,1}} \right\|_\infty \leqslant \left(M - \|\omega\|_\infty^2 \right)^{1-1/p} \|u\|_{\omega,p}, \quad p < 1. \tag{1.34}$$

Notice that the only possible dependence on N is through the ∞-norm of the multivariate gPC basis. This appears to only restrict the regime of sizes of index sets where this estimate is viable and the necessary summability of u. This dependence should be clarified further though as we begin linking the ideas of index sets, weights, sparsity, and polynomial degree.

The proof of this result relies on a weighted Stechkin estimate which will involve estimates on the following intermediate quantity.

Definition 1.8. *For a sequence* x*, let* $\gamma = (\gamma_j)_{j\in\mathbb{N}}$ *denote the non-increasing rearrangement of the sequence* $(|x_\nu|^p \omega_\nu^{-p})_{\nu\in\mathbb{N}_0^N}$ *and* $\alpha : \mathbb{N} \to \mathbb{N}_0^N$ *the corresponding permutation. We then define* S *to correspond to the first* k *indices of* γ *satisfying* $\sum_{j=1}^k \omega_{\alpha(j)}^2 \leqslant M$*, and let the* error in the quasi-best M-term approximation *be*

$$\tilde{\sigma}_M(x)_{\omega,p} = \|x - x_S\|_{\omega,p} = \|x_{S^c}\|_{\omega,p}.$$

Thus, x_S can be thought of in some sense as taking only the terms of x corresponding to the M largest "weighted terms." Note that by definition $\sigma_M(x)_{\omega,p} \leqslant \tilde{\sigma}_M(x)_{\omega,p}$. This allows us to prove a Stechkin estimate similarly to [**15**, Proposition 2.3].

Theorem 1.5 ([**24**], Theorem 3.2). *For* $p < q \leqslant 2$*, let* $x \in \ell_{\omega,p}$*. Then for* $M > \|\omega\|_\infty^2$*,*

$$\sigma_M(x)_{\omega,q} \leqslant \tilde{\sigma}_M(x)_{\omega,q} \leqslant \left(M - \|\omega\|_\infty^2 \right)^{1/q-1/p} \|x\|_{\omega,p}.$$

Proof. We work only with the weighted quasi-best estimate for x, x_S. We first wish to bound $\tilde{\sigma}_M(x)^q_{\omega,q} = \|x_{S^c}\|^q_{\omega,q}$ by an expression involving $\|x\|_{\omega,p}$. To this end, we consider

$$\begin{aligned}
\|x_{S^c}\|^q_{\omega,q} &= \sum_{\nu\notin S} |x_\nu|^q \omega_\nu^{2-q} \\
&\leqslant \sup_{\nu\notin S} |x_\nu|^{q-p} \omega_\nu^{p-q} \sum_{\nu\notin S} |x_\nu|^p \omega_\nu^{2-p} \\
&\leqslant \left(\frac{1}{\omega(S)} \sum_{\eta\in S} \omega_\eta^2 \sup_{\nu\notin S} |x_\nu|^p \omega_\nu^{-p} \right)^{(q-p)/p} \qquad \|x\|^p_{\omega,p} \\
&\leqslant \left(\frac{1}{\omega(S)} \sum_{\eta\in S} \omega_\eta^2 |x_\eta|^p \omega_\eta^{-p} \right)^{(q-p)/p} \qquad \|x\|^p_{\omega,p} \\
&\leqslant \omega(S)^{(p-q)/p} \|x\|^q_{\omega,p}.
\end{aligned}$$

All that remains is to estimate $\omega(S)$ from below, since $p - q < 0$.

Since $\omega(S) \leqslant M$ and S is the longest subset of reordered indices that maintains this weighed sparsity, $\omega(S) + \omega_\alpha^2 > M$ where $\alpha \notin S$. Since $\omega_\alpha^2 \leqslant \|\omega\|^2_\infty$, we have

$$M - \|\omega\|^2_\infty \leqslant \omega(S).$$

Combining with the above and taking the $1/q$ power gives the desired bound. □

We now apply this bound to obtain L^∞ estimates on the convergence of u to its multivariate gPC using *only* weighted ℓ_p summability of the Fourier coefficients.

Proof of Theorem 1.4. We bound the ∞-norm using the weights as

$$\begin{aligned}
\left\| u - u_{\Lambda^{\text{opt}}_{M,1}} \right\|_\infty &\leqslant \sum_{\nu\notin\Lambda^{\text{opt}}_{M,1}} |\hat{u}_\nu| \|\phi_\nu\|_\infty \\
&\leqslant \sum_{\nu\notin\Lambda^{\text{opt}}_{M,1}} |\hat{u}_\nu| \omega_\nu \qquad (1.35) \\
&= \sigma_M(u)_{\omega,1}.
\end{aligned}$$

Theorem 1.5 with $q = 1$ then gives

$$\sigma_M(u)_{\omega,1} \leqslant \left(M - \|\omega\|^2_\infty \right)^{1-1/p} \|u\|_{\omega,p},$$

as desired. □

In practice, we cannot use this estimate to reduce our truncation error to a desired tolerance, as the coefficients $(\hat{u}_\nu)_{\nu\in\mathbb{N}_0^N}$ are not known. Additionally, though (1.35) suggests that $\Lambda^{\text{opt}}_{M,1}$ should consist of the M largest values of $|\hat{u}_\nu|\omega_\nu$, we would again need to know the gPC coefficients in advance of any calculations.

As an alternative, we may consider *quasi-optimal approximations*, that is, calculating $\Lambda_M^{q-\text{opt}}$ with only knowledge of sharp upper bounds on the coefficients $(\hat{u}_\nu)_{\nu\in\mathbb{N}_0^N}$. With the help of computation friendly upper bounds, it is possible to optimize the Stechkin estimate (1.34) for a given M to calculate q and hence determine the quasi-optimal index set. In cases where the upper bounds are not so friendly, an alternative approach forgoing the Stechkin estimate was developed, directly working on bounding the sum of the coefficients outside of a quasi-optimal index set.

We will revisit the idea of optimal index sets later, and discuss approaches which proceed to calculate best M-term approximations with no prior knowledge of the optimal index set.

1.5.4. gPC Expansions of Banach Space Valued Functions. For the remainder of these notes, the gPC expansion will be considered for Banach space valued functions $u : \Gamma \to \mathcal{X}$ where $\mathcal{X}$ is a Banach space of functions. In this case, the only difference is that the coefficients

$$\hat{u}_\nu = \int_\Gamma u(z)\phi_\nu(z)\pi(z)\ dz \tag{1.36}$$

now take values in $\mathcal{X}$. Unless otherwise specified, we will fix $\mathcal{X} = H^k(\Omega)$ and identify the Bochner space $L^p_\pi(\Gamma;\mathcal{X})$ with $H^k(\Omega) \otimes L^p_\pi(\Gamma)$. In this case, the coefficients are obtained by integrating out the parameter, resulting in

$$\hat{u}_\nu(x) = \int_\Gamma u(x,z)\phi_\nu(z)\pi(z)\ dz \in H^k(\Omega).$$

All previous notions of convergence still hold in some sense with the spatial domain considered. For example, the spectral convergence result (1.27) becomes a pointwise estimate in Ω. Additionally, the notions of Definition 1.6 still apply, where the absolute value of any coefficient is replaced by $\|\cdot\|_{H^k(\Omega)}$. Thus, the analogous result of Theorem 1.4 must take this into account, giving a pointwise bound on $\|\cdot\|_{H^k(\Omega)}$. This then reads

$$\sup_{z\in\Gamma} \left\| u(z) - u_{\Lambda^{\text{opt}}_{M,1}}(z) \right\|_{H^k(\Omega)} \leqslant \left(M - \|\omega\|^2_\infty \right)^{1-1/p} \left(\sum_{\nu\in\Lambda} \|\hat{u}\|^p_{H^k(\Omega)} \omega_\nu^{2-p} \right)^{1/p}.$$

1.6. Stochastic Galerkin

With our in-depth understanding of gPC expansions out of the way, let's now use them to produce approximate solutions to PDE again returning to the idea of weak solutions. We will first introduce stochastic Galerkin methods. Unfortunately, we will not be able to maintain the same amount of generality, as the formulation is often highly problem dependent. The general structure though follows the same setup as our derivation of weak solutions.

The main idea is to replace the infinite dimensional stochastic space $L^q_\pi(\Gamma)$ with a finite dimensional approximation. We will use a finite dimensional subspace spanned by a gPC basis,

$$\mathcal{P}_{\Lambda(p)} = \text{span}\left\{ \prod_{n=1}^N \phi_{\nu_n} \mid \nu \in \Lambda(p) \right\},$$

where $\Lambda(p)$ is an index set parameterized by some degree parameter p.

Thus, in the joint weak problem given in Definition 1.4, instead of $\mathcal{X}(\Omega) \otimes L^q_\pi(\Gamma)$ used as the solution and test space, we use $\mathcal{X}(\Omega) \otimes \mathcal{P}_{\Lambda(p)}$. Thus, when we take test functions $w \in H^1_0(\Omega) \otimes \mathcal{P}_{\Lambda(p)}$, we can represent these functions in the $H^1_0(\Omega)$-valued gPC expansion

$$w = \sum_{\nu\in\Lambda(p)} \hat{w}_\nu \phi_\nu,$$

with coefficients $\hat{w}_\nu \in H_0^1(\Omega)$. Since the operators T_k in (1.18) are linear and don't depend on z, we can write $\mathfrak{L}(u, w)$ (without loss of generality considering only one term) as

$$\begin{aligned}
\mathfrak{L}(u, w) &= \int_\Gamma \int_\Omega S(u(x, z); z) T(w(x, z)) \pi(z) \, dx \, dz \\
&= \int_\Gamma \int_\Omega S(u(x, z); z) T \left(\sum_{\nu \in \Lambda(p)} \hat{w}_\nu(x) \phi_\nu(z) \right) \pi(z) \, dx \, dz \\
&= \sum_{\nu \in \Lambda(p)} \int_\Gamma \int_\Omega S(u(x, z); z) T(\hat{w}_\nu(x)) \phi_\nu(z) \pi(z) \, dx \, dz \\
&= \sum_{\nu \in \Lambda(p)} \int_\Omega \tilde{S}(u, \phi_\nu; x) T(\hat{w}_\nu(x)) \, dx \\
&= \sum_{\nu \in \Lambda(p)} \left(\tilde{S}(u, \phi_\nu; x), T(\hat{w}_\nu(x)) \right)_{L^2(\Omega)} \\
&=: \sum_{\nu \in \Lambda(p)} \tilde{\mathfrak{L}}(u, w_\nu)_\nu .
\end{aligned}$$

where

$$\tilde{S}(u, \phi_\nu; x) = \int_\Gamma S(u(x, z); z) \phi_\nu(z) \pi(z) \, dz.$$

As for the right hand side,

$$\begin{aligned}
\int_\Gamma \int_\Omega f(x, z) w(x, z) \pi(z) \, dx \, dz &= \sum_{\nu \in \Lambda(p)} \int_\Gamma \int_\Omega f(x, z) \hat{w}_\nu(x) \phi_\nu(z) \pi(z) \, dx \, dz \\
&= \sum_{\nu \in \Lambda(p)} \int_\Omega F(\phi_\nu; x) \hat{w}_\nu(x) \, dx \\
&= \sum_{\nu \in \Lambda(p)} (F(\phi_\nu; x), \hat{w}_\nu(x))_{L^2(\Omega)} \\
&=: \sum_{\nu \in \Lambda(p)} \left({}_{H^{-1}(\Omega)} \langle F, w_\nu \rangle_{H_0^1(\Omega)} \right)_\nu ,
\end{aligned}$$

where

$$F(\phi_\nu; x) = \int_\Gamma f(x, z) \phi_\nu(z) \pi(z) \, dz.$$

Since w is arbitrary in $H_0^1(\Omega) \otimes L_\pi^q(\Gamma)$, we now assume that $\hat{w}_\nu$ above is arbitrary in $H_0^1(\Omega)$. Additionally, we have equality of these two quantities after summing so long as they hold for all $\nu \in \Lambda(p)$. This leads us to the stochastic Galerkin formulation.

Definition 1.9. *Let* $u_M = \sum_{\eta \in \Lambda(p)} \hat{u}_\eta \phi_\eta$. *The stochastic Galerkin formulation of the joint weak problem* (1.19) *is to solve the following weak system of PDE: Find* $\hat{u} = \{\hat{u}_\eta\}_{\eta \in \Lambda(p)} \in (H_0^1(\Omega))^M$ *such that for any* $w \in H_0^1(\Omega)$,

$$\left(\tilde{\mathfrak{L}}(u_M, w) \right)_\nu = \left({}_{H^{-1}(\Omega)} \langle F, w \rangle_{H_0^1(\Omega)} \right)_\nu \quad \textit{for all } \nu \in \Lambda(p).$$

Thus, the stochastic Galerkin formulation removes any stochastic element to the problem, converting it into a system of *deterministic* PDE. If one has the ability to solve the weak system of PDE in some manner, then the solution of the original parameterized problem can also be obtained. This process is often done using finite element methods which take advantage of the weak formulation of the problem, and use the Galerkin method further in the space of $H_0^1(\Omega)$ test functions.

EXAMPLE 1.7. *We provide a concrete example where we start with the original pointwise PDE and derive the Galerkin formulation the same way we derived the weak formulation. For some variety, we'll work with the* time-dependent *problem*

$$\begin{cases} D_t u - D_x \cdot (A(x, Z) D_x u) = f(x, t, Z), & \Omega \times (0, T] \times \Gamma \\ u = 0, & \partial\Omega \times (0, T] \times \Gamma \\ u = u_0, & \Omega \times \{t = 0\} \times \Gamma \end{cases} . \tag{1.37}$$

We'll assume a KL expansion in A,

$$A(x, Z) = \hat{A}_0(x) + \sum_{i=1}^{N} \hat{A}_i(x) Z_i = \sum_{i=0}^{N} \hat{A}_i(x) Z_i$$

where $Z_0 \equiv 1$.

Now we assume a truncated gPC expansion for the solution

$$u_M = \sum_{\eta \in \Lambda(p)} \hat{u}_\eta(x, t) \phi_\eta(Z) =: \sum_{j=1}^{M} \hat{u}_j(x, t) \phi_j(Z),$$

where we relabel $\Lambda(p)$ *through some indexing scheme. In the same fashion as before, we integrate against test functions, but since we've discretized our stochastic space, we only need to use the basis functions* ϕ_k:

$$\sum_{j=1}^{M} D_t \hat{u}_j \int_\Gamma \phi_j \phi_k \pi \, dz - \sum_{0=1}^{N} \sum_{j=1}^{M} D_x(\hat{A}_i(X) D_x \hat{u}_j) \int_\Gamma z_i \phi_j \phi_k \pi \, dz = \int_\Gamma f \phi_k \pi \, dz.$$

Letting

$$\hat{f}_k = \int_\Gamma f \phi_k \pi \, dz$$

be the gPC coefficients of f,

$$e_{ijk} = \int_\Gamma z_i \phi_j \phi_k \pi \, dz, \qquad B_{jk} = \sum_{i=0}^{N} \hat{A}_i e_{ijk}, \quad 1 \leqslant j, k \leqslant M,$$

and using orthonormality in the gPC basis, we can rewrite this equation as

$$D_t \hat{u}_k - \sum_{j=1}^{M} D_x(B_{jk} D_x \hat{u}_j) = \hat{f}_k.$$

In matrix form, we arrive at the deterministic system of PDE

$$D_t \hat{u} - D_x(B D_x \hat{u}) = \hat{f}$$

where $\hat{u} = (\hat{u}_k)_{k=1}^M$, $B = \left(B_{jk}\right)_{1 \leqslant j, k \leqslant M}$, *and* $\hat{f} = \left(\hat{f}_k\right)_{k=1}^M$.

Notice that from this point forward, we can consider the weak formulation of the system and arrive at a setup like that given in Definition 1.9. Thus, this is a more general formulation of the problem. Any solver can be used to solve the deterministic problem, however, if spectral or finite element methods are being used, we reduce the system into its weak form in the spatial variable anyways and make a discretization in the space of test functions. Since [16] *chooses to break the spatial variable into a finite element expansion, Definition 1.9 (which mimics their setup and notation) goes directly to the spatial weak form.*

We remark here that the coefficients recovered from exactly solving the resulting deterministic system may not be *and are likely not* the gPC coefficients of the true solution. There are special cases where we do obtain exact recovery, such as operators which are linear in the parameters. Proving the consistency of a stochastic Galerkin method necessitates showing some relationship between the recovered coefficients and the true gPC coefficients. This becomes problem dependent and is beyond the scope of these notes. Instead, we focus on sampling methods such as stochastic collocation.

1.7. Stochastic Collocation

Let us review the path we've taken so far for solutions of pointwise parameterized PDE in the form of (1.4). If we first consider the analogous concept of a weak solution in only the parameter domain (which has not yet been discussed), we can view this process as requiring that a solution u induces a functional $\mathcal{L}(x, Z)u(x, Z) - f(x, Z)$ which is zero on $L^2_\pi(\Gamma)$. In order to test that this functional is in fact zero, we apply it to test functions in $w \in L^2_\pi(\Gamma)$. We can then require that the resulting problem in the physical domain is satisfied pointwise or weakly (where the latter results in the joint weak formulation of Definition 1.4).

However, instead of using all of $L^2_\pi(\Gamma)$ as the space of test functions, it suffices to use a basis $\{\phi_\nu\}_{\nu\in\mathbb{N}_0^N}$ (perhaps a gPC basis?). Thus, we can view the weak problem in only the parameter domain as requiring

$$(\mathcal{L}(x)u(x), \phi_\nu)_{L^2_\pi(\Gamma)} = (f(x), \phi_\nu)_{L^2_\pi(\Gamma)} \quad \text{for all } \phi_\nu, \text{ with } \nu \in \mathbb{N}_0^N. \tag{1.38}$$

Notice that this relationship requires that $\mathcal{L}u - f$ is orthogonal to $\operatorname{span}\{\phi_\nu \mid \nu \in \mathbb{N}_0^N\} = L^2_\pi(\Gamma)$. Thus, $\mathcal{L}u - f \equiv 0$ as a functional and by the Riesz-representation theorem is therefore zero as a function in $L^2_\pi(\Gamma)$.

Now in the stochastic Galerkin method, we no longer use the entire space

$$\operatorname{span}\left\{\phi_\nu \mid \nu \in \mathbb{N}_0^N\right\},$$

and instead truncate it. In the case of gPC bases, we can consider various degree rules (1.28)-(1.32) to construct a finite dimensional subspace to test the functional against and produce a discrete solution. Since (at least in the elliptic case), we require the space of test functions to be the same as the space of solutions, we also discretize the parameter dependence of the solution u representing it as a truncated gPC expansion. The system then results in simply finding the coefficients of u in the truncated gPC expansion, and (1.38) reduces to a system of deterministic PDE. In some convenient cases, we can take advantage of structure in the gPC basis to evaluate the integrals arising in (1.38). Often, the integrals involve only polynomials in the parameter domain and can therefore be evaluated using a numerical quadrature rule which is exact for some polynomial degree. However, in general, such techniques are not necessarily available.

With this process in mind as motivation, we now discuss stochastic collocation. There are many ways to link collocation (e.g. pseudo-spectral) techniques to spectral/Galerkin techniques which we may touch on, but for now, we begin by discussing the interpolation approach. For more information, see [**7**, Section 4.4] As previously discussed, the systems resulting from a Galerkin discretization become complicated and very dependent on the gPC basis used.

To keep things simple and remove any dependence on a gPC basis, we instead work pointwise in Z. Let $\left\{Z^{(i)}\right\}_{i=1}^M \subset \Gamma$ be a set of parameter values where we wish to fix the system (1.4) and solve in physical space. We denote these as *collocation points*. We follow the setup of [**16**] where all solutions in physical space are approximated using finite element methods, however, any convergent method can suffice (in fact, if we don't wish to analyze any discretization error in the physical domain, we can assume these solutions are exact). This creates an ensemble $\{u_{J_h}(\cdot, Z^{(i)})\}_{i=1}^{M}$ of approximate solutions in $W_h(\Omega)$ where $W_h(\Omega)$ is a finite element space corresponding to a triangulation $\mathcal{T}_h$ of Ω with maximum mesh size $h > 0$, and $J_h = \dim W_h(\Omega)$. Our goal is to use these sample solutions to reconstruct an approximation to u in the parameter domain.

In light of our insistence on polynoimal expansions of the solution in the parameter domain, we will use this ensemble to construct a global polynomial approximation for $u(Z)$,

$$u_{J_h,M}(x,y) = \sum_{m=1}^{M} c_m(x)\phi_m(Z).$$

For now, we neither specify $\{\phi_m\}_{m=1}^{M}$ as gPC polynomials nor the c_m as the gPC coefficients of $u(\cdot, Z)$. In fact, even when we do let our polynomial basis be a gPC basis, the c_m will not in general correspond to gPC coefficients. We do recall some previous notation however. Supposing that M is determined by the length of the polynomial expansion which is parameterized by some degree parameter p through an index set $\Lambda(p) \subset \mathbb{N}_0^N$, that is, $M := |\Lambda(p)|$, we have $u_{J_h,M}(x,\cdot) \in \mathcal{P}_{\Lambda(p)}(\Gamma)$ for any $x \in \Omega$. Note that we will freely swap between the equivalent indexing schemes $\{\phi_m\}_{m=1}^{M}$ and $\{\phi_\nu\}_{\nu\in\Lambda(p)}$ depending on which is most convenient.

Now, since we consider M collocation points and our expansion of $u_{J_h,M}$ has M terms, we can invert the linear system

$$\sum_{m=1}^{M} c_m(x)\phi_m\left(Z^{(i)}\right) = u_{J_h}\left(x, Z^{(i)}\right) \quad \text{for } i = 1,\ldots,M. \tag{1.39}$$

to obtain the coefficients c_m as linear combinations of the sample solutions. Thus, we arrive at a fully discrete approximation

$$u_{J_h,M} \in W_h(\Omega) \otimes \mathcal{P}_{\Lambda(p)}(\Gamma).$$

where the discretization in the spatial domain is through the finite element space and the discretization in the parameter space is through the polynomial subspace. Notice that the system (1.39) imposes the interpolation condition

$$u_{J_h,M}\left(\cdot, Z^{(i)}\right) = u_{J_h}\left(\cdot, Z^{(i)}\right) \quad \text{for all } i = 1,\ldots,M.$$

Notice that our only use of of the basis $\{\phi_m\}_{m=1}^{M}$ is in the solution of system (1.39). Thus, if this is our only concern, it would make sense to choose the basis as interpolatory over the set of collocation points, that is,

$$\phi_m\left(Z^{(i)}\right) = \delta_{m,i}. \tag{1.40}$$

This reduces solving (1.39) to setting the coefficient c_m to be the approximated solution at $Z^{(m)}$,

$$c_m(x) = u_{J_h}\left(x, Z^{(m)}\right).$$

For an arbitrary sequence of collocation points in *one-dimension*, the Lagrange fundamental polynomials satisfying the delta condition (1.40) can be calculated as

$$\ell_m(Z) = \prod_{\substack{i=1\\ i\neq m}}^{M} \frac{Z - Z^{(i)}}{Z^{(m)} - Z^{(i)}}, \tag{1.41}$$

where each ϕ_m is of degree $M-1$. In the case of multiple dimensions, the tensor product

$$\ell_\nu(Z) = \prod_{n=1}^{N} \ell_{\nu_n}(Z_n)$$

will satisfy (1.40)

$$\ell_\nu\left(Z^{(\eta)}\right) = \prod_{n=1}^{N} \ell_{\nu_n}\left(Z^{(\eta_n)}\right) = \prod_{n=1}^{N} \delta_{\nu_n,\eta_n} = \delta_{\nu,\eta}$$

over the tensor product grid of collocation points,

$$\bigotimes_{n=1}^{N} \left\{Z^{(i)}\right\}_{i=1}^{M} = \left\{\left(Z^{(\eta_n)}\right)_{n=1}^{N}\right\}_{\eta\in\Lambda_{TP}(M)} =: \left\{Z^{(\eta)}\right\}_{\eta\in\Lambda_{TP}(M)}, \tag{1.42}$$

where $\Lambda_{TP}(M)$ is the tensor product index set (1.28).

We can at this point suggest a comparison with Monte-Carlo sampling. In both methods, we obtain an ensemble of solutions on Γ, then reconstruct an approximation to the solution over the entire parameter domain. In Monte-Carlo approximation, our reconstruction is simply averaging over random samples, whereas with stochastic collocation, we are polynomially interpolating *structured samples.* It remains to be discussed how the samples are structured and whether the rate of convergence can beat $O(M^{-1/2})$ obtained in $L^2_\pi(\Gamma)$ for Monte-Carlo sampling. As we have analyzed the convergence of gPC expansions with respect to the regularity of the approximated function, we expect that if our solutions are smooth in the parameter domain (which they often are), we should be able to obtain faster convergence than Monte-Carlo which simply requires that the solution is in $L^2_\pi(\Gamma)$.

1.8. Sparse Grids

The analysis of the interpolation approach of stochastic collocation depends on our choice of collocation points. This section is devoted to a discussion of constructing sparse grids, which carry significantly fewer points than the tensor product grid (1.42). We will also comment on the rates of convergence of the interpolation operator over these sparse grids, offering a comparison with convergence rates of full tensor product grids.

1.8.1. Sparse Grid Construction. The sparse grids used in [**16**] are generalizations of those first proposed in [**26**]. The idea is to construct a sparse proxy to an interpolation operator as a linear combination of tensor product interpolation operators. We begin by clarifying our one-dimensional notation that we combine for multiple dimensions.

For each dimension, we define a one-dimensional *level* of the approximation, $l_n \in \mathbb{N}_+$, $n = 1, \dots N$. This level indexes the one-dimensional set of collocation points used in the nth dimension,

$$\left\{Z^{(k)}_{n,l_n}\right\}_{k=1}^{m(l_n)} \subset \Gamma_n,$$

where $m(l_n)$ is the total number of collocation points used in the nth direction. We require that m satisfies

$$m(l) : \mathbb{N}_+ \to \mathbb{N}_+, \quad m(0) = 0, \quad m(1) = 1, \quad m(l) < m(l+1).$$

The Lagrange interpolation operator in the nth dimension on this collocation grid

$$\mathcal{U}_n^{m(l_n)} : C^0(\Gamma_n) \to \mathcal{P}_{m(l_n)-1}(\Gamma_n)$$

is defined as above,

$$\mathcal{U}_n^{m(l_n)}[v](Z_n) = \sum_{k=1}^{m(l_n)} v\left(Z_{n,l_n}^{(k)}\right) \phi_{n,l_n}^{(k)}(Z_n),$$

where $\phi_{n,l_n}^{(k)} \in \mathcal{P}_{m(l_n)-1}$ is the Lagrange interpolant as in (1.41) corresponding to the point $Z_{n,l_n}^{(k)}$. We also define $\mathcal{U}_n^{m(0)} = 0$, and define the *difference* or *detail operator*

$$\Delta_n^{m(l_n)} = \mathcal{U}_n^{m(l_n)} - \mathcal{U}_n^{m(l_n-1)}.$$

Note then that we have the telescoping identity

$$\mathcal{U}_n^{m(l)} = \sum_{i=1}^{l} \Delta_n^{m(i)}. \tag{1.43}$$

Now, we make our definitions in the multivariate case. We take $l \in \mathbb{N}_+^N$ to be a multi-index capturing the levels in each dimension, and $L \in \mathbb{N}_+$ is the *total level* of the sparse grid. We now tensor product the detail operators

$$\Delta^{m(l)} = \bigotimes_{n=1}^{N} \Delta_n^{m(l_n)}$$

into the N-*dimensional hierarchical surplus operator*, and define the L*th level generalized sparse grid operator* by

$$\mathcal{I}_L^{m,g} = \sum_{g(l) \leqslant L} \Delta^{m(l)}.$$

The function $g : \mathbb{N}_+^N \to \mathbb{N}$ is a strictly increasing function (in the sense that for $i, j \in \mathbb{N}_+^N$, $i < j$ pointwise implies $g(i) < g(j)$) constructing the index set for a given level L. Letting g take on the corresponding rules in (1.28), (1.30), and (1.32), we interpolate over levels in the tensor product, total degree, and hyperbolic cross index sets for "degree" L respectively. The fact that g is strictly increasing implies that the corresponding index set is a *lower set*, that is, if $i < j \in \{\ell \mid g(\ell) \leqslant L\}$, then $i \in \{\ell \mid g(\ell) \leqslant L\}$. We discuss the relationship between the index sets used for collocation and corresponding polynomial index sets below.

Notice that with the tensor product rule (defined for level index sets in (1.48)), applying (1.43) in each dimension gives

$$\mathcal{I}_L^{m,\mathrm{TP}} = \bigotimes_{n=1}^{N} \mathcal{U}_n^{m(L)}, \tag{1.44}$$

that is, we obtain the N-dimensional interpolation operator on the tensor grid (1.42). Thus, by truncating the sum of hierarchical surplus operators over the tensor product level index set, we are truncating the decomposition of the tensor product grid interpolation operator, where the decomposition is in terms of multi-level tensor product interpolating operators (since the hierarchical surplus operators are linear combinations of tensor product interpolating operators, see Proposition 1.2 below). The benefits of the sparse grid interpolation are generally analyzed by considering the amount of error that the level index set captures in its corresponding interpolations [**27**].

Proposition 1.2. *We can write the sparse grid operator as the linear combination of tensor product interpolating operators*

$$\mathcal{I}_L^{m,g} = \sum_{g(l) \leqslant L} \sum_{k \in \{0,1\}^N} (-1)^{|k|} \mathcal{U}^{m(l-k)} = \sum_{g(j) \leqslant L} \sum_{\substack{k \in \{0,1\}^N \\ g(j+k) \leqslant L}} (-1)^{|k|} \mathcal{U}^{m(j)},$$

where

$$\mathcal{U}^{m(l)} = \bigotimes_{n=1}^{N} \mathcal{U}_n^{m(l_n)}.$$

Proof. The first equality is given by [27, Proposition 1.2]. The second is found from the change of variables $j = l - k$ and the fact that $g(j+k) \leqslant L$ implies $g(j) \leqslant L$. □

In light of Proposition 1.2, the necessary set of collocation points for the sparse grid operator (that is, the titular *sparse grid*), is given as

$$\mathcal{H}_L^{m,g} = \bigcup_{g(l)\leqslant L} \bigotimes_{n=1}^{N} \left\{ Z_{n,l_n}^{(k)} \right\}_{k=1}^{m(l_n)}, \tag{1.45}$$

that is, the union over the index set of levels of all tensor grids defined by these level multi-indices. We will use M_L to denote the size of the sparse grid.

Now that we are only using interpolants of the solution rather than gPC expansions, the coefficients no longer directly translate to solution statistics. On the other hand, the interpolating formula gives a natural quadrature formula to calculate these QoI:

$$\mathbb{E}[u_{J_h,M}](x) = \sum_{i=1}^{M_L} u_h\left(x, Z^{(i)}\right) w_i,$$

with quadrature weights

$$w_i = \mathbb{E}\phi_i = \int_\Gamma \phi_i\,(z)\,\pi(z)\;dz,$$

and

$$\mathbb{V}[u_{J_h,M}](x) = \sum_{i=1}^{M_L} \tilde{w}_i u_{J_h}^2\left(x, Z^{(i)}\right) - \mathbb{E}[u_{J_h,M}]^2(x)$$

with weights

$$\tilde{w}_i = \mathbb{V}\phi_i,$$

so long as the interpolating polynomials are independent (which does not seem like a natural assumption). We can then simply pre-compute the weights on the desired sparse grid and then plug in the interpolating coefficients resulting from an ensemble of solutions to the desired problem.

It is also possible to describe the associated polynomial basis corresponding to the sparse grid operator. This was first used [5] to provide a fair comparison between stochastic Galerkin and collocation where both methods made use of the same finite dimensional polynomial subspace of $L^2_\pi(\Gamma)$. In the case where the level l is simply the number of collocation points used, that is, $m(l) = l$, the polynomial index set corresponding to the sparse grid is

$$\Lambda_g(L) = \left\{ \nu \in \mathbb{N}_0^N \mid g(\nu + 1) \leqslant L \right\}, \tag{1.46}$$

where rather than a gPC basis, the underlying polynomial basis is

$$\mathcal{P}_{\Lambda_g(L)} = \operatorname{span}\left\{ \prod_{n=1}^{N} Z_n^{\nu_n} \mid \nu \in \Lambda_g(L) \right\}, \tag{1.47}$$

the tensor product of Taylor monomials. Thus, for

$$g(l) = \max_{n=1,\dots,N} (l_n - 1), \tag{1.48}$$

$$g(l) = \sum_{n=1}^{N} (l_n - 1), \tag{1.49}$$

$$g(l) = \sum_{n=1}^{N} \lg(l_n) \tag{1.50}$$

we obtain the tensor-product, total degree, and hyperbolic cross polynomials respectively. Further modifications can be made for more complicated m (like the definition discussed in Section 1.8.2 for Clenshaw-Curtis points) which can describe the popular sparse Smolyak subspace as well. See [5, Proposition 1] for further details.

1.8.2. Choice of Collocation Points. The survey [16] discusses two types of collocation points on the interval $\Gamma_n = [-1, 1]$: Clenshaw-Curtis points and Gaussian points. *Clenshaw-Curtis* points are derived as extrema of Chebyshev polynomials,

$$Z_l^{(i)} = -\cos\left(\frac{\pi(i-1)}{m(l)-1}\right), \quad \text{for } i = 1, \dots, m(l),$$

where $Z_l^{(1)} = 0$ if $m(l) = 1$. In this case, we should choose M to grow exponentially,

$$m(l) = \begin{cases} 1, & \text{if } l = 1 \\ 2^{l-1} + 1 & \text{if } l > 1 \end{cases}.$$

Then the Clenshaw-Curtis (CC) points are *nested* at each level. This is a very convenient, since all corresponding tensor product grids must then necessarily be nested, and points are reused at all levels of the sparse grid construction. Additionally, sparse grids at different levels are nested. Thus, if one wishes to increase the level of approximation, deterministic solutions need only be calculate on the new nodes, reusing the previous computations. Another reason to use CC points (or any set of nested points) is that when g is the total degree rule, the sparse grid operator does in fact interpolate the points in the sparse grid [6]. Additionally, as we have near logarithmic bounds on the Lebesgue constant (that is, the operator norm of the one-dimensional interpolation operator $\mathcal{U} : C_0 \to C_0$), sparse grid interpolation using CC points is optimal up to logarithmic factors in terms of the number points in the sparse grid [6]. Finally CC points have nice quadrature properties for "not too analytic" functions. In particular, Clenshaw-Curtis knots allow for convergence near the level of Gaussian points [28].

Gaussian points are derived as the "optimal" quadrature abscissas for polynomials integrated against a weight π_n in the sense that univariate Gaussian quadrature with $M+1$ nodes is exact for polynomials of up to degree $2M+1$. In the context of stochastic problems, we let π be the density of the sequence of random variables Z satisfying the finite dimensional noise assumption. In the case that the components of Z actually are independent, we define each π_n as the density of Z_n. The *Gaussian points* $\left\{Z_l^{(i)}\right\}_{i=1}^{m(l)}$ are defined as the zeros of $\phi_{m(l)}(Z)$, the $m(l)$th gPC polynomial corresponding to the weight π_n. On the other hand, if the components of Z are not independent, π, the density of Z does not factor. We instead consider an auxiliary product density $\hat{\pi} : \Gamma \to \mathbb{R}^+$ given by

$$\hat{\pi}(y) = \prod_{n=1}^{N} \hat{\pi}_n(Z_n).$$

We also assume that $\|\pi/\hat{\pi}\|_\infty < \infty$. This allows us to estimate quantities in the original probability space such as

$$\|u\|^2_{H^1_0(\Omega)\otimes L^2_\pi(\Gamma)} \leqslant \left\|\frac{\pi}{\hat{\pi}}\right\|_\infty \|u\|^2_{H^1_0(\Omega)\otimes L_{\hat{\pi}}(\Gamma)},$$

the latter of which can be approximated and analyzed by considering a sparse grid with respect to the Gaussian points of the factors of $\hat{\pi}$.

1.8.3. Convergence Results. We now discuss the error in the exact solution and the solution arising from stochastic collocation with Lagrange interpolation on tensor and sparse grids. As in the survey [**16**], we provide the general outline and do not consider the proofs of the main results which in general require detailed results from polynomial approximation theory which is outside the scope of these notes.

1.8.3.1. *Regularity Assumptions.* We begin by assuming that the true solution admits an analytic extension to the complex plane in each dimension in a neighborhood of the polyellipse

$$\mathcal{E}_\mu = \bigotimes_{1\leqslant n\leqslant N} \operatorname{Int}\{z_n \in \mathbb{C} \mid z_n = \cosh(\mu_n)\cos(\theta) + i\sinh(\mu_n)\sin(\theta),\ \theta \in [0, 2\pi)\} \tag{1.51}$$

for all $x \in \Omega$. In general, this is product of the regions of convergence of the one-dimensional Chebyshev or Legendre series for an analytic function where μ_n is the smallest quasi-radial coordinate of the polyellipse passing through a singularity of the restriction of u to $\mathbb{C}$ in the nth dimension [**7**, Theorem 7].

In the special case of differential operators $\mathcal{L}$ which have *affine* dependence on the parameter sequence Z, that is, we can express the operator as

$$\mathcal{L} = \mathcal{L}_0 + \sum_{n=1}^{N} Z_n \mathcal{L}_n, \tag{1.52}$$

this assumption of analyticity can be clarified as being separated into each parameter dimension as follows. We restrict our solution u to vary only in the nth parametric dimension by fixing a point in

$$\Gamma^*_n = \prod_{j=1 j\neq n}^{N} \Gamma_j$$

and $Z^*_n \in \Gamma^*_n$. Then there exist regions $\Sigma_n := \{z \in \mathbb{C} \mid \operatorname{dist}(z, \Gamma_n) \leqslant \tau_n\}$ for some constants $\tau_n > 0$ (independent of our choice of Z^*_n) such that $u(x, Z^*_n, \cdot)$ admits an analytic extension on Σ_n. Another way of stating this is that the function $u : \Gamma_n \to C^0(\Gamma^*_n; H^1_0(\Omega))$ admits an analytic extension on Σ_n. It is this affine case to which we restrict our current (and the majority of our future) discussion. Note that the stationary (that is, time *independent*) diffusion equation (cf. (1.37) where the time dependence is removed) fits this model, where we take

$$\mathcal{L}_n = -D \cdot (\hat{A}_n(x)D).$$

1.8.3.2. *Typical Error Analysis.* We now consider our discrete solution obtained from stochastic collocation on a sparse grid $u_{J_h,M_L} := \mathcal{I}^{m,g}_L[u_{J_h}]$ and provide bounds on its difference from the exact solution. We proceed by splitting this error into separate pieces:

$$\begin{aligned}\|u - u_{J_h,M}\| &\leqslant \|u - u_N\| + \|u_N - u_{N,J_h}\| + \left\|u_{N,J_h} - \mathcal{I}^{m,g}_L[u_{N,J_h}]\right\| \\ &:= I + II + III,\end{aligned}$$

where we let $\|\cdot\| := \|\cdot\|_{L^2_\pi(\Gamma)\otimes H^1_0(\Omega)}$. Notice that by bounding the error in this norm, we can obtain physical bounds on the expectation of the error as

$$\|\mathbb{E}[u - u_{J_h,M_L}]\|_{H^1_0(\Omega)} \leqslant \mathbb{E}\|u - u_{J_h,M_L}\|_{H^1_0(\Omega)} \leqslant \|u - u_{J_h,M_L}\|_{L^2_\pi(\Gamma)\otimes H^1_0(\Omega)},$$

by Jensen's and Hölder's inequalities.

The error I is a result of truncating the original differential equation to depend on only N parameters. As discussed in section 1.2.2, we can generally use a KL expansion to parameterize any stochastic process in the original equation in terms of a countable sequence Z. Truncating this sequence gives some amount of "problem level" truncation error which requires further analysis. The structure of the KL expansion and knowledge of the underlying covariance function are useful for bounding this type of error [**20**], but we take the perspective of the *finite-dimensional* noise assumption, thus assuming $I = 0$. Thus, in the remaining errors, we replace u_N by u.

The error II is due to the finite element approximation in the spatial variable x. Since all sampling methods considered will necessitate a spatial solver, this error will persist in each analysis. As such we refer to the finite element method theory which gives, depending on the spatial regularity of $u(\cdot, Z)$,

$$II = O(h^t)$$

where we recall that h parameterizes the mesh size of the finite element approximation. In general, t is related to the polynomial degree of the finite element basis. The specific regime for t depends on the spatial regularity of u, and the implicit constant is also dependent on u and its relationship to the parameter domain.

Thus, we arrive at the error III resulting from the interpolation of the sample solutions found from the collocation grid. Again, as in [**16**], we consider the sparse grid corresponding to CC points with g determined by the total degree rule. Analogous results hold for Gaussian abscissas where the error considered is *only* able to be bounded in $\|\cdot\|_{H^1_0(\Omega)\otimes L^2_\pi(\Gamma)}$ rather than the more restrictive $L^\infty(\Gamma)$ norm in the parameter domain as we consider ahead. A standard reduction at this point is to assume that the spatially discretized solution u_h is unaffected in its parametric regularity, and thus, all assumptions of analyticity of u carry over to u_h. Thus, we consider III with u_h replaced by u. As previously mentioned, since the errors being bounded are from polynomial interpolation and the traditional literature handles bounding the error of approximating continuous functions by polynomials we are motivated to instead consider all error in $\|\cdot\| := \|\cdot\|_{H^1_0(\Omega)\otimes L^\infty(\Gamma)}$. Results in $L^2_\pi(\Gamma)$ follow immediately since $(\Gamma, \mathcal{B}(\Gamma), \pi\ dz)$ is assumed to be a probability space.

The first step is to provide bounds for one-dimensional interpolation, that is, consider

$$\left\| u - \mathcal{U}_n^{m(l_n)}[u] \right\|_\infty,$$

for $u \in C^0(\Gamma_n; H^1_0(\Omega))$. For $\mathbb{L}_{m(l_n)}$ the Lebesgue constant of $\mathcal{U}_n^{m(l_n)}$, (that is, the operator norm of $\mathcal{U}_n^{m(l_n)} : C^0(\Gamma_n) \to \mathcal{P}_{m(l_n)-1}$ where $\mathcal{P}_{m(l_n)-1}$ is treated as subspace of $C^0(\Gamma_n)$) which for CC points has the upper bound

$$\begin{aligned}
\mathbb{L}_{m(l_n)} &\leqslant \frac{2}{\pi}\log(m(l_n) - 1) + 1 \\
&= \frac{2}{\pi}\log\left(2^{l_n - 1} + 1 - 1\right) + 1 \\
&= \frac{2}{\pi}\log(2)(l_n - 1) + 1 \\
&\leqslant l_n.
\end{aligned}$$

Letting $E_{m(l_n)-1} = \min_{w \in \mathcal{P}_{m(l_n)-1}} \|u - w\|_{C^0}$ and w^* the minimizing polynomial, we have

$$\begin{aligned} \left\|u - \mathcal{U}_n^{m(l_n)}[u]\right\|_{C^0} &\leqslant \|u - w^*\|_{C^0} + \left\|\mathcal{U}_n^{m(l_n)}[u - w^*]\right\|_{C^0} \\ &\leqslant \left(1 + \mathbb{L}_{m(l_n)}\right) E_{m(l_n)-1}. \end{aligned} \tag{1.53}$$

We now make use of the following lemma on $E_{m(l_n)-1}$ and best polynomial approximations.

Lemma 1.1 ([20], Lemma 4.4). *For* $u \in C^0(\Gamma_n; H_0^1(\Omega))$ *admitting an analytic extension in* $\Sigma_n = \{z \in \mathbb{C} \mid \text{dist}(z, \Gamma_n) \leqslant \tau_n\}$*, then*

$$E_{m(l_n)-1} \leqslant \frac{2}{e^{2r_n} - 1} e^{-2r_n(m(l_n)-1)} \max_{z \in \Sigma_n} \|u(z)\|_{H_0^1(\Omega)}, \tag{1.54}$$

where

$$0 < r_n = \frac{1}{2} \log\left(\frac{2\tau_n}{|\Gamma_n|} + \sqrt{1 + \frac{4\tau_n^2}{|\Gamma_n|^2}}\right).$$

Proof Sketch. The main idea is to bound the error by the error of a Chebyshev polynomial approximation of degree $m(l_n) - 1$ denoted $T_{m(l_n)-1}(z)$. The domain of convergence given in (1.51) as $\mathcal{E}_{r_n}$ is the ellipse of quasi-radial coordinate r_n to be determined. Under the change of variables $z = \cos(\theta)$, $T_{m(l_n)-1}(\cos(\theta))$ becomes the truncation of the Fourier series for $u(\cos(\theta))$. In order for this series to converge, since the imaginary part of θ is the quasi-radial coordinate r_n, the Fourier coefficients a_k must decay at a rate of $O(e^{-kr_n})$. The constant in this convergence can be determined in terms of u as $\|a_k\|_{H_0^1(\Omega)} \leqslant 2e^{-kr_n} \max_{z \in \Sigma_n} \|u(z)\|_{H_0^1(\Omega)}$ (see the argument used in Proposition 3.3 for a rigorous proof using only the formula for the Chebyshev coefficients) . Thus, truncating the Chebyshev series and using that the L^∞ norm of Chebyshev polynomials is bounded by 1,

$$\begin{aligned} E_{m(l_n)-1} &\leqslant \sum_{k=m(l_n)}^{\infty} \|a_k\|_{H_0^1(\Omega)} \\ &\leqslant \sum_{k=m(l_n)}^{\infty} 2e^{-kr_n} \max_{z \in \Sigma_n} \|u(z)\|_{H_0^1(\Omega)} \\ &\leqslant \frac{2e^{-(m(l_n)-1)r_n - r_n}}{1 - e^{-r_n}} \max_{z \in \Sigma_n} \|u(z)\|_{H_0^1(\Omega)} \\ &= \frac{2e^{-(m(l_n)-1)r_n}}{e^{r_n} - 1} \max_{z \in \Sigma_n} \|u(z)\|_{H_0^1(\Omega)}. \end{aligned}$$

The last step is solving for the quasi-radial coordinate r_n as defining the largest ellipse $\mathcal{E}_{r_n}$ which fits in Σ_n. This is determined to be

$$r_n = \log\left(\frac{2\tau_n}{|\Gamma_n|} + \sqrt{1 + \frac{4\tau_n^2}{|\Gamma_n|^2}}\right).$$

Taking $r_n \leftarrow \frac{r_n}{2}$ gives the desired result. □

Combining (1.53) and (1.54) gives

$$\left\|u - \mathcal{U}_n^{m(l_n)}[u]\right\|_{\infty} \leqslant \frac{4}{e^{2r_n} - 1} l_n e^{-2r_n 2^{l_n}} \theta(u), \tag{1.55}$$

and we can bound the one-dimensional detail operator by twice the larger error of the two interpolating operators,

$$\left\|\Delta_n^{m(l_n)}[u]\right\|_\infty \leqslant \frac{8}{e^{2r_n}-1} l_n e^{-2r_n 2^{l_n-1}}\theta(u),$$

where we take $\theta(u)$ to be the maximum of the quantities $\max_{z\in\Sigma_n}\|u\|_{H_0^1(\Omega)}$ over all N dimensions. In the below estimates, we assume without loss of generality that $u \leftarrow \frac{u}{\theta(u)}$ and thus $\theta(u)=1$.

1.8.3.3. *Tensor Grid Error.* In this section, we use the one dimensional results to derive error bounds on tensor grid stochastic collocation. The result given is for CC full tensor grids, whereas the original result in [4] is for Gaussian full tensor grids.

Theorem 1.6 (cf. [4], Theorem 4.1). *For CC tensor product grid (that is, the sparse grid using* g *defined by* (1.48)*) of level* L *and* $u \in L_\pi^2(\Gamma)\otimes H_0^1(\Omega)$ *satisfying the affine version of analyticity with quasi-radial analyticity parameters* r_n*, we have*

$$\left\|u - \mathcal{I}_L^{m,TP}\right\| \leqslant C(r)L^N M_L^{-r/N}$$

where $r = \min_{1\leqslant n\leqslant N} r_n$.

Proof. We proceed as in the style of the proof of [20, Lemma 3.4] which extended the original result from [4]. Though a bit more involved, this method allows a simple derivation of the factor L^N (which we do not compare to the original suppressed constant factor depending on N in [4, Theorem 4.1]).

We first make note of the implicit identification

$$\mathcal{U}_{(n)}^{m(L)} : C^0(\Gamma; H_0^1(\Omega)) \to C^0(\Gamma; H_0^1(\Omega))$$
$$\mathcal{U}_{(n)}^{m(L)} := \mathcal{U}_n^{m(L)} \otimes \bigotimes_{\substack{i=1\\ i\neq n}}^N I_i$$

where $I_i : C^0(\Gamma_n; H_0^1(\Omega)) \to C^0(\Gamma_n; H_0^1(\Omega))$ is the one-dimensional identity operator in the ith dimension. We then have

$$\left\|\mathcal{U}_{(n)}^{m(L)}\right\| \leqslant \left\|\mathcal{U}_n^{m(L)}\right\|$$

where $\|\cdot\|$ denotes the operator norm. Henceforth, we thus drop the parentheses on the n-dimensional operator.

Now, we calculate

$$\begin{aligned} I - \mathcal{I}_L^{m,TP} &= I - \bigotimes_{n=1}^N \mathcal{U}_n^{m(L+1)} \qquad \text{by (1.44)}\\ &= I - \bigotimes_{n=1}^{N-1} \mathcal{U}_n^{m(L+1)} \otimes \left(\mathcal{U}_N^{m(L+1)} - I\right) - \bigotimes_{n=1}^{N-1} \mathcal{U}_n^{m(L+1)} \otimes I\\ &= \bigotimes_{n=1}^{N-1} \mathcal{U}_n^{m(L+1)} \otimes \left(I - \mathcal{U}_N^{m(L+1)}\right) + \left(I - \bigotimes_{n=1}^{N-1} \mathcal{U}_n^{m(L+1)}\right) \otimes I. \end{aligned}$$

Applying this identity recursively on the last term gives

$$I - \mathcal{I}_L^{m,TP} = \sum_{d=1}^N \bigotimes_{n=1}^{d-1} \mathcal{U}_n^{m(L+1)} \otimes \left(I - \mathcal{U}_d^{m(L+1)}\right) \otimes I^{N-d}.$$

By the previous discussion and our one-dimensional estimates

$$\begin{aligned}\left\|\left(I-\mathcal{J}_L^{m,TP}\right)[u]\right\|_{C^0(\Gamma;H_0^1(\Omega))} &\leqslant \sum_{d=1}^{N}\prod_{n=1}^{d-1}\left\|\mathcal{U}_n^{m(L+1)}\right\|\left\|I-\mathcal{U}_d^{m(L+1)}\right\|\|I\|^{N-d}\\ &\leqslant \sum_{d=1}^{N}\mathbb{L}_{m(L+1)}^{d-1}\frac{4}{e^{2r_d-1}}(L+1)e^{-2r_d 2^{L+1}}\\ &\leqslant \frac{4}{e^{2r}-1}e^{-2r2^{L+1}}\sum_{d=1}^{N}(L+1)^d\\ &\leqslant C(r)L^N e^{-r2^{L+2}}.\end{aligned}$$

Note that we have made use of the embedding $C^0(\Gamma_n;H_0^1(\Omega))\subset C^0(\Sigma_n;H_0^1(\Omega))$ to identify $I-\mathcal{U}_n^{m(L+1)}$ with its operator on the region of analyticity of u, Σ_n. As a result, its norm is multiplied by $\|u\|_{C^0(\Sigma_n;H_0^1(\Omega))}$ which we have assumed is less or equal to $\theta(u)\leqslant 1$.

The final step is to relate the level to the number of points in the tensor grid M_L. Since each dimension uses $m(L+1)=2^L+1$ points, we have $M_L=(2^L+1)^N$. Taking logarithms gives

$$\frac{1}{N}\log(M_L)=\log\left(2^L+1\right)\leqslant 2^L-1\leqslant 2^{L+2}.$$

Finally then

$$C(r)L^N e^{-r2^{L+2}}\leqslant C(r)L^N M_L^{-r/N},$$

which gives us the desired bound. □

1.8.3.4. *Isotropic Sparse Grid Error.* We now give the result of applying this one-dimensional argument dimension by dimension over the sparse grid operator. The argument proceeds much the same as the proof of Theorem 1.6, but more care is necessary to efficiently bound the sums of one-dimensional interpolation operators over the total degree index sets.

Theorem 1.7 ([**20**], Theorems 3.10 and 3.11). *For isotropic CC sparse grids with total degree rule* (1.49)*, and* $u\in L_\pi^2(\Gamma)\otimes H_0^1(\Omega)$ *satisfying the affine version of analyticity, we have the following estimates.*

- *Algebraic convergence* ($0<L\leqslant\frac{N}{\log(2)}$)

$$\left\|u-\mathcal{J}_L^{m,g}[u]\right\|_\infty\leqslant C_0(r)\max\{1,C_1(r)\}^N M_L^{-\mu_1}$$
$$\textit{with } \mu_1=\frac{r}{1+\log(2N)},\quad r=\min_{1\leqslant n\leqslant N} r_n.$$

- *Subexponential convergence* ($L>\frac{N}{\log(2)}$)

$$\left\|u-\mathcal{J}_L^{m,g}[u]\right\|_\infty\leqslant C_2(r)\max\{1,C_1(r)\}^N M_L^{\mu_3}e^{-\frac{Nr}{2^{1/N}}M_L^{\mu_2}}$$
$$\textit{with } \mu_2=\frac{\log(2)}{N(1+\log(2N))}\ \textit{and}\ \mu_3=\frac{C_4(r)}{1+\log(2N)}.$$

1.8.3.5. *Anisotropic Sparse Grid Error.* The results corresponding to sparse grids can be further accelerated with anisotropy. So far, we have provided equal weighting in each dimension when constructing our grid indexing scheme, producing symmetric grids. However, if we know that our function is "less analytic" in certain dimensions as measured by the quasi-radii r_n of analyticity regions, we can shift our attention to that dimension by providing more collocation points. We reflect this by weighting the indices in the $g(l):\mathbb{N}_+^N\to\mathbb{N}$ formulae. Thus, we introduce a weight vector $\alpha \in \mathbb{R}^N_+$ with $\alpha_{\min} = \min_{1\leqslant n\leqslant N} \alpha_n$, and use it to weight the indices in the level index sets $\{l \mid g(l) \leqslant L\}$. In the case of the sparse Smolyak grid with CC points and total degree rule, we take

$$g(l) = \sum_{n=1}^{N} \alpha_n(l_n - 1) \leqslant \alpha_{\min}L. \tag{1.56}$$

A dimension with a higher weight is thus less likely to contribute indices to the index set.

As mentioned above, a natural (and as proved in [**19**], optimal) choice of weights is to take $\alpha_n = r_n$, $\alpha_{\min} = r = \min_{1\leqslant n\leqslant N} r_n$, and $\mathcal{R}(N) = \sum_{n=1}^{N} r_n$, since those dimensions in which the function is smoother get penalized more in the weighting scheme. This will mean that the dimensions in which the exponential rate of convergence in the one dimensional approximation (1.55) are lower will be used at higher levels than their peers, boosting the exponential rate of convergence. Combining these one-dimensional estimates in a similar manner as the isotropic sparse grid case where in particular, care must be taken to account for the reduced anisotropic number of points resulting from the anisotropic sparse grid, we obtain an improvement on Theorem 1.7 given as the following convergence results.

Theorem 1.8 ([**19**] Theorem 3.8). *For anisotropic CC sparse grids with total degree rule* (1.56)*, and* $u \in L^2_\pi(\Gamma) \otimes H^1_0(\Omega)$ *satisfying our assumption of analyticity in the affine case, we have the following estimates.*

- *Algebraic convergence* ($0 < L \leqslant \frac{\mathcal{R}(N)}{r\log(2)}$)
$$\left\|u - \mathcal{I}^{m,g}_L[u]\right\|_\infty \leqslant \hat{C}(r,N)M_L^{-\mu_1}$$
$$\textit{with } \mu_1 = \frac{r(\log(2)e - 1/2)}{\log(2) + \sum_{n=1}^{N} r/r_n}$$
- *Subexponential convergence* ($L > \frac{\mathcal{R}(N)}{r\log(2)}$)
$$\left\|u - \mathcal{I}^{m,g}_L[u]\right\|_\infty \leqslant \hat{C}(r,N)\max\{1, C_1(r)\}^N M_L^{\mu_2} e^{-\mathcal{R}(N)M_L^{\mu_2}}$$
$$\textit{with } \mu_2 = \frac{\log(2)}{\mathcal{R}(N)\left(\log(2) + \sum_{n=1}^{N} r/r_n\right)}.$$

For sequences $r_n \to \infty$*, the constant* $\hat{C}(r,N)$ *is bounded as* $N \to \infty$*.*

The final remark that $r_n \to \infty$ is a natural assumption for problems in the setup we've considered. It is generally the case that for affine parametric operators derived from KL expansions, the size of the analyticity regions of resulting solutions are inversely proportional to the product of the infinity norm of eigenfunctions and the square root of eigenvalues in the KL expansion. Since this product tends to zero, we observe increasingly large regions of analyticity in N.

1.8.3.6. *Discussion, Comparison, and the Curse of Dimensionality.* We first notice that the subexponential decay is always asymptotically faster than the algebraic decay. However, the subexponential decay usually requires a high level sparse grid which is less computationally feasible. Thus, we restrict our attention to comparing the algebraic convergence.

We first notice that the tensor grid result suffers from the curse of dimensionality. As N increases, the algebraic rate of convergence decays, even worse than Monte-Carlo (when exactly this occurs depends on the analyticity region of the solution u). On the other hand, the exponent in the algebraic convergence for the isotropic sparse grid decays in N at a rate of $\log^{-1}(2N) \geqslant \frac{1}{N}$. Thus, the sparse grid is able to reduce the curse of dimensionality to only logarithmic decay in N. Additionally, for certain amounts of analyticity the term $\max(1, C_1(r))^N$ may be bounded in N which further reduces any deterioration within high dimensional problems. Finally, in the anisotropic sparse grid, we see that if $\sum_{n=1}^{\infty} r/r_n < \infty$, the exponent in the algebraic error is no longer dependent on N, and only depends on the bound of this sum. Since the constant $\widehat{C}(r,N)$ is also bounded in N when $r_n \to \infty$, we know then that the entire convergence rate is *independent* of N, and there is no curse of dimensionality. For functions with large and quickly growing regions of analyticity, the algebraic convergence rates can then rival and surpass those of Monte-Carlo. It turns out that a similar rate of convergence holds even for anisotropic full tensor grids, however, the constant is worse.

1.8.4. Sparse Grids for Optimal Polynomial Subspaces. By the previously considered setup, the resulting interpolation is in a polynomial subspace entirely determined by the sparse grid (recall the space spanned by Taylor monomials defined by (1.47) indexed by (1.46)). However, as we discussed in Section 1.5.3, we can attempt to restrict our approximate solution in parameter space to lie in the span of a set of multivariate polynomials indexed by an optimal or at least quasi-optimal index set $\Lambda_M^{q-\text{opt}}$. For the following discussion, we again consider the space of Taylor monomials, that is

$$\mathcal{P}_{\Lambda_M^{q-\text{opt}}} = \text{span}\left\{ \phi_\nu(Z) = \prod_{n=1}^{N} Z_n^{\nu_n} \mid \nu \in \Lambda_M^{q-\text{opt}} \right\}.$$

Assuming that this polynomial subspace is quasi-optimal in the L^∞ sense, that is

$$\left\| u - u_{\Lambda_M^{q-\text{opt}}} \right\|_{L^\infty(\Gamma)} \approx \inf_{w \in \mathcal{P}_\Lambda \text{ s.t. } |\Lambda| = M} \|u - w\|_{L^\infty(\Gamma)},$$

we can hope that the sparse grid polynomial subspace captures the quasi-optimal subspace. However, the index set $\Lambda_M^{q-\text{opt}}$ may be much smaller than a sparse grid index set containing it, making our estimates on the quasi-best M-term error useless. Our goal then is to produce a sparse grid method conforming to the quasi-optimal index set.

We assume that the quasi-optimal index set $\Lambda_M^{q-\text{opt}}$ is a lower set, and define the quasi-optimal sparse grid interpolating operator similarly to the previous discussion as

$$\mathcal{I}_{\Lambda_M^{q-\text{opt}}} = \sum_{\nu \in \Lambda_M^{q-\text{opt}}} \Delta^{m(\nu+1)}.$$

and the resulting sparse grid

$$\mathcal{H}_{\Lambda_M^{q-\text{opt}}} = \bigcup_{\nu \in \Lambda_M^{q-\text{opt}}} \bigotimes_{n=1}^{N} \left\{ Z_{n,\nu_n+1}^{(k)} \right\}_{k=1}^{m(\nu_n+1)},$$

for some collection of one-dimensional collocation points to be determined. By virtue of an analog of Proposition 1.2 and the fact that $\Lambda_M^{q-\text{opt}}$ is a lower set, we immediately obtain $\mathcal{I}_{\Lambda_M^{q-\text{opt}}} : C^0(\Gamma) \to \mathcal{P}_{\Lambda_M^{q-\text{opt}}}$. Additionally, we want as little waste as possible in the number of interpolation points we use, and so we require $M = \left| \mathcal{H}_{\Lambda_M^{q-\text{opt}}} \right|$. This is satisfied when we take m to be the identity, and require that successive levels of collocation points are nested, i.e., increasing from ν_n to $\nu_n + 1$ just adds one additional collocation point.

We can then use the argument to derive (1.53) on the quasi-optimal sparse grid operator giving

$$\left\| u - \mathcal{I}_{\Lambda_M^{q-\text{opt}}}[u] \right\|_{L^\infty(\Gamma)} \leqslant \left(1 + \mathbb{L}_{\Lambda_M^{q-\text{opt}}} \right) \left\| u - u_{\Lambda_M^{q-\text{opt}}} \right\|_{L^\infty(\Gamma)},$$

where $\mathbb{L}_{\Lambda_M^{q-\text{opt}}}$ is the Lebesgue constant for the quasi-optimal sparse grid interpolating operator. Thus, the goal is to add new points to our running list of collocation points in a manner that keeps the Lebesgue constant low as a function of M. The discussion in [**16**] appeals to greedy searches, where for the $K-1$ one-dimensional collocation points

$$\left\{Z^{(k)}\right\}_{k=1}^{K-1} =: \mathcal{Z}_{K-1}$$

(where we are suppressing any dependence on a dimension n), the next point is chosen using rules such as

$$\textbf{(i)} \quad Z^{(K)} = \arg\max_{\xi\in\Gamma} \prod_{k=1}^{K-1} \left|\xi - Z^{(k)}\right|$$

$$\textbf{(ii)} \quad Z^{(K)} = \arg\max_{\xi\in\Gamma} \lambda_{\mathcal{Z}_{K-1}}(\xi)$$

$$\textbf{(iii)} \quad Z^{(K)} = \arg\min_{\xi\in\Gamma} \max_{y\in\Gamma} \lambda_{\mathcal{Z}_{K-1},\xi}(y),$$

where $\lambda_{\mathcal{Z}}$ is the Lebesgue function

$$\lambda_{\mathcal{Z}}(\zeta) := \max_{\|u\|_{L^\infty(\Gamma)}\leqslant 1} |\mathcal{U}_{\mathcal{Z}}[u](\zeta)| = \sum_{k=1}^{K} |\ell_k(\zeta)|,$$

for ℓ_k the kth Lagrange interpolating polynomial corresponding to the collocation points $\mathcal{Z}$. Note that $\mathbb{L}_{\mathcal{Z}} = \max_{\zeta\in\Gamma}\lambda_{\mathcal{Z}}(\zeta)$. We can summarize these rules as follows:

(i) The next point should be farthest away from the current points in a sense similar to geometric mean.

(ii) The next point should be that which maximizes the Lebesgue function. We can think of this point as (at least) one of the points that produces the L^∞ error in the interpolation.

(iii) The next point should minimize the resulting Lebesgue constant.

As of 2015, no theoretical bounds were available for the Lebesgue constants, but experiments show promising results. A final remark is that attempting to produce an N-dimensional set of interpolating points using the aforementioned algorithms forgoing the sparse grid structure entirely is "complex, ill-conditioned, and computationally impractical for more than $N = 2$" [**16**]. This is also somewhat supported by a previous assertion that even constructing the N-dimensional Lagrange interpolating polynomials and corresponding interpolation operator through these points is "not an easy matter," but "there exist means for doing so" [**16**].

CHAPTER 2

Compressive Sensing for Function Approximation

2.1. Sparse Polynomial Interpolation

We now return to polynomial interpolation on bases other than Lagrange interpolating polynomials. For simplicity, we restrict our attention to scalar valued functions $u : \Gamma \to \mathbb{C}$ (where, to match the analysis in [24], we allow our functions to take values over the complex numbers) and suppose that u is represented as a gPC expansion

$$u(Z) = \sum_{\nu \in \Lambda \subset \mathbb{N}_0^N} \hat{u}_\nu \phi_\nu(Z),$$

where $\{\phi_\nu\}_{\nu \in \mathbb{N}_0^N}$ is orthonormal with respect to the orthogonalization measure $\pi\ dz$. Supposing that we have truncated our expansion to a finite index set with $|\Lambda| = M$, and given a set of interpolating points $\left\{Z^{(k)}\right\}_{k=1}^{K} \subset \Gamma$ (where we now allow $K \neq M$), if we obtain a vector of function values $y = \left(u\left(Z^{(k)}\right)\right)_{k=1}^{K}$, we can calculate the coefficients of the gPC expansion by inverting the system

$$\sum_{\nu \in \Lambda} \hat{u}_\nu \phi_\nu\left(Z^{(k)}\right) = u\left(Z^{(k)}\right) \text{ for all } k \in [K].$$

Assembling the *sampling matrix*

$$A_{k,\nu} := \phi_\nu\left(Z^{(k)}\right)$$

and the vector of coefficients $\hat{u} = (\hat{u}_\nu)_{\nu \in \Lambda}$, we write the system in matrix form as

$$A\hat{u} = y.$$

Recall that in Section 1.7, we took the polynomial basis as the Lagrange interpolating polynomials for $\left\{Z^{(k)}\right\}_{k=1}^{M}$, so that $A = I$, the $M \times M$ identity matrix. However, since we now allow $K \neq M$, our system may be under- or over-determined. Since, in our application, obtaining the ensemble of function values is equivalent to obtaining the expensive deterministic finite element solution to a PDE, we hope to take K small. Thus, we will only be considering under-determined systems with $K \ll M$, making A a wide matrix.

In order to select a solution from the possibly infinitely many, we impose further constraints on the coefficient vector. Since in the context of the affine parametric PDE, solutions are analytic in Γ, we should expect fast decay of the gPC coefficients. Indeed, in the context of weighted cardinality and summability of the gPC coefficients (which we should view as a proxy for smoothness and will later be shown to be a consequence of analyticity), the weighted Stechkin estimate of Theorem 1.5 shows decay at the rate of $s^{1-1/p}$ for approximating $\hat{u}$ with coefficients supported on sets of weighted cardinality s. Thus, a natural assumption is to suppose that the coefficient vector $\hat{u}$ is *weighted* s*-sparse*, that is, $\|\hat{u}\|_{\omega,0} := \sum_{\nu \in \operatorname{supp} \hat{u}} \omega_\nu^2 \leqslant s \ll M$.

Before considering the weighted setup, we first review the traditional results for standard sparsity. These results are summarized in the following theorem.

Theorem 2.1 ([**15**] Corollary 12.34). *If* $W = \max_{\nu\in\Lambda} \|\varphi_\nu\|_{L^\infty(\Gamma)}$ *for* $|\Lambda| = M$ *and we draw*

$$K \gtrsim sW^2 \log^4(M)$$

i.i.d. samples $\{Z^{(k)}\}_{k=1}^K$ *from the orthogonalization measure* $\pi\, dz$*, then with probability exceeding* $1 - M^{-\log^3(M)}$*, we can approximately recover* u *from the polluted samples* $y = A\hat{u} + e$ *with error satisfying* $\|e\|_2 \leqslant \eta$ *as the solution* $\hat{u}^\sharp$ *of*

$$\underset{z\in\mathbb{C}^N}{\text{minimize}}\, \|z\|_1 \ \textit{subject to}\ \|Az - y\| \leqslant \eta,$$

in the sense that for $u^\sharp = \sum_{\nu\in\Lambda} \hat{u}^\sharp_\nu \varphi_\nu$*,*

$$\left\|u - u^\sharp\right\|_2 \leqslant \frac{A_1}{\sqrt{s}}\sigma_s(u)_1 + A_2\frac{\eta}{\sqrt{K}}$$

for $\sigma_s(\hat{u})_1 = \inf_{\|z\|_0 \leqslant s} \|\hat{u} - z\|_1$ *the* ℓ_1 *error in the best* s*-term approximation of* $\hat{u}$.

The main issue with this theorem is the reliance on the L^∞ norm of the gPC basis, as, even for nice bases, this produces exponential blow-up as the number of dimensions N increases requiring too many samples. In a following section, we will see examples of these issues and the benefit of introducing weighted ℓ_1 minimization and weighted sparsity.

2.2. Coefficient Weighting

We return to the setup defined in 1.5.3, where we consider weighted coefficient and function norms. For convenience, we recall Definitions 1.6 and 1.7, and Theorems 1.4 and 1.5 (where the latter is used to prove the former).

Definition 1.6. *For a sequence of weights* $\omega = (\omega_\nu)_{\nu\in\Lambda}$ *indexed over the same index set as the gPC basis, we define the* weighted ℓ_p space

$$\ell_{\omega,p} := \left\{ x = (x_\nu)_{\nu\in\Lambda} \mid \|x\|_{\omega,p} := \left(\sum_{\nu\in\Lambda} \omega_\nu^{2-p}|x_\nu|^p\right)^{1/p} < \infty \right\}, \quad 0 < p \leqslant 2,$$

with the weighted ℓ_0 *norm as*

$$\|x\|_{\omega,0} = \sum_{\nu\in\text{supp}(x)} \omega_\nu^2.$$

Additionally, we define associated function quasi-normed function space

$$S_{\omega,p} := \left\{ u(x) = \sum_{\nu\in\Lambda} \hat{u}_\nu\varphi_\nu(x) \mid \|u\|_{\omega,p} := \|\hat{u}\|_{\omega,p} < \infty \right\}, \quad 0 < p < 1.$$

Definition 1.7. *For a given basis of weighted cardinality* s*, we define the* error in the best weighted s-term approximation to a vector $x \in \ell_{\omega,p}$ *as*

$$\sigma_s(x)_{\omega,p} = \inf_{z:\|z\|_{\omega,0}\leqslant s} \|x - z\|_{\omega,p},$$

and associated error in the best weighted s-term approximation to a function $u \in S_{\omega,p}$ *as*

$$\sigma_s(u)_{\omega,p} = \sigma_s(\hat{u})_{\omega,p}.$$

We then take $u_{\Lambda^{\text{opt}}_{s,p}}$ *as the minimizer (if it exists), and*

$$\Lambda^{\text{opt}}_{s,p} = \text{supp}\, u_{\Lambda^{\text{opt}}_{s,p}}.$$

Theorem 1.4. *Suppose that the weight sequence* ω *satisfies* $\omega_\nu \geqslant \|\phi_\nu\|_\infty$ *on* $\mathbb{N}_0^N$ *and* $s \geqslant \|\omega\|_\infty^2$. *If* $u \in S_{\omega,p}$,

$$\left\|u - u_{\Lambda_{s,1}^{\mathrm{opt}}}\right\|_\infty \leqslant \left(s - \|\omega\|_\infty^2\right)^{1-1/p} \|u\|_{\omega,p}, \quad p < 1. \tag{2.1}$$

Theorem 1.5 ([24], Theorem 3.2). *For* $p < q \leqslant 2$, *let* $x \in \ell_{\omega,p}$. *Then for* $s \geqslant \|\omega\|_\infty^2$,

$$\sigma_s(x)_{\omega,q} \leqslant \tilde{\sigma}_s(x)_{\omega,q} \leqslant \left(s - \|\omega\|_\infty^2\right)^{1/q-1/p} \|x\|_{\omega,p}.$$

With the these alternatives to traditional ℓ_p norms, we are able to state a revised version of Theorem 2.1.

Theorem 2.2 ([24], Theorem 6.1). *For a finite index set* $|\Lambda| = M$ *and weights* $\omega_\nu \geqslant \|\phi_\nu\|_\infty$, *if* $s \geqslant 2\|\omega\|_\infty^2$ *and we draw*

$$K \gtrsim s \log^3(s) \log(M)$$

i.i.d. samples $\{Z^{(k)}\}_{k=1}^K$ *from the orthogonalization measure* $\pi\, dz$, *then with probability* $1 - M^{-\log^3(s)}$, *we can approximately recover* u *from the polluted samples* $y = A\hat{u} + e$ *with error satisfying* $\|e\|_2 \leqslant \eta$ *as the solution* $\hat{u}^\sharp$ *of*

$$\underset{z \in \mathbb{C}^M}{\text{minimize}}\, \|z\|_{\omega,1} \text{ subject to } \|Az - y\|_2 \leqslant \eta,$$

in the sense that for $u^\sharp = \sum_{\nu\in\Lambda} \hat{u}_\nu^\sharp \phi_\nu$,

$$\|u - u^\sharp\|_\infty \leqslant \|u - u^\sharp\|_{\omega,1} \leqslant B_1 \sigma_s(u)_{\omega,1} + B_2 \eta \sqrt{\frac{s}{K}}, \tag{2.2}$$

$$\|u - u^\sharp\|_2 \leqslant \frac{C_1}{\sqrt{s}} \sigma_s(u)_{\omega,1} + C_2 \frac{\eta}{\sqrt{K}}. \tag{2.3}$$

2.3. Weighted Null Space Property and Robust Sparse Recovery

In order to prove Theorem 2.2, we must make use of weighted versions of the traditional conditions which provide robust sparse recovery. We begin with the weighted robust null space property (NSP).

Definition 2.1 ([24], Defintion 4.1). *For a weight sequence* ω, *a matrix* $A \in \mathbb{C}^{K,M}$ *is said to satisfy the* weighted robust null space property of order s with constants $\rho \in (0, 1)$ and $\tau > 0$ *if*

$$\|v_S\|_2 \leqslant \frac{\rho}{\sqrt{s}} \|v_{S^C}\|_{\omega,1} + \tau \|Av\|_2 \text{ for all } v \in \mathbb{C}^M \text{ and all } S \subset [M] \text{ with } \omega(S) \leqslant s$$

where subscripting a vector with an index set (e.g., S*) denotes either restricting that vector to the lower dimensional subset corresponding to* S *(i.e.,* $v_S \in \mathbb{R}^{|S|}$ *with* $(v_S)_i = v_{S(i)}$*) or setting the entries at the indices chosen by the index set to be zero (i.e.,* $(v_S)_i = 0$ *for all* $i \in S$ *and* $(v_S)_i = v_i$ *for all* $i \notin S$*).*

The robust weighted null space property allows for the robust sparse recovery of vectors via weighted ℓ_1 minimization with estimates as in Theorem 2.2.

Theorem 2.3 ([24], Corollary 4.3). *Let* $A \in \mathbb{C}^{K,M}$ *satisfy the weighted robust null space property of order* s *with constants* $\rho \in (0, 1)$ *and* $\tau > 0$. *For* $x \in \mathbb{C}^M$ *and* $y = Ax + e$ *with* $\|e\|_2 \leqslant \eta$, *let* $x^\sharp$ *solve the weighted* ℓ_1 *minimization program*

$$\underset{z \in \mathbb{C}}{\text{minimize}}\, \|z\|_{\omega,1} \text{ subject to } \|Az - y\| \leqslant \eta.$$

Then the error in the recovered solution can be bounded in terms of the error in the weighted best s*-term estimate and measurement error as*

$$\|x - x^\sharp\|_{\omega,1} \leqslant B_1\sigma_s(x)_{\omega,1} + B_2\eta\sqrt{s},$$
$$\|x - x^\sharp\|_2 \leqslant \frac{C_1}{\sqrt{s}}\sigma_s(x)_{\omega,1} + C_2\eta,$$

where the latter bound holds under the assumption that $s \geqslant 2\|\omega\|_\infty^2$*, and all constants depend only on* ρ *and* τ*.*

Proof. In order to prove this theorem, we will use the weighted robust NSP on $x - x^\sharp$. In particular, we use the weighted robust NSP to prove a bound on the difference between x and any vector z in terms of the weighted best s term estimate, measurement error, and a quantity which acts nicely when $z = x^\sharp$ solves (2.3). Additionally, when $z = x^\sharp$ solves (2.3), we know that $Az - y = A(z - x)$ which has 2-norm bounded by η, and so our bound can more generally involve this quantity. The proper bounds end up being given by the following $\ell_{\omega,1}$ and ℓ_2 distance bounds.

Lemma 2.1 ([24], Theorem 4.1). *If* $A \in \mathbb{C}^{K,M}$ *satisfies the weighted robust null space property of order* s *with constants* $\rho \in (0,1)$ *and* $\tau > 0$*, then for all* $x, z \in \mathbb{C}^M$*, we have*

$$\|x - z\|_{\omega,1} \leqslant \frac{1+\rho}{1-\rho}\left(\|z\|_{\omega,1} - \|x\|_{\omega,1} + 2\sigma_s(x)_{\omega,1}\right) + \frac{2\tau\sqrt{s}}{1-\rho}\|A(x-z)\|_2, \tag{2.4}$$

and if $s \geqslant 2\|\omega\|_\infty^2$*,*

$$\|x - z\|_2 \leqslant \frac{C_1}{\sqrt{s}}\left(\|z\|_{\omega,1} - \|x\|_{\omega,1} + 2\sigma_s(x)_{\omega,1}\right) + C_2\|A(x-z)\|_2. \tag{2.5}$$

Proof of Lemma 2.1. We first claim that $y = \arg\min_{\|z\|_{\omega,0}\leqslant s}\|x - z\|_{\omega,1}$ is composed of entries of x on some index set S with weighted cardinality $\omega(S) \leqslant s$, and thus, the minimizer of this problem exists by minimizing over the finite set $\{x_S \mid \omega(S) \leqslant s\}$. Indeed, if $\operatorname{supp}(y) = S$,

$$\|x - x_S\|_{\omega,1} = \sum_{\nu\notin S}|x_\nu|\omega_\nu \leqslant \sum_{\nu\notin S}|x_\nu|\omega_\nu + \sum_{\nu\in S}|x_\nu - y_\nu|\omega_\nu = \|x - y\|_{\omega,1}.$$

Thus, we now take $y = x_S$ giving

$$\sigma_s(x)_{\omega,1} = \|x - x_S\|_{\omega,1} = \|x_{S^c}\|_{\omega,1}.$$

Now since we expect to obtain a bound involving $-\|x\|_{\omega,1}$, we keep the identity

$$0 = \|x_S\|_{\omega,1} + \|x_{S^c}\|_{\omega,1} - \|x\|_{\omega,1} = \|x_S\|_{\omega,1} + \sigma_s(x)_{\omega,1} - \|x\|_{\omega,1} \tag{2.6}$$

in our pocket to apply when necessary.

Since we wish to apply the weighted robust NSP, we begin by letting $x - z = v$ and consider

$$\|v\|_{\omega,1} = \|v_{S^c}\|_{\omega,1} + \|v_S\|_{\omega,1}.$$

If we can obtain a bound from the NSP involving the weighted ℓ_1 norm of v_S, we can apply it to the second term. Indeed, the Cauchy-Schwarz inequality followed by the weighted robust NSP gives

$$\|v_S\|_{\omega,1} = \sum_{\nu\in S}|v_\nu|\omega_\nu \leqslant \sqrt{\sum_{\nu\in S}|v_\nu|^2}\sqrt{\sum_{\nu\in S}\omega_\nu^2} = \sqrt{s}\|v_S\|_2 \leqslant \rho\|v_{S^c}\|_{\omega,1} + \tau\sqrt{s}\|Av\|_2.$$

Thus,

$$\|v\|_{\omega,1} \leqslant (1+\rho)\|v_{S^c}\|_{\omega,1} + \tau\sqrt{s}\|Av\|_2. \tag{2.7}$$

For the weighted ℓ_1 estimate, it just remains to bound $\|v_{S^c}\|_{\omega,1}$. We use our "pocket-identity" (2.6) and the weighted robust NSP again to obtain

$$\begin{aligned}\|(x-z)_{S^c}\|_{\omega,1} &\leqslant \|x_{S^c}\|_{\omega,1} + \|z_{S^c}\|_{\omega,1} + \sigma_s(x)_{\omega,1} + \|x_S\|_{\omega,1} - \|x\|_{\omega,1}\\ &= 2\sigma_s(x)_{\omega,1} + \|(x-z)_S + z_S\|_{\omega,1} + \|z_{S^c}\|_{\omega,1} - \|x\|_{\omega,1}\\ &\leqslant 2\sigma_s(x)_{\omega,1} + \|v_S\|_{\omega,1} + \|z_S\|_{\omega,1} + \|z_{S^c}\|_{\omega,1} - \|x\|_{\omega,1}\\ &\leqslant 2\sigma_s(x)_{\omega,1} + \rho\|v_{S^c}\|_{\omega,1} + \tau\sqrt{s}\|Av\|_2 + \|z\|_{\omega,1} - \|x\|_{\omega,1}.\end{aligned}$$

Moving $\rho\|v_{S^c}\|_{\omega,1}$ to the other side and dividing by $1-\rho$ gives a bound for $\|v_{S^c}\|_{\omega,1}$ which we plug into (2.7) to obtain

$$\begin{aligned}\|x-z\|_{\omega,1} &\leqslant \frac{1+\rho}{1-\rho}\left(2\sigma_s(x)_{\omega,1} + \|z\|_{\omega,1} - \|x\|_{\omega,1} + \tau\sqrt{s}\|Av\|_2\right) + \frac{1-\rho}{1-\rho}\tau\sqrt{s}\|Av\|_2\\ &= \frac{1+\rho}{1-\rho}\left(2\sigma_s(x)_{\omega,1} + \|z\|_{\omega,1} - \|x\|_{\omega,1}\right) + \frac{2\tau\sqrt{s}}{1-\rho}\|Av\|_2.\end{aligned}$$

Thus, (2.4) holds as desired.

We now prove (2.5) by way of the weighted Stechkin inequality in Theorem 1.5. Now letting S be the index set for the quasi-best s term estimate to v, that is $\tilde{\sigma}_s(v)_{\omega,2} = \|v_{S^c}\|_{\omega,2}$ with $\omega(S) \leqslant s$, we have

$$\|v\|_2 \leqslant \tilde{\sigma}_s(v)_{\omega,2} + \|v_S\|_2.$$

The former can be bounded by the weighted Stechkin inequality as

$$\tilde{\sigma}_s(v)_{\omega,2} \leqslant \frac{1}{\sqrt{s-\|\omega\|_\infty^2}}\|v\|_{\omega,1},$$

and the latter by the robust weighted NSP as

$$\begin{aligned}\|v_S\|_2 &\leqslant \frac{\rho}{\sqrt{s}}\|v_{S^c}\|_{\omega,1} + \tau\|Av\|_2\\ &\leqslant \frac{\rho}{\sqrt{s-\|\omega\|_\infty^2}}\|v\|_{\omega,1} + \tau\|Av\|_2.\end{aligned}$$

Combining, and using the ℓ_1 bound (2.4) derived above, we obtain

$$\begin{aligned}\|v\|_2 &\leqslant \frac{1+\rho}{\sqrt{s-\|\omega\|_\infty^2}}\|v\|_{\omega,1} + \tau\|Av\|_2\\ &\leqslant \frac{(1+\rho)^2}{(1-\rho)\sqrt{s-\|\omega\|_\infty^2}}\left(2\sigma_s(x)_{\omega,1} + \|z\|_{\omega,1} - \|x\|_{\omega,1}\right) + \left(\tau + \frac{2\tau(1+\rho)\sqrt{s}}{(1-\rho)\sqrt{s-\|\omega\|_\infty}}\right)\|Av\|_2.\end{aligned}$$

The last step is rewriting the denominator $\sqrt{s-\|\omega\|_\infty^2}$ in terms of $\sqrt{s}$. But since $s - 2\|\omega\|_\infty^2 \geqslant 0$, adding s to both sides, taking square roots and then reciprocals gives

$$\frac{1}{\sqrt{s}} \geqslant \frac{1}{\sqrt{2}\sqrt{s-\|\omega\|_\infty^2}}.$$

Thus, adding a factor of $\sqrt{2}$ to each fraction allows us to rewrite the denominator as $\sqrt{s}$ giving (2.5) with constants

$$C_1 = \frac{\sqrt{2}(1+\rho)^2}{1-\rho}, \qquad C_2 = \tau + \frac{2\sqrt{2}\tau(1+\rho)}{1-\rho}.$$

□

We now return to the proof of Theorem 2.3 using the bounds in Lemma 2.1. When $z = x^\sharp$ the solution of the minimization program (2.3), we know that since x is feasible for (2.3), $\|x^\sharp\|_{\omega,1} \leqslant \|x\|_{\omega,1}$. The differences of norms in (2.4) and (2.5) are then negative and we immediately obtain the desired reconstruction bounds by the fact that $A(x - x^\sharp) = y - Ax^\sharp$ which has 2-norm bounded by η.

□

2.4. Weighted Restricted Isometry Property

Instead of directly showing the weighted robust NSP for the sampling matrix, we appeal to the weighted restricted isometry property (RIP) which implies the weighted robust NSP. In the next section, we will then show that with high probability, the sampling matrix satisfies the weighted RIP. We begin with the definition of weighted RIP constants.

Definition 2.2 ([**24**], Definition 1.3). *For* $A \in \mathbb{C}^{K,M}$, $s \geqslant 1$, *and a weight sequence* ω, *the* ω-RIP constant $\delta_{\omega,s}$ *for* A *is the smallest number for which*

$$(1 - \delta_{\omega,s})\|x\|_2^2 \leqslant \|Ax\|_2^2 \leqslant (1 + \delta_{\omega,s})\|x\|_2^2$$

for all $x \in \mathbb{C}^M$ *with* $\|x\|_{\omega,0} \leqslant s$. *We say that* A *satisfies the* weighted restricted isometry property *(ω-RIP) if* $\delta_{\omega,s}$ *is sufficiently small (where the "sufficient" depends on the context).*

Proposition 2.1. *We can write the weighted restricted isometry constant as*

$$\delta_{\omega,s} = \max_{\omega(S)\leqslant s} \|A_S^* A_S - I\|, \tag{2.8}$$

where the norm is the operator 2-norm, and subscripting by S *represents restriction of* A *to the columns of the index set* S.

Proof. For any $\omega(S) \leqslant s$, we have $A_S^* A_S - I \in \mathbb{C}^{|S|,|S|}$ is Hermitian. By the spectral theorem, $A_S^* A_S - I = Q^* D Q$ is unitarily diagonalizable. Since Q is unitary $\|A_S^* A_S - I\| = \|D\|$, where for any $z \in \mathbb{S}^{|S|-1}$,

$$\|Dz\|_2^2 = \sum_{i=1}^{|S|} \lambda_i^2 z_i^2 \leqslant \lambda_{\max}^2,$$

and thus, $\|D\| = |\lambda_{\max}|$ (where equality is realized by taking z to be the canonical basis vector corresponding to the index of $\lambda_{\max}$). Additionally,

$$\sup_{z\in\mathbb{S}^{|S|-1}} |z^*(A_S^* A_S - I)z| = \sup_{z\in\mathbb{S}^{|S|-1}} |(Qz)^* D Q z| = \sup_{w\in\mathbb{S}^{|S|-1}} |w^* D w| = \sup_{w\in\mathbb{S}^{|S|-1}} \left|\sum_{i=1}^{|S|} \lambda_i w_i^2\right|.$$

By an analogous argument to the one showing $\|D\| = |\lambda_{\max}|$, this quantity must also equal $|\lambda_{\max}| = \|A_S^* A_S - I\|$.

By identifying the set of vectors $x \in \mathbb{C}^M$ with $\|x\|_{\omega,0} \leqslant s$ as the set of all vectors in $\mathbb{C}^{|S|}$ for some $\omega(S) \leqslant s$ by restricting to their support and rewriting (2.8) as

$$\frac{|x^*(A^*A - I)x|}{\|x\|_2^2} \leqslant \delta_{\omega,s},$$

$\delta_{\omega,s}$ can be written

$$\delta_{\omega,s} = \max_{\omega(S)\leqslant s} \sup_{x\in\mathbb{C}^{|S|}} \frac{|x^*(A_S^* A_S - I)x|}{\|x\|_2^2} = \max_{\omega(S)\leqslant s} \sup_{z\in\mathbb{S}^{|S|-1}} |z^*(A_S^* A_S - I)z| = \max_{\omega(S)\leqslant s} \|A_S^* A_S - I\|$$

where the last equality follows from the previous paragraph. □

Now, we prove that the ω-RIP implies the ω-NSP.

Theorem 2.4 ([**24**],Theorem 4.5). *Let $A \in \mathbb{C}^{K,M}$ have ω-RIP constant*

$$\delta_{\omega,3s} < 1/3$$

where we only consider $s \geqslant 2\|\omega\|_\infty^2$. Then A has the robust weighted null space property of order s with constants $\tau = \sqrt{1+\delta_{\omega,3s}}/(1-\delta_{\omega,3s})$ and $\rho = 2\delta_{\omega,3s}/(1-\delta_{\omega,3s})$.

Proof. The goal will be for any $\omega(S) \leqslant s$ to bound $\|v_S\|_2$ by a quantity involving the weighted ℓ_1 norm of v_{S^C}. It seems likely that in the process of using the ω-RIP to bound $\|v_S\|_2$, we will only encounter $\|v_{S^C}\|_2$. This ends up being the biggest sticking point and motivates the start of the bounding, so we will start by providing bounds for (almost) this quantity in terms of the weighted ℓ_1 norm.

Recall that in the proof of the weighted Stechkin estimate, Theorem 1.5, we were able to exchange from the ℓ_2 norm to the weighted ℓ_1 norm by paying a power of $(s - \|\omega\|_\infty^2)^{-1}$. The technique used here was to make sure that the ℓ_2 norm being bounded was of a vector whose entries were sorted to be smaller than those considered in the norm bounding it which allowed us to insert the weight sequence. Thus, we analogously start by constructing the non-increasing rearrangement of $|v_\nu|\omega_\nu^{-1}$, and then segmenting S^C reordered in this manner into blocks $S_1, S_2, \ldots$ of size to be determined. At the very least, we will want to apply ω-RIP or derive bounds involving s, so we will assume that $\omega(S_\ell) \leqslant s$.

Now restricting v to one of these index sets, we wish to bound

$$\|v_{S_\ell}\|_2^2 = \sum_{\nu \in S_\ell} v_\nu^2 = \sum_{\nu \in S_\ell} \left(|v_\nu|\omega_\nu^{-1}\right)^2 \omega_\nu^2 \leqslant s \left(\max_{\eta \in S_\ell} |v_\eta|\omega_\eta^{-1}\right)^2 .$$

Taking square roots, our hierarchical reordering of $|v_\nu|\omega_\nu^{-1}$ then comes in handy to bound last term. Again since our goal is to get to the weighted ℓ_1 norm, we multiply by ω_ν^2 on top and bottom and sum over the block with larger entries (which allows us to make use of weighted cardinality as well). To summarize, for $\ell \geqslant 2$,

$$\begin{aligned}\|v_{S_\ell}\|_2 &\leqslant \sqrt{s}\frac{1}{\omega(S_{\ell-1})} \sum_{\nu \in S_{\ell-1}} \omega_\nu^2 \max_{\eta \in S_\ell} |v_\eta|\omega_\eta^{-1} \\ &\leqslant \sqrt{s}\frac{1}{\omega(S_{\ell-1})} \sum_{\nu \in S_{\ell-1}} \omega_\nu^2 |v_\nu|\omega_\nu^{-1} \\ &= \sqrt{s}\frac{1}{\omega(S_{\ell-1})} \|v_{S_{\ell-1}}\|_{\omega,1}.\end{aligned}$$

A lower bound on $\omega(S_{\ell-1})$ will make this estimate ready to use in the remainder of the proof. In order to handle as much information in each block, we make sure that the weighted cardinality of $S_{\ell-1}$ is maximal, that is

$$s - \|\omega\|_\infty^2 \leqslant \omega(S_{\ell-1}) \leqslant s,$$

since if this lower bound did not hold, we could add the next weight in the non-increasing rearrangement and still maintain $\omega(S_{\ell-1}) \leqslant s$ which we have already assumed. Note that in general, this lower bound is only possible to enforce for every block but one. However, we only need it to hold for the last blocks (since it is impossible to step lower than the first block), and so we allow the first block to simply contain the remaining indices necessary without necessarily reaching weighted cardinality above $s - \|\omega\|_\infty^2$. With this choice and our assumption that $s \geqslant 2\|\omega\|_\infty^2$, we obtain

$$\|v_{S_\ell}\|_2 \leqslant \frac{\sqrt{s}}{s - \|\omega\|_\infty^2}\|v_{S_{\ell-1}}\|_{\omega,1} \leqslant \frac{2\sqrt{s}}{s + (s - 2\|\omega\|_\infty^2)}\|v_{S_{\ell-1}}\|_{\omega,1} \leqslant \frac{2}{\sqrt{s}}\|v_{S_{\ell-1}}\|_{\omega,1}. \tag{2.9}$$

Note that we've recovered the factor of $s^{-1/2}$ required by the ω-NSP as well!

Now that we have this bound, we may proceed with the start of bounding $\|v_S\|_2$. Since (2.9) is unavailable for $\ell = 1$, we prefer to work with

$$\|v_S + v_{S_1}\|_2 = \left\| v - \sum_{\ell \geqslant 2} v_{S_\ell} \right\|_2,$$

where the left hand side bounds $\|v_S\|_2$ as $S \cap S_1 = \emptyset$. We now make use of the ω-RIP to give

$$\begin{aligned}
\|v_S + v_{S_1}\|_2^2 &\leqslant \frac{1}{1-\delta_{\omega,2s}}\|A(v_S + v_{S_1})\|_2^2 \\
&= \frac{1}{1-\delta_{\omega,2s}}\langle A(v_S + v_{S_1}), A(v_S + v_{S_1})\rangle \\
&= \frac{1}{1-\delta_{\omega,2s}}\left\langle A(v_S + v_{S_1}), Av - \sum_{\ell \geqslant 2} Av_{S_\ell}\right\rangle \\
&\leqslant \frac{1}{1-\delta_{\omega,2s}}\left[|\langle A(v_S + v_{S_1}), Av\rangle| + \sum_{\ell \geqslant 2} |\langle A(v_S + v_{S_1}), Av_{S_\ell}\rangle| \right] \\
&\leqslant \frac{\sqrt{1+\delta_{\omega,2s}}}{1-\delta_{\omega,2s}}\|v_S + v_{S_1}\|_2\|Av\|_2 + \frac{1}{1-\delta_{\omega,2s}}\sum_{\ell \geqslant 2} |\langle A(v_S + v_{S_1}), Av_{S_\ell}\rangle|,
\end{aligned} \tag{2.10}$$

where the last line results from applying Cauchy-Schwarz followed by the ω-RIP.

We may be tempted to apply this same argument to the second term, however, the resulting constant cannot be bounded by one, a requirement for the ω-NSP. Instead, we apply a more nuanced argument taking advantage of the fact that S, S_1, and S_ℓ are mutually disjoint and therefore the restrictions of v on each of these sets are orthogonal. Indeed, this gives

$$\begin{aligned}
|\langle A(v_S + v_{S_1}), Av_{S_\ell}\rangle| &= \left|\left\langle A^*_{S\cup S_1\cup S_\ell}A_{S\cup S_1\cup S_\ell}(v_S + v_{S_1}), v_{S_\ell}\right\rangle\right| \\
&= \left|\left\langle A^*_{S\cup S_1\cup S_\ell}A_{S\cup S_1\cup S_\ell}(v_S + v_{S_1}), v_{S_\ell}\right\rangle + \langle v_S + v_{S_1}, v_{S_\ell}\rangle\right| \\
&= \left|\left\langle \left(A^*_{S\cup S_1\cup S_\ell}A_{S\cup S_1\cup S_\ell} - I\right)(v_S + v_{S_1}), v_{S_\ell}\right\rangle\right| \\
&\leqslant \delta_{\omega,3s}\|v_S + v_{S_1}\|_2\|v_{S_\ell}\|_2.
\end{aligned}$$

Combining with our ℓ_2 bound (2.9) we find,

$$|\langle A(v_S + v_{S_1}), Av_{S_\ell}\rangle| \leqslant \frac{2\delta_{\omega,3s}}{\sqrt{s}}\|v_S + v_{S_1}\|_2\|v_{S_{\ell-1}}\|_{\omega,1}.$$

Plugging into (2.10), dividing by $\|v_S + v_{S_1}\|_2$, and bounding all ω-RIP constants by the ones for largest weighted sparsity level $\delta_{\omega,3s}$, we have

$$\begin{aligned}
\|v_S\|_2 &\leqslant \|v_S + v_{S_1}\|_2 \\
&\leqslant \frac{\sqrt{1+\delta_{\omega,3s}}}{1-\delta_{\omega,3s}}\|Av\|_2 + \frac{2\delta_{\omega,3s}}{(1-\delta_{\omega,3s})\sqrt{s}}\sum_{\ell \geqslant 1}\|v_{S_\ell}\|_{\omega,1} \\
&\leqslant \frac{\sqrt{1+\delta_{\omega,3s}}}{1-\delta_{\omega,3s}}\|Av\|_2 + \frac{2\delta_{\omega,3s}}{(1-\delta_{\omega,3s})\sqrt{s}}\|v_{S^c}\|_{\omega,1}.
\end{aligned}$$

Thus, we arrive at the ω-NSP with constants $\tau = \sqrt{1+\delta_{\omega,3s}}/(1-\delta_{\omega,3s})$ and $\rho = 2\delta_{\omega,3s}/(1-\delta_{\omega,3s})$ where the latter is bounded by one by the assumption that $\delta_{s,3s} < 1/3$. □

2.5. Tools from High Dimensional Probability

In this section, we present the tools necessary for proving the ω-RIP property for the sampling matrix with high probability in Section 2.6. We begin with the characterization of sub-gaussian random variables following the treatment in [29].

Proposition 2.2 ([29], Proposition 2.5.2). *The following are equivalent:*

(1) The tails of X *satisfy*

$$\mathbb{P}(|X| \geqslant t) \leqslant 2\exp\left(-t^2/K^2\right) \quad \textit{for all } t \geqslant 0.$$

(2) The moments of X *satisfy*

$$(\mathbb{E}|X|^p)^{1/p} \leqslant K\sqrt{p} \quad \textit{for all } p \geqslant 1.$$

(3) The MGF of X^2 *satisfies*

$$\mathbb{E}\exp\left(\lambda^2 X^2\right) \leqslant \exp\left(K^2\lambda^2\right) \quad \textit{for all } \lambda \textit{ such that } |\lambda| \leqslant \frac{1}{K}.$$

(4) The MGF of X^2 *is bounded at some point, namely*

$$\mathbb{E}\exp\left(X^2/K^2\right) \leqslant 2$$

(5) Additionally, if $\mathbb{E}X = 0$*, the MGF of* X *satisfies*

$$\mathbb{E}\exp(\lambda X) \leqslant \exp\left(K^2\lambda^2\right) \quad \textit{for all } \lambda \in \mathbb{R}.$$

The constants in each property are within absolute multiplicative constants of one another.

Proof. We always start by setting $X \to X/K$ noting that each property is multiplicatively homogeneous.

(1) $\implies$ (2): Using the layer cake representation of $|X|^p$, we find

$$\begin{aligned}
\mathbb{E}|X|^p &= \int_0^\infty \mathbb{P}(|X|^p \geqslant s)\, ds \\
&= \int_0^\infty pt^{p-1}\mathbb{P}(|X| \geqslant t)\, dt \\
&\leqslant 2\int_0^\infty pt^{p-1}\exp\left(-t^2\right)\, dt \\
&= p\int_0^\infty 2t(t^2)^{p/2-1}\exp\left(-t^2\right)\, dt \\
&= p\int_0^\infty u^{p/2-1}\exp(-u)\, du \\
&= p\Gamma(p/2).
\end{aligned}$$

We make use of Stirling's formula

$$\Gamma(x) = \sqrt{2\pi}x^{x-1/2}e^{-x}\exp\left(\frac{\theta(x)}{12x}\right) \text{ for all } x > 0, \tag{2.11}$$

where $0 \leqslant \theta(x) \leqslant 1$. We then obtain the upper bound

$$\Gamma(x) \leqslant \sqrt{2\pi}x^{x-1/2}\exp\left(-x+\frac{1}{12x}\right).$$

Dividing by x^x and taking logarithms, we obtain

$$\log\left(\frac{\Gamma(x)}{x^x}\right) = \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(x) - x + \frac{1}{12x}$$

which is clearly negative at 2π and has derivative

$$-\frac{1}{2x} - 1 - \frac{1}{12x^2} \leqslant 0.$$

Thus, for any $x \geqslant 2\pi$, we know that $\Gamma(x) \leqslant x^x$. On the compact interval $[1/2, 2\pi]$ $\Gamma(x)$ is uniformly bounded from above, and $x^x > 0$ from below. Then for all $x \geqslant 1/2$, we have that $\Gamma(x) \leqslant C^p x^x$ for some absolute constant $C \geqslant 1$. Applying this to our bound for the moment of X, for all $p \geqslant 1$,

$$\begin{aligned}\mathbb{E}|X|^p &\leqslant C^p p \left(\frac{p}{2}\right)^{p/2} \\ &\leqslant C^p p p^{p/2}.\end{aligned}$$

Taking $1/p$ powers and using that $p^{1/p}$ has a global maximum at e, we have

$$\|X\|_{L^p} \leqslant C e^{1/e}\sqrt{p}$$

as desired.

(2) $\implies$ (3): Considering

$$\mathbb{E}\exp(\lambda^2 X^2) = 1 + \sum_{p=1}^{\infty} \frac{\lambda^{2p}\mathbb{E}X^{2p}}{p!} \leqslant 1 + \sum_{p=1}^{\infty} \frac{\lambda^{2p}(2p)^p}{p!},$$

and using Stirling's approximation to give $p! = \Gamma(p+1) \geqslant (Cp)^p$ for some $C > 0$ an absolute constant, we have

$$\mathbb{E}\exp(\lambda^2 X^2) \leqslant 1 + \sum_{p=1}^{\infty} (C\lambda^2)^p = \frac{1}{1 - C\lambda^2} \text{ for all } C\lambda^2 \leqslant 1.$$

Now since since $1 - C\lambda^2 \geqslant e^{-2C\lambda^2}$ for $C\lambda^2 < c \leqslant 1/2$ small enough, we know

$$\mathbb{E}\exp(\lambda^2 X^2) \leqslant \exp(2C\lambda^2),$$

for all $C\lambda^2 \leqslant c \equiv |\lambda| \leqslant \sqrt{2c}\frac{1}{\sqrt{2C}} \leqslant \frac{1}{\sqrt{2C}}$ as desired.

(3) $\implies$ (4): Take $\lambda = \log(2)/K$.

(4) $\implies$ (1): Taking squares and exponentials and applying Markov's inequality,

$$\begin{aligned}\mathbb{P}[|X| \geqslant t] &= \mathbb{P}[\exp(X^2) \geqslant \exp(t^2)] \\ &\leqslant \exp(-t^2)\mathbb{E}\exp(X^2) \\ &\leqslant 2\exp(-t^2)\end{aligned}$$

as desired.

(3) $\implies$ (5): We begin by proving the magic inequality

$$e^x \leqslant x + e^{x^2} \quad \text{for all } x \in \mathbb{R}.$$

We do this by showing $f(x) = x + e^{x^2} - e^x$ is strictly convex with global minimum $f(0) = 0$. Taking second derivatives $f''(x) = 2e^{x^2}(1 + 2x^2) - e^x$. Dividing by e^x, we then wish to show

$$1 < 2e^{x^2 - x}(1 + 2x^2).$$

This follows from the fact that $1 + 2x^2 > 1$ and the minimum value of $x^2 - x$ is $-1/4$. Thus,

$$\begin{aligned} 1 &< \left(\frac{16}{e}\right)^{1/4} \\ &= 2e^{-1/4} \\ &\leqslant 2e^{x^2-x} \\ &< 2e^{x^2-x}(1+2x^2). \end{aligned}$$

Multiplying by e^x then gives

$$0 < 2e^{x^2}(1+2x^2) - e^x = f''(x)$$

as desired. Since $f'(x) = 1 + 2xe^{x^2} - e^x = 0$ when $x = 0$, we know that $x = 0$ is the lone critical point of f and is indeed the global minimizer by strict convexity. Thus, the inequality holds.

Now, this inequality allows us to compare the moment generating function of X and X^2, and Property (3),

$$\mathbb{E}\exp(\lambda X) \leqslant \mathbb{E}\lambda X + \mathbb{E}\exp\left(\lambda^2 X^2\right) = \mathbb{E}\exp\left(\lambda^2 X^2\right) \leqslant \exp\left(\lambda^2\right)$$

when $|\lambda| \leqslant 1$, since $\mathbb{E}X = 0$. Now, we just consider the case when $|\lambda| > 1$. By Cauchy's inequality,

$$\mathbb{E}\exp(\lambda X) \leqslant \exp\left(\lambda^2/2\right)\mathbb{E}\exp\left(X^2/2\right) \leqslant \exp\left(\lambda^2/2\right)\exp(1/2) \leqslant \exp\left(\lambda^2/2 + \lambda^2/2\right) = \exp\left(\lambda^2\right).$$

(5) $\implies$ (1): We end with a similar argument to (4) $\implies$ (1), but must be a bit more careful. Multiplying by $\lambda > 0$ and exponentiating, we have

$$\begin{aligned} \mathbb{P}[X \geqslant t] &= \mathbb{P}[\exp(\lambda X) \geqslant \exp(\lambda t)] \\ &\leqslant \exp(-\lambda t)\mathbb{E}\exp(\lambda X) \\ &\leqslant \exp\left(-\lambda t + \lambda^2\right). \end{aligned}$$

Optimizing for λ, we have a minimizer at $\lambda = t/2$, giving

$$\mathbb{P}[X \geqslant t] \leqslant \exp\left(-t^2/4\right).$$

Repeating for $-X$ and using that (5) holds for all $\lambda \in \mathbb{R}$, we know that $\mathbb{P}[-X \geqslant t] \leqslant \exp\left(-t^2/4\right)$ as well. Property (1) follows by the union bound. □

Definition 2.3 ([29], Definition 2.5.6). *A random variable satisfying any of Properties (1)–(5) is called* sub-gaussian. *We define the* sub-gaussian norm of X, $\|X\|_{\psi_2}$, *to be the smallest* K *for which Property (1) holds. By the equivalence of (1)–(5), we can replace* K *in each property by* $C\|X\|_{\psi_2}$, *where* $C > 0$ *is an absolute constant depending on the property. Thus, up to this absolute constant* $\|X\|_{\psi_2}$ *is the smallest constant for which each of the properties holds.*

Example 2.1. *A Rademacher random variable* ε *is a mean zero, unit variance, symmetric Bernoulli random variable with distribution*

$$\varepsilon = \begin{cases} 1, & \textit{with probability } \frac{1}{2} \\ -1, & \textit{with probability } \frac{1}{2}. \end{cases}$$

Since $\mathbb{P}(|\varepsilon| \geqslant t) = 0 \leqslant 2\exp\left(-t^2/K^2\right)$ *for all* $t > 1$, *and since* $\mathbb{P}(|\varepsilon| \geqslant t) = 1$ *for all* $t \leqslant 1$, *to determine* $\|\varepsilon\|_{\psi_2}$, *it suffices to consider* $t = 1$ *which is where the exponential in Property (1)* *is smallest. Solving for* $\|\varepsilon\|_{\psi_2}$ *in*

$$\mathbb{P}(|\varepsilon| \geqslant 1) = 1 = 2\exp\left(-\frac{1}{\|\varepsilon\|_{\psi_2}^2}\right),$$

we obtain $\|\varepsilon\|_{\psi_2} = 1/\sqrt{\log(2)}$.

More generally, this argument holds for any bounded random variable X *with* $\|X\|_\infty = L$. *The sub-gaussian norm is entirely determined by* $t = L$, *and rearranging*

$$\mathbb{P}(|X| \geqslant L) = 1 = 2\exp\left(-\frac{L^2}{\|X\|_{\psi_2}^2}\right)$$

gives

$$\|X\|_{\psi_2} = \frac{L}{\sqrt{\log(2)}}.$$

A very important property of sub-gaussian random variables is that sums of independent mean-zero sub-gaussian random variables are still sub-gaussian. This is characterized in the following proposition.

Proposition 2.3. *Suppose* $X_1, \ldots, X_N$ *are independent, mean-zero, sub-gaussian random variables. Then* $\sum_{i=1}^N X_i$ *is also sub-gaussian with*

$$\left\|\sum_{i=1}^N X_i\right\|_{\psi_2}^2 \leqslant C\sum_{i=1}^N \|X_i\|_{\psi_2}^2$$

for C *an absolute constant.*

Proof. Using Property (5) in the definition of sub-gaussian and independence,

$$\begin{aligned}\mathbb{E}\exp\left(\lambda\sum_{i=1}^N X_i\right) &= \prod_{i=1}^N \mathbb{E}\exp(\lambda X_i)\\ &\leqslant \prod_{i=1}^N \exp\left(C\|X_i\|_{\psi_2}^2\lambda^2\right)\\ &= \exp\left(\lambda^2 C\sum_{i=1}^N \|X_i\|_{\psi_2}^2\right).\end{aligned}$$

Again by Property (5), the sum is sub-gaussian with

$$\left\|\sum_{i=1}^N X_i\right\|_{\psi_2}^2 \leqslant C\sum_{i=1}^N \|X_i\|_{\psi_2}^2$$

as desired. □

The equivalent properties of sub-gaussian random variables along with the previous proposition produce some nice inequalities of sums of random variables which we will make use of later.

Theorem 2.5 (Kintchine's Inequality, [29], Exercise 2.6.5). *For* $X_1, \ldots, X_N$ *independent sub-gaussian random variables with zero mean, let* $a = (a_i)_{i=1}^N \in \mathbb{R}^N$. *Then for any* $p \in [2, \infty)$,

$$\left\|\sum_{i=1}^N X_i a_i\right\|_{L^p} \leqslant CK\sqrt{p}\|a\|_2$$

where $K = \max_{i \in [N]} \|X_i\|_{\psi_2}$*, and* C *is an absolute constant.*

Proof. By the moment condition on the sum (which is sub-gaussian by the previous proposition),

$$\begin{aligned} \left\| \sum_{i=1}^{N} a_i X_i \right\|_{L^p} &\leqslant C\sqrt{p} \left\| \sum_{i=1}^{N} a_i X_i \right\|_{\psi_2} \\ &\leqslant C\sqrt{p} \sqrt{\sum_{i=1}^{N} a_i^2 \|X_i\|_{\psi_2}^2} \\ &\leqslant C\sqrt{p} \max_{i \in [N]} \|X_i\|_{\psi_2} \|a\|_2, \end{aligned}$$

as desired. □

We now consider sub-gaussian processes which will become important when bounding the ω-RIP constant of the sampling matrix.

Definition 2.4. *A stochastic process* $(X_t)_{t \in T}$ *on a metric space* (T, d) *is* sub-gaussian *if there exists a uniform constant* $K > 0$ *such that the increments satisfy*

$$\|X_s - X_t\|_{\psi_2} \leqslant K d(s, t).$$

It turns out that we can bound the suprema of such processes by quantities involving covering numbers of the index set T.

Definition 2.5. *Let* $(\mathcal{X}, d)$ *be a metric space. Consider a subset* $K \subset \mathcal{X}$ *and let* $t > 0$*. A subset* $\mathcal{N} \subseteq K$ *is called an* t-net of K *if every point in* K *is within a distance* t *of some point of* $\mathcal{N}$*, that is,*

$$\textit{for all } x \in K, \ \textit{there exists some } x_0 \in \mathcal{N} \textit{ such that } d(x, x_0) \leqslant t.$$

The smallest possible cardinality of a t*-net of* K *is called the* covering number *of* K *and is denoted* $\mathcal{N}(K, d, t)$*.*

Theorem 2.6 (Dudley's Inequality, [29], Thoerem 8.1.3). *A sub-gaussian process* $(X_t)_{t \in T}$ *satisfies*

$$\mathbb{E} \sup_{t \in T} |X_t - X_{t_0}| \leqslant CK \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \epsilon)}\, d\epsilon,$$

where K *is the constant with respect to which* X_t *is sub-gaussian. For measurability purposes, we consider the* lattice supremum *of the stochastic process,*

$$\mathbb{E} \sup_{t \in T} |X_t - X_{t_0}| = \sup_{\substack{S \subseteq T \\ S \textit{ finite}}} \mathbb{E} \sup_{t \in S} |X_t - X_{t_0}|.$$

In order to prove Dudley's inequality we need to make use of the following proposition on the maximum of a collection of sub-gaussian random variables.

Proposition 2.4 ([29], Exercise 2.5.10). *For every* $N \geqslant 2$ *and* $X_1, X_2, \ldots, X_N$ *sub-gaussian,*

$$\mathbb{E} \max_{i \in [N]} |X_i| \leqslant CK\sqrt{\log N}$$

where $K = \max_{i \in [N]} \|X_i\|_{\psi_2}$.

Proof. We first consider the random variable

$$Z := \max_{i\in[N]} Z_i := \max_{i\in[N]} \frac{|X_i|}{K\sqrt{1+\log i}},$$

where in particular, the Z_i have sub-gaussian norm bounded by $1/\sqrt{1+\log i}$. Then applying the union bound,

$$\begin{aligned}\mathbb{P}[|Z| \geqslant t] &\leqslant \sum_{i=1}^{N} \mathbb{P}[|Z_i| \geqslant t] \\ &\leqslant 2\sum_{i=1}^{N} \exp\big(-t^2(1+\log i)\big) \\ &= 2\exp\big(-t^2\big)\sum_{i=1}^{N} i^{-t^2}.\end{aligned}$$

For $t^2 \geqslant 2$, the sum is bounded by $\sum_{i=1}^{\infty} i^{-2} = \pi^2/6$. For $t^2 \leqslant 2$,

$$\mathbb{P}[|Z| \geqslant t] \leqslant 1 = e^2\exp(-2) \leqslant 2\frac{e^2}{2}\exp\big(-t^2\big).$$

Thus, for all $t \geqslant 0$, taking $C' = \max\{\pi^2/6, e^2/2\} = e^2/2$,

$$\mathbb{P}[|Z| \geqslant t] \leqslant 2C'\exp\big(-t^2\big)$$

Running the argument through for Property (1) $\implies$ Property (2) in the definition of a sub-gaussian random variable, we find $\mathbb{E}|Z| \leqslant C$. Since $|X_i|/\sqrt{1+\log i} \geqslant |X_i|/\sqrt{1+\log N}$, we obtain

$$\mathbb{E}\max_{i\in[N]} |X_i| \leqslant CK\sqrt{1+\log N} \leqslant \sqrt{2}CK\sqrt{\log N}$$

since $N \geqslant 2$. □

Proof of Theorem 2.6. The idea of the proof is to replace the expectation of the supremum over all of T into one over just ϵ-nets of T which can be bounded with a union bound. This was actually already done in the context of expectations of supremeums of sub-gaussian random variables in Proposition 2.4. So we will just want to handle the expectation of the supremum of increments over ϵ-nets since we know X_t is a sub-gaussian process.

We start with a summed discrete version of the upper bound which we then realize as an integral which may be easier to calculate in practice. Since we are working with the lattice supremum, if we can prove a uniform upper bound over supremums over all finite subsets of T, the bound holds for the lattice supremum as well. Thus, without loss of generality, we assume T is finite.

We will want to relate the expectation of the supremum over T to the expectation over just ϵ-nets of T. To this end, we construct a sequence of nets of increasing fineness. Under the assumption that T is finite, if we make our net fine enough then, it eventually must then contain every point in T. And if we start our sequence of nets at a large enough scale, every point in T will be be able to be contained in the ball, so our first net can contain the single point t_0 chosen in the statement of the theorem.

It turns out that in order to relate the resulting sum to an integral, we should choose the scale of our nets to be dyadic, that is, define

$$\epsilon_k = 2^{-k} \text{ for some } k \in \mathbb{Z},$$

and let a corresponding net achieving the covering number at this scale to be T_k (that is, $|T_k| = \mathcal{N}(T, d, \epsilon_k)$). As reasoned above, for some κ small enough, we may choose $T_\kappa = \{t_0\}$, and for some K large enough, we may choose $T_K = T$.

Now, we will relate an arbitrary X_t to these nets, by defining $\phi_k(t)$ to be the closest point to t in T_k. By definition of an ϵ_k-net, $d(t, \phi_k(t)) \leqslant \epsilon_k$. Now, we walk from t_0 to t along these finer and finer nets, giving the telescopic representation

$$X_t - X_{t_0} = \sum_{k=\kappa+1}^{K} X_{\phi_k(t)} - X_{\phi_{k-1}(t)},$$

since we know $X_{\phi_K(t)} = X_t$ and $X_{\phi_\kappa(t)} = t_0$. Thus, we may rewrite the supremum (which is a maximum over finite T) we wish to bound as

$$\begin{aligned}
\mathbb{E}\max_{t\in T}|X_t - X_{t_0}| &\leqslant \mathbb{E}\max_{t\in T}\sum_{k=\kappa+1}^{K}\left|X_{\phi_k(t)} - X_{\phi_{k-1}(t)}\right| \\
&\leqslant \sum_{k=\kappa+1}^{K}\mathbb{E}\max_{t\in T}\left|X_{\phi_k(t)} - X_{\phi_{k-1}(t)}\right| \\
&\leqslant \sum_{k=\kappa+1}^{K}\mathbb{E}\max_{t_1\in T_k, t_2\in T_{k-1}}|X_{t_1} - X_{t_2}|.
\end{aligned}$$

Now, since the increments are sub-gaussian, Proposition 2.4 guarantees that

$$\begin{aligned}
\mathbb{E}\max_{t_1\in T_k, t_2\in T_{k-1}}|X_{t_1} - X_{t_2}| &\leqslant C\max_{t_1\in T_k, t_2\in T_{k-1}}\|X_{t_1} - X_{t_2}\|_{\psi_2}\sqrt{\log(|T_k||T_{k-1}|)} \\
&\leqslant CK\max_{t_1\in T_k, t_2\in T_{k-1}} d(t_1, t_2)\sqrt{\log|T_k|^2} \\
&\leqslant CK\max_{t_1\in T_k, t_2\in T_{k-1}}[d(t_1, t) + d(t, t_2)]\sqrt{\log|T_k|} \\
&\leqslant CK[\epsilon_k + \epsilon_{k-1}]\sqrt{\log|T_k|} \\
&\leqslant CK\epsilon_{k-1}\sqrt{\log|T_k|}.
\end{aligned}$$

Inserting into the previous sum,

$$\mathbb{E}\sup_{t\in T}|X_t - X_{t_0}| \leqslant CK\sum_{k=\kappa+1}^{K}\epsilon_{k-1}\sqrt{\log|T_k|} = 2CK\sum_{k=\kappa+1}^{K}2^{-k}\sqrt{\log\mathcal{N}(T, d, 2^{-k})}.$$

Finally, we make use of our dyadic scale to say that

$$2^{-k} = 2\int_{2^{-(k+1)}}^{2^{-k}} d\epsilon.$$

And since for any $\epsilon \in (2^{-(k+1)}, 2^{-k})$, $\epsilon \leqslant 2^{-k}$,

$$\mathcal{N}(T, d, 2^{-k}) \leqslant \mathcal{N}(T, d, \epsilon).$$

Thus, the sum can be bounded as

$$\mathbb{E}\sup_{t\in T}|X_t - X_{t_0}| \leqslant CK\sum_{k\in\mathbb{Z}}\int_{2^{-(k+1)}}^{2^{-k}}\sqrt{\log\mathcal{N}(T, d, \epsilon)}\, d\epsilon = CK\int_0^\infty\sqrt{\log\mathcal{N}(T, d, \epsilon)}\, d\epsilon,$$

as desired. □

We now discuss symmetrization, a technique which allows us to convert a sum of random variables into a sum weighted by independent Rademacher variables which we can consider separately. For example, we can take the conditional expectation with respect to the original random variables, and view the symmetrized sum as a weighted Rademacher sum to which we can apply nice theorems (e.g., Bernstein or Kintchine inequalities).

Lemma 2.2 (Symmetrization, [29], Lemma 6.4.2). *Let* $X_1, \dots, X_N$ *be independent random vectors in a normed space, and* $\varepsilon_1, \dots \varepsilon_N$ *an independent Rademacher sequence also independent of the* X_i. *Then*

$$\mathbb{E}\left\|\sum_{i=1}^{N}(X_i - \mathbb{E}X_i)\right\| \leqslant 2\mathbb{E}\left\|\sum_{i=1}^{N}\varepsilon_i X_i\right\|.$$

Proof. Let X_i' be an independent copy of X_i and for any random element ξ in an expression containing only jointly independent random elements, denote expectation conditional on all random elements except ξ as $\mathbb{E}_\xi$. Then since $\mathbb{E}X_i = \mathbb{E}X_i'$,

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^{N}(X_i - \mathbb{E}X_i)\right\| &= \mathbb{E}\left\|\sum_{i=1}^{N}(X_i - \mathbb{E}_{X_i'}X_i')\right\| \\ &\leqslant \mathbb{E}\mathbb{E}_{X'}\left\|\sum_{i=1}^{N}(X_i - X_i')\right\| \\ &= \mathbb{E}\left\|\sum_{i=1}^{N}(X_i - X_i')\right\|.\end{aligned}$$

Since X_i and X_i' are i.i.d., for any $B \in \mathcal{B}(\mathbb{R}^n)$ Borel measurable on $\mathbb{R}^n$

$$\begin{aligned}\mathbb{P}(X - X' \in B) &= \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} 1_{\{x_1 - x_2 \in B\}}(x_1, x_2)\ dP_X(x_1)\ dP_{X'}(x_2) \\ &= \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} 1_{\{x_1 - x_2 \in B\}}(x_1, x_2)\ dP_{X'}(x_1)\ dP_X(x_2) = \mathbb{P}(X' - X \in B).\end{aligned}$$

Thus $X - X'$ is symmetric. Additionally, for any symmetric random vector ξ, we have that $\varepsilon\xi$ and ξ are equal in distribution as

$$\begin{aligned}\mathbb{P}[\varepsilon\xi \in B] &= \mathbb{P}\left[((\varepsilon = -1) \cap (-\xi \in B)) \cup ((\varepsilon = 1) \cap (\xi \in B))\right] \\ &= \mathbb{P}\left[\varepsilon = -1\right]\mathbb{P}\left[-\xi \in B\right] + \mathbb{P}\left[\varepsilon = 1\right]\mathbb{P}\left[\xi \in B\right] \\ &= \mathbb{P}\left[\xi \in B\right](\mathbb{P}[\varepsilon = -1] + \mathbb{P}[\varepsilon = 1]) \\ &= \mathbb{P}\left[\xi \in B\right].\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^{N}(X_i - X_i')\right\| &= \mathbb{E}\left\|\sum_{i=1}^{N}\varepsilon_i(X_i - X_i')\right\| \\ &\leqslant \mathbb{E}\left\|\sum_{i=1}^{N}\varepsilon_i X_i\right\| + \mathbb{E}\left\|\sum_{i=1}^{N}\varepsilon_i X_i'\right\| \\ &= 2\mathbb{E}\left\|\sum_{i=1}^{N}\varepsilon_i X_i\right\|\end{aligned}$$

since X_i and X_i' are also equal in distribution. □

As an example of symmetrization, we prove a version of *Maurey's lemma* on bounding covering numbers of the convex hull of a set of points. We will later use this lemma to bound the Dudley integral which in turn bounds the supremum of a sub-gaussian process.

Lemma 2.3 (Maurey's Lemma, [17], Lemma 4.2). *Let $\mathcal{X}$ be a normed space, and consider a finite set $\mathcal{U} \subset \mathcal{X}$ of N points. Assume that for every $L \in \mathbb{N}$ and $(u_1, \dots, u_L) \in \mathcal{U}^L$, $\mathbb{E}_\varepsilon \left\| \sum_{i=1}^L \varepsilon_i u_i \right\| \leqslant A\sqrt{L}$ for a Rademacher sequence ε as above. Then for every $t > 0$*

$$\log \mathcal{N}(\operatorname{conv}(\mathcal{U}), \|\cdot\|, t) \leqslant C(A/t)^2 \log N,$$

where $C > 0$ is an absolute constant.

Proof. Take $x \in \operatorname{conv} \mathcal{U}$ with $x = \sum_{j=1}^N u_j \theta_j$ with $\sum_{j=1}^N \theta_j$. We use this convex combination as a probability distribution, and we define a random element X to take the value u_j with probability θ_j. Thus $\mathbb{E}X = x$. Since we have a nice bound on the Rademacher sum of entries in $\mathcal{U}^L$, we wish to use symmetrization on a sum of length L to be determined of centered random elements taking values in $\mathcal{U}$. Since X takes values in $\mathcal{U}$, and we know its expectation, we consider

$$\mathbb{E}\left\| \sum_{i=1}^L (X_i - x) \right\| \leqslant 2\mathbb{E}\mathbb{E}_\varepsilon \left\| \sum_{i=1}^L \varepsilon_i X_i \right\| \leqslant 2A\sqrt{L},$$

where we condition on X_i using that the assumed bound is uniform over $\mathcal{U}^L$. Since we wish to bound the distance of x to some t-net, we normalize this equation to produce

$$\mathbb{E}\left\| x - \frac{1}{L}\sum_{i=1}^L X_i \right\| = \frac{1}{L}\mathbb{E}\left\| \sum_{i=1}^L (X_i - x) \right\| \leqslant \frac{2A}{\sqrt{L}},$$

and denote the random point $\frac{1}{L}\sum_{i=1}^L X_i = X_0$. If we choose $L \sim (2A/t)^2$, we then have that in expectation $\|x - X_0\|$ is bounded by t, and therefore there must exist some realization of X_0, x_0 that also satisfies this bound. We also know that

$$x_0 \in \left\{ \frac{1}{L}\sum_{i=1}^L u_i \mid u_i \in \mathcal{U} \text{ for all } i \in [L] \right\}.$$

Since x was arbitrary, this set is then a t-net for $\operatorname{conv} \mathcal{U}$ and its cardinality $N^L \leqslant N^{C(A/t)^2}$ bounds the covering number of $\operatorname{conv} \mathcal{U}$. Taking logarithms gives the desired bound. □

We end with a very powerful theorem giving a Bernstein style one-sided concentration inequality on the supremum of an empirical process, that is, the supremum of a stochastic process indexed by a class of functions. We follow the lead of [22] and do not consider its proof as it is very lengthy and outside the scope of these notes which attempt to give only an overview of basic techniques used in proving results from high dimensional probability. Interested readers can see [15, Section 8.9] for the proof and many corollaries.

Theorem 2.7 (Bernstein's Inequality for Suprema of Empirical Processes). *Let $\mathcal{F}$ be a countable set of functions $F : \mathbb{C}^M \to \mathbb{R}$. Let $X_1, \dots X_K$ be independent random vectors in $\mathbb{C}^M$ such that $\mathbb{E}F(X_k) = 0$, and $F(X_k) \leqslant L$ almost surely for all $k \in [K]$ and for all $F \in \mathcal{F}$ for some constant $L > 0$. For the supremum of an empirical process*

$$Z = \sup_{F \in \mathcal{F}} \sum_{k=1}^K F(X_k),$$

which satisfies the bound on the variance of the summed terms $\mathbb{E}[F(X_k)^2] \leqslant \sigma_k^2$ *for all* $F \in \mathcal{F}$ *and* $k \in [K]$*, we have for any* $t > 0$*,*

$$\mathbb{P}(Z \geqslant \mathbb{E}Z + t) \leqslant \exp\left(-\frac{t^2/2}{\sigma^2 + 2L\mathbb{E}Z + tL/3}\right),$$

where $\sigma^2 = \sum_{k=1}^K \sigma_k^2$ *is the variance of the sum.*

As a corollary, we have the traditional Bernstein's inequality (which we remark may be proved my much simpler means, similar to a proof of Hoeffding's inequality for example).

Corollary 2.1 (Bernstein's Inequality, [**29**], Theorem 2.8.4). *Let* $X_1, \dots, X_K$ *be independent, mean-zero random variables such that* $|X_k| \leqslant L$ *for all* $k \in [K]$*. Then for every* $t \geqslant 0$*,*

$$\mathbb{P}\left[\left|\sum_{k=1}^K X_k\right| \geqslant t\right] \leqslant 2\exp\left(-\frac{t^2/2}{\sigma^2 + tL/3}\right)$$

where $\sigma^2 = \sum_{k=1}^K \mathbb{E}X_k^2$ *is the variance of the sum.*

Proof. In Theorem 2.7, take $\mathcal{F} = \{I\}$, just the identity. Then $Z = \sum_{k=1}^K X_k$ is also mean-zero, giving

$$\mathbb{P}\left(\sum_{k=1}^K X_k \geqslant t\right) \leqslant \exp\left(-\frac{t^2/2}{\sigma^2 + tL/3}\right).$$

A second application with $\mathcal{F} = \{-I\}$ gives the result for the negative sum, and the union bound proves the desired bound for the absolute value. □

2.6. Weighted Restricted Isometry Property for Sampling Matrix

With our tools from high dimensional probability in hand, we are ready to prove the initial implication in the chain that when combined with with the previously proven theorems on the ω-RIP, the ω-NSP, and their relation to sparse recovery, will prove Theorem 2.2. Thus, we now prove that the sampling matrix for our gPC basis satisfies the ω-RIP with high probability.

Theorem 2.8 ([**24**], Theorem 5.2). *Fix* $\delta, \gamma \in (0,1)$*, and let* $(\phi_\nu)_{\nu \in \Lambda}$ *be a gPC basis on a finite index set with* $|\Lambda| = M$*. Taking a weight sequence such that* $\omega_\nu \geqslant \|\phi_\nu\|_\infty$ *and*

$$K \gtrsim \delta^{-2} s \max\left\{\log^3(s)\log(M), \log(1/\gamma)\right\} \tag{2.12}$$

i.i.d. sampling points $\left\{Z^{(k)}\right\}_{k=1}^K$ *drawn from the orthogonalization measure* π*, with probability exceeding* $1 - \gamma$*, the normalized sampling matrix* $\tilde{A} \in \mathbb{C}^{K,M}$ *with entries* $\tilde{A}_{k,\nu} = \frac{1}{\sqrt{K}}\phi_\nu(Z^{(k)})$ *has* ω*-RIP constant* $\delta_{\omega,s} \leqslant \delta$*.*

Proof. Recall that in Proposition 2.1, we may rewrite the ω-RIP constant of order s of the normalized sampling matrix $\tilde{A}$ as

$$\delta_{\omega,s} = \max_{\omega(S) \leqslant s} \left\|\tilde{A}_S^* \tilde{A}_S - I\right\|.$$

Instead of restricting $\tilde{A}$ to S, we move a step back to the original definition, taking

$$T_\omega^{s,M} = \left\{z \in \mathbb{C}^M \mid \|z\|_2 \leqslant 1,\ \|z\|_{\omega,0} \leqslant s\right\}$$

and therefore

$$\delta_{\omega,s} = \sup_{z \in T_\omega^{s,M}} |\langle(\tilde{A}^*\tilde{A} - I)z, z\rangle|. \tag{2.13}$$

Now notice that to bound the expectation of $\delta_{\omega,s}$, we just need to bound the supremum of a stochastic process indexed over $T = T_{\omega}^{s,M}$. However, in order to use Dudley's inequality, we would need this process to be sub-gaussian, and since we are discussing a general gPC basis, we have no information on how the random matrix $\frac{1}{m}\tilde{A}^*\tilde{A}$ deviates from its mean I. In order to make use of Dudley's inequality, we will instead symmetrize to introduce sub-gaussian Rademacher random variables. However, in order to use symmetrization in the context of the $\mathbb{E}\delta_{\omega,s}$, we need to represent $\delta_{\omega,s}$ as the norm of the sum of centered random vectors.

To this end, we construct a semi-norm on $\mathbb{R}^{M,M}$ as

$$\|B\|_s = \sup_{z\in T_{\omega}^{s,M}} |\langle Bz, z\rangle|.$$

Thus, we can rewrite (2.13) as

$$\delta_{\omega,s} = \left\|\tilde{A}^*\tilde{A} - I\right\|_s.$$

Decomposing A^* as a sum of columns $\sum_{k=1}^{K} X_k e_k^*$ (where $X_k = \left(\overline{\phi_\nu(Z^{(k)})}\right)_{\nu\in\Lambda}$), we have

$$\begin{aligned} A^*A &= \left(\sum_{i=1}^{K} X_i e_i^*\right)\left(\sum_{j=1}^{K} X_j e_j^*\right)^* \\ &= \sum_{i,j=1}^{K} X_i e_i^* e_j X_j^* \\ &= \sum_{k=1}^{K} X_k X_k^*, \end{aligned}$$

where by the fact that ϕ is a gPC basis, $\mathbb{E}X_kX_k^* = I$. Therefore we obtain $\mathbb{E}\delta_{\omega,s}$ as the expectation of the norm of a sum of centered random matrices

$$\mathbb{E}\delta_{\omega,s} = \frac{1}{K}\mathbb{E}\left\|\sum_{k=1}^{K}(X_kX_k^* - I)\right\| = \frac{1}{K}\mathbb{E}\left\|\sum_{k=1}^{K}(X_kX_k^* - \mathbb{E}X_kX_k^*)\right\|.$$

A simple application of Lemma 2.2 on symmetrization implies

$$\begin{aligned} \mathbb{E}\delta_{\omega,s} &\leqslant \frac{2}{K}\mathbb{E}_X\mathbb{E}_\varepsilon\left\|\sum_{k=1}^{K}\varepsilon_kX_kX_k^*\right\| \\ &= \frac{2}{K}\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{z\in T_{\omega}^{s,M}}\left|\sum_{k=1}^{K}\langle\varepsilon_kX_kX_k^*z, z\rangle\right| \\ &= \frac{2}{K}\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{z\in T_{\omega}^{s,M}}\left|\sum_{k=1}^{K}\varepsilon_k|\langle X_k, z\rangle|^2\right| \\ &=: \frac{2}{K}\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{z\in T_{\omega}^{s,M}}|Y_z| \end{aligned} \tag{2.14}$$

where the ε_k are independent (with the $X_kX_k^*$ as well), Rademacher random variables.

We now find ourselves in prime territory for Dudley's inequality (conditional on X). As a process, the Rademacher sum has increments

$$Y_z - Y_x = \sum_{k=1}^{K}\left(|\langle X_k, z\rangle|^2 - |\langle X_k, x\rangle|^2\right)\varepsilon_k.$$

By Proposition 2.3 and the fact that Rademacher random variables have sub-gaussian norm $\|\varepsilon\|_{\psi_2} = C$,

$$\begin{aligned}\|Y_z - Y_x\|_{\psi_2}^2 &\leqslant C \sum_{k=1}^{K} \left(|\langle X_k, z\rangle|^2 - |\langle X_k, x\rangle|^2\right)^2 \|\varepsilon_k\|_{\psi_2}^2 \\ &\leqslant C \sum_{k=1}^{K} \left(|\langle X_k, z\rangle|^2 - |\langle X_k, x\rangle|^2\right)^2 .\end{aligned}$$

Defining the pseudo-metric

$$d(z, x) = \left(\sum_{k=1}^{K} \left(|\langle X_k, z\rangle|^2 - |\langle X_k, x\rangle|^2\right)^2\right)^{1/2},$$

we have that Y_z is sub-gaussian with respect to the metric space $(T_\omega^{s,M}, d)$. Choosing $0 = t_0 \in T_\omega^{s,M}$ and applying Theorem 2.6, Dudley's inequality,

$$\mathbb{E}_\varepsilon \sup_{z\in T_\omega^{s,M}} |Y_z| \leqslant C \int_0^\infty \sqrt{\log \mathcal{N}(T_\omega^{s,M}, d, u)}\, du. \tag{2.15}$$

The majority of the remainder of the proof consists of bounding the integrand for covering numbers at different scales.

The first step to bounding the covering numbers is to rewrite $d(z, x)$ in terms of the more friendly 2-norm to which we can apply volumetric arguments. Applying Hölder's inequality,

$$\begin{aligned}d(z, x) &= \left(\sum_{k=1}^{K} \left(|\langle X_k, z\rangle|^2 - |\langle X_k, x\rangle|^2\right)^2\right)^{1/2} \\ &= \left(\sum_{k=1}^{K} (|\langle X_k, z\rangle| + |\langle X_k, x\rangle|)^2 (|\langle X_k, z\rangle| - |\langle X_k, x\rangle|)^2\right)^{1/2} \\ &\leqslant \left(\sum_{k=1}^{K} (|\langle X_k, z\rangle| + |\langle X_k, x\rangle|)^2 |\langle X_k, z - x\rangle|^2\right)^{1/2} \\ &\leqslant \left(\sum_{k=1}^{K} (|\langle X_k, z\rangle| + |\langle X_k, x\rangle|)^{2p}\right)^{1/2p} \left(\sum_{k=1}^{K} |\langle X_k, z - x\rangle|^{2q}\right)^{1/2q} \\ &\leqslant 2 \sup_{z\in T_\omega^{s,M}} \left(\sum_{k=1}^{K} |\langle X_k, z\rangle|^{2p}\right)^{1/2p} \left(\sum_{k=1}^{K} |\langle X_k, z - x\rangle|^{2q}\right)^{1/2q} .\end{aligned}$$

In the first term, we can use that $|(X_k)_\nu| \leqslant \omega_\nu$, $z \in T_\omega^{s,M}$, and the weighted Cauchy-Schwarz inequality to give

$$|\langle X_k, z\rangle| \leqslant \sum_{\nu\in\Lambda} \omega_\nu |z_\nu| \leqslant \|z\|_2 \sqrt{\sum_{\nu\in \text{supp}(z)} \omega_\nu^2} \leqslant \sqrt{s}\|z\|_2 \leqslant \sqrt{s} \tag{2.16}$$

Rather than applying this to the entire first term, we separate summand and apply giving

$$|\langle X_k, z\rangle|^{2p} = |\langle X_k, z\rangle|^2 |\langle X_k, z\rangle|^{2(p-1)} \leqslant s^{p-1} |\langle X_k, z\rangle|^2.$$

This will later end up allowing a better logarithmic factor in required number of measurements than standard arguments. Thus,

$$d(z,x) \leqslant 2s^{(p-1)/2p} \sup_{z \in T_\omega^{s,M}} \left(\sum_{k=1} |\langle X_k, z\rangle|^2 \right)^{1/2p} \left(\sum_{k=1}^{K} |\langle X_k, z-x\rangle|^{2q} \right)^{1/2q}.$$

The last piece defines our new semi-norm

$$\|z\|_{X,q} = \left(\sum_{k=1}^{K} |\langle X_k, z\rangle|^{2q} \right)^{1/2q}$$

which we will use in the Dudley integral, since

$$d(z,x) \leqslant C(s,p,X)\|z-x\|_{X,q}, \quad \text{with } C(s,p,X) = 2s^{(p-1)/2p} \sup_{z \in T_\omega^{s,M}} \left(\sum_{k=1}^{K} |\langle X_k, z\rangle|^2 \right)^{1/2p}.$$

For any metrics satisfying $d(x,z) \leqslant cd'(x,z)$, we know that

$$u \geqslant d'(x,z) \geqslant \frac{1}{c} d'(x,z) \implies cu \geqslant d(x,z), \tag{2.17}$$

and thus, any u-net with respect to d' is a cu-net with respect to d. Thus the covering number for d at the scale cu must be bounded by the covering number for d' at the scale u, that is

$$\mathcal{N}(T,d,cu) \leqslant \mathcal{N}(T,d',u).$$

In the Dudley integral, this bound and a change of variables gives

$$\begin{aligned} \int_0^\infty \sqrt{\log \mathcal{N}(T,d,u)}\, du &= \int_0^\infty \sqrt{\log \mathcal{N}\left(T,d,\frac{cu}{c}\right)}\, du \\ &= c\int_0^\infty \sqrt{\log \mathcal{N}(T,d,cu)}\, du \\ &\leqslant c\int_0^\infty \sqrt{\log \mathcal{N}(T,d',u)}\, du. \end{aligned}$$

Thus, for our scenario,

$$\int_0^\infty \sqrt{\log \mathcal{N}(T_\omega^{s,M},d,u)}\, du \leqslant C(s,p,X) \int_0^\infty \sqrt{\log \mathcal{N}(T_\omega^{s,M},\|\cdot\|_{X,q},u)}\, du \tag{2.18}$$

We will eventually split the Dudley integral into two pieces, so that we integrate over smaller u and larger u using separate bounds on the covering number. For smaller u, we proceed as follows. By the same reasoning in (2.16), we know that for $z \in T_\omega^{s,M}$,

$$\|z\|_{X,q} \leqslant K^{1/2q}\sqrt{s}\|z\|_2. \tag{2.19}$$

Thus, if we have a u-net with respect to $K^{1/2q}\sqrt{s}\|\cdot\|_2$, this is also a u-net with respect to $\|\cdot\|_{X,q}$, allowing us to instead bound

$$\mathcal{N}(T_\omega^{s,M}, K^{1/2q}\sqrt{s}\|\cdot\|_2, u) \geqslant \mathcal{N}(T_\omega^{s,M}, \|\cdot\|_{X,q}, u).$$

Additionally, applying (2.17) (with equality instead of an inequality), we find

$$\mathcal{N}(T_\omega^{s,M}, K^{1/2q}\sqrt{s}\|\cdot\|_2, u) = \mathcal{N}(T_\omega^{s,M}, \|\cdot\|_2, K^{-1/2q}s^{-1/2}u),$$

allowing us to consider standard covering numbers in the friendlier ℓ_2-norm. The standard way to do this is by a volumetric argument.

We will start by splitting $T^{s,M}_{\omega}$ into the union of lower dimensional balls

$$T^{s,M}_{\omega} = \bigcup_{\substack{S\subseteq\Lambda\\ \omega(S)\leqslant s}} B_S, \qquad \text{for } B_S = \{z \in \mathbb{C}^M \mid \operatorname{supp}(x) \subseteq S,\ \|x\|_2 \leqslant 1\}.$$

If we denote the $\mathbb{C}^m$ ℓ_2-ball of radius r centered at $z \in \mathbb{C}^m$ by $B^m(x,r)$, note that we can identify $B_S \equiv B^{|S|}(0,1)$. Then if we have a u-net for each B_S, the union of these nets is a u-net for $T^{s,M}_{\omega}$, giving

$$\mathcal{N}(T^{s,M}_{\omega}, \|\cdot\|_2, K^{-1/2q}s^{-1/2}u) \leqslant \sum_{\substack{S\subseteq\Lambda\\ \omega(S)\leqslant s}} \mathcal{N}(B_S, \|\cdot\|_2, K^{-1/2q}s^{-1/2}u).$$

Our final step is to estimate the covering number of B_S at the scale $u' = K^{-1/2q}s^{-1/2}u$. In order to do this, we consider a maximal u' packing of B_S, that is the maximal set $\{z_1, \dots, z_P\} \subseteq B_S$ such that $B^{|S|}(z_i, u'/2) \cap B^{|S|}(z_j, u'/2) = \emptyset$ for all $i \neq j \in [P]$. This maximal packing must also be a u'-net, since if not, there is some z^* further than u' away from every element of the maximal packing, contradicting the fact that the maximal packing is maximal. Additionally, since every element in the packing has norm bounded by one, $B^{|S|}(z_i, u'/2) \subseteq B^{|S|}(0, 1+u'/2)$ for every element in the packing. These balls being disjoint implies

$$P\left|B^{|S|}(0,u'/2)\right| = \sum_{i=1}^{P}\left|B^{|S|}(z_i,u'/2)\right| = \left|\bigcup_{i=1}^{P} B^{|S|}(z_i,u'/2)\right| \leqslant \left|B^{|S|}(0,1+u'/2)\right|.$$

Identifying $\mathbb{C}^{|S|}$ balls with $\mathbb{R}^{2|S|}$ and using that $\left|B^{|S|}(0,r)\right| = r^{2|S|}B^{|S|}(0,1)$, we then find

$$\mathcal{N}(B_S, \|\cdot\|_2, u') \leqslant \frac{\left|B^{|S|}(0,1+u'/2)\right|}{\left|B^{|S|}(0,u'/2)\right|} = \frac{(1+u'/2)^{2|S|}}{(u'/2)^{2|S|}} = \left(1 + \frac{2K^{1/2q}\sqrt{s}}{u}\right)^{2|S|}.$$

Piecing together, we then find

$$\mathcal{N}(T^{s,M}_{\omega}, \|\cdot\|_{X,q}, u) \leqslant \sum_{\substack{S\subseteq\Lambda\\ \omega(S)\leqslant s}} \left(1 + \frac{2K^{1/2q}\sqrt{s}}{u}\right)^{2|S|}.$$

However, since $\omega_\nu \geqslant 1$, we have $\omega(S) \geqslant |S|$, and so we can instead sum over all unweighted up to s-sparse index sets. We can enumerate the exactly r-sparse index sets by simply taking all M indices and choosing r of them, giving

$$|\{S \subseteq \Lambda \mid \omega(S) \leqslant s\}| \leqslant \sum_{r=1}^{s} |\{S \subseteq \Lambda \mid \omega(S) = r\}| \leqslant \sum_{r=1}^{s} \binom{M}{r}.$$

We can then approximate

$$\binom{M}{r} = \frac{M(M-1)\cdots(M-(r+1))}{r!} \leqslant \frac{M^r}{r^r}\frac{r^r}{r!} \leqslant \left(\frac{eM}{r}\right)^r. \tag{2.20}$$

Taking derivatives, we see that this last quantity is increasing in r, and so

$$|\{S \subseteq \Lambda \mid \omega(S) \leqslant s\}| \leqslant s\left(\frac{eM}{s}\right)^s \leqslant (eM)^s.$$

Finally, we obtain

$$\mathcal{N}(T^{s,M}_{\omega}, \|\cdot\|_{X,q}, u) \leqslant (eM)^s \left(1 + \frac{2K^{1/2q}\sqrt{s}}{u}\right)^{2s}. \tag{2.21}$$

Now, we need to derive a bound for large values of u. Here, we will use Lemma 2.3, Maurey's lemma. After we get a bound on the covering number on the convex hull of some points $\mathcal{U}$, we will want to link this to the covering number of $T_{\omega}^{s,M}$. We will choose $\mathcal{U}$ so that $T_{\omega}^{s,M}$ is (almost) a subset, and then the best way to link the covering numbers is to show that for any sets $T \subseteq S$, we have

$$\mathcal{N}(T, \|\cdot\|, u) \leqslant \mathcal{N}(S, \|\cdot\|, u/2). \tag{2.22}$$

From a previous discussion, we know that the covering number of T at scale u is bounded by the packing number at the same scale. This packing will then be a packing of S, so bounds the packing number of S at that scale from below. The last step is to show that this packing number bounds the covering number at scale $u/2$. But if we have a u packing and a $u/2$ net, each packing point must be contained in a "net ball". In fact, based on the u separation of point in the packing, there can be at most one packing point in each net ball. Thus, there must be more net points than packing points, implying the same relationship for the covering and packing numbers, finishing the argument for (2.22).

Now we need to come up with the proper set of points to use for Maurey's lemma. We at least need to be sure that we can satisfy the hypothesis of Maurey's lemma, that for any $L \in \mathbb{N}_+$ and collection of points $(u_1, \ldots, u_L) \in \mathcal{U}^L$, $\mathbb{E}_\varepsilon \left\| \sum_{i=1}^L \varepsilon_i u_i \right\| \leqslant A\sqrt{L}$ for a Rademacher sequence ε. To this end, let us keep $\mathcal{U}$ general and see what this bound means in terms of our norm $\|\cdot\|_{X,q}$.

We calculate

$$\begin{aligned}
\mathbb{E}_\varepsilon \left\| \sum_{i=1}^L \varepsilon_i u_i \right\|_{X,q} &= \mathbb{E}_\varepsilon \left(\sum_{k=1}^K \left| \left\langle X_k, \sum_{i=1}^L \varepsilon_i u_i \right\rangle \right|^{2q} \right)^{1/2q} \\
&\leqslant \left(\sum_{k=1}^K \mathbb{E}_\varepsilon \left| \sum_{i=1}^L \varepsilon_i \langle X_k, u_i \rangle \right|^{2q} \right)^{1/2q} \\
&\leqslant C\sqrt{2q} \left(\sum_{k=1}^K \left\| (\langle X_k, u_i \rangle)_{i=1}^L \right\|_2^{2q} \right)^{1/2q},
\end{aligned}$$

by Kintchine's inequality. If we are able to choose $|\langle X_k, u_i \rangle| = D$ constant, we would obtain,

$$\mathbb{E}_\varepsilon \left\| \sum_{i=1}^L \varepsilon_i u_i \right\|_{X,q} \leqslant CD\sqrt{2q}K^{1/2q}\sqrt{L},$$

as desired. In order to do this, a simple choice is to consider what happens when u_i is some multiple of of a canonical basis vector, $u_i = c_\nu e_\nu$. Then

$$|\langle X_k, u_i \rangle| = c_\nu |(X_k)_\nu| = c_\nu \left| \phi_\nu(Z^{(k)}) \right| \leqslant c_\nu \omega_\nu.$$

Choosing $c_\nu = \omega_\nu^{-1}$ then gives $D = 1$ uniformly for all pairs of i and k as desired.

Our first pass is to then choose

$$\mathcal{U}' = \{\omega_\nu^{-1} e_\nu \mid \nu \in \Lambda\}.$$

However, it is clear that $T_{\omega}^{s,M}$ cannot be contained in $\operatorname{conv} \mathcal{U}'$. In particular, at the very least, $T_{\omega}^{s,M}$ contains negative and imaginary numbers. So let us revise our point set to be

$$\mathcal{U} = \{\pm\omega_\nu^{-1} e_\nu, \pm i\omega_\nu^{-1} e_\nu \mid \nu \in \Lambda\},$$

noting that $|\langle X_k, u_i\rangle| \leqslant 1$ for points in this set as well. However, we again find that $T_\omega^{s,M}$ is not necessarily contained in $\operatorname{conv}(\mathcal{U})$. Indeed, for some ν with $\sqrt{s} \geqslant \omega_\nu > 1$, we have $\|e_\nu\|_{\omega,0} = \omega_\nu^2 \leqslant s$, with $\|e_\nu\|_2 = 1$ giving that $e_\nu \in T_\omega^{s,M}$. However, $\|e_\nu\|_2 > \left\|\omega_\nu^{-1} e_\nu\right\|$ which is the longest vector in $\operatorname{conv}(\mathcal{U})$ which "points" in the ν direction, giving that $e_\nu \notin \operatorname{conv}(\mathcal{U})$.

Let us see then where our ability to represent $z \in T_\omega^{s,M}$ as a convex combination of points in $\mathcal{U}$ fails. We begin by writing $z = (z_\nu)_{\nu\in\Lambda} = (x_\nu + iy_\nu)_{\nu\in\Lambda}$. Then we have

$$\begin{aligned} z &= \sum_{\nu\in\Lambda} \operatorname{sgn}(x_\nu)|x_\nu|e_\nu + i\operatorname{sgn}(y_\nu)|y_\nu|e_\nu \\ &= \sum_{\nu\in\Lambda} \operatorname{sgn}(x_\nu)\omega_\nu^{-1}e_\nu(\omega_\nu|x_\nu|) + i\operatorname{sgn}(y_\nu)\omega_\nu^{-1}e_\nu(\omega_\nu|y_\nu|). \end{aligned}$$

Since $\sum_{\nu\in\Lambda} \omega_\nu x_\nu + \omega_\nu y_\nu$ is not required to sum up to one, this is not then a convex combination. As a compromise, we can make this a convex combination by normalizing by this sum, which we represent $\|z\|_1^*$. Then $z/\|z\|_1^* \in \operatorname{conv}(\mathcal{U})$. This gives us the idea to no longer consider $T_\omega^{s,M}$ being a subset of just $\operatorname{conv}(\mathcal{U})$, but as a rescaled version. In order to do this, we need a uniform bound on $\|z\|_1^*$ over $T_\omega^{s,M}$. Our weighted version of Cauchy-Schwarz again does the trick, as

$$\|z\|_1^* = \sum_{\nu\in\Lambda} \omega_\nu|x_\nu| + \omega_\nu|y_\nu| \leqslant \sqrt{s}(\|x\|_2 + \|y\|_2) \leqslant 2\sqrt{s},$$

since $\|z\|_2 \leqslant 1$. Now, since $0 \in \operatorname{conv}(\mathcal{U})$ (which is convex) and $z/\|z\|_1^* \in \operatorname{conv}(\mathcal{U})$, we must have the shorter $z/(2\sqrt{s}) \in \operatorname{conv}(\mathcal{U})$. Thus, $T_\omega^{s,M} \subseteq 2\sqrt{s}\operatorname{conv}(\mathcal{U})$.

Combining (2.22) and the bound from Maurey's lemma, we find

$$\begin{aligned} \mathcal{N}(T_\omega^{s,M}, \|\cdot\|_{X,q}, u) &\leqslant \mathcal{N}(2\sqrt{s}\operatorname{conv}(\mathcal{U}), \|\cdot\|_{X,q}, u/2) \\ &\leqslant \mathcal{N}(\operatorname{conv}(\mathcal{U}), \|\cdot\|_{X,q}, u/(4\sqrt{s})) \\ &\leqslant \exp\Big(C(sqK^{1/q})/u^2\log(4M)\Big), \end{aligned}$$

and in terms of the Dudley integrand,

$$\sqrt{\log\mathcal{N}(T_\omega^{s,M}, \|\cdot\|_{X,q}, u)} \leqslant C\sqrt{sqK^{1/q}\log(4M)}/u. \tag{2.23}$$

We are now ready to estimate the Dudley integral. First, we note that since any $z \in T_\omega^{s,M}$ has $\|z\|_2 \leqslant 1$, by the Cauchy-Schwarz argument giving (2.19), we know that $\|z\|_{X,q} \leqslant K^{1/2q}\sqrt{s}$. Thus, for any u-net of $T_\omega^{s,M}$ with $u \geqslant K^{1/2q}\sqrt{s}$, we will be able to choose just one point in the net, and therefore the root-logarithm of the covering number will be zero. So it suffices to integrate up to $K^{1/2q}\sqrt{s}$. As previously mentioned, we have derived (2.21) for small values of u and (2.23) for large values of u. Thus, we split the integral at some $\alpha \in (0, K^{1/2q}\sqrt{s})$ to be determined, giving

$$\begin{aligned}
\int_0^{K^{1/2q}\sqrt{s}} \sqrt{\log \mathcal{N}(T_\omega^{s,M}, \|\cdot\|_{X,q}, u)}\ du &= \int_0^{\alpha} \sqrt{\log \mathcal{N}(T_\omega^{s,M}, \|\cdot\|_{X,q}, u)}\ du \\
&\quad + \int_\alpha^{K^{1/2q}\sqrt{s}} \sqrt{\log \mathcal{N}(T_\omega^{s,M}, \|\cdot\|_{X,q}, u)}\ du \\
&\leqslant \int_0^{\alpha} \sqrt{s \log(eM) + 2s \log\left(1 + 2K^{1/2q}\sqrt{s}u^{-1}\right)}\ du \\
&\quad + C\sqrt{sqK^{1/q}\log(4M)} \int_\alpha^{K^{1/2q}\sqrt{s}} u^{-1}\ du \\
&= \int_0^{\alpha} \sqrt{s \log(eM) + 2s \log\left(1 + 2K^{1/2q}\sqrt{s}u^{-1}\right)}\ du \\
&\quad + C\sqrt{sqK^{1/q}\log(4M)} \log\left(K^{1/2q}\sqrt{s}\alpha^{-1}\right).
\end{aligned}$$

In order to handle the remaining integral I, we can first use the fact that for positive a, b, $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$. Splitting the square root and integrating the first term gives

$$I \leqslant \alpha\sqrt{s\log(eM)} + \sqrt{2s}\int_0^\alpha \sqrt{\log\left(1 + 2K^{1/2q}\sqrt{s}u^{-1}\right)}\ du. \tag{2.24}$$

In order to handle the remaining integral, we let $2K^{1/2q}\sqrt{s} = \beta$, and use Cauchy-Schwarz to find

$$\int_0^\alpha \sqrt{\log(1+\beta u^{-1})}\ du \leqslant \sqrt{\alpha}\sqrt{\int_0^\alpha \log(1+\beta u^{-1})\ du.}$$

The substitution $t = \beta u^{-1}$ gives

$$\int_0^\alpha \log\left(1+\beta u^{-1}\right)\ du = \beta \int_{\beta/\alpha}^\infty t^{-2}\log(1+t)\ dt,$$

and integrating by parts gives

$$\begin{aligned}
\int_0^\alpha \log\left(1+\beta u^{-1}\right)\ du &= \beta\left[t^{-1}\log(1+t)\Big|_\infty^{\beta/\alpha} + \int_{\beta/\alpha}^\infty \frac{1}{t(t+1)}\ dt\right] \\
&\leqslant \beta\left[\frac{\alpha}{\beta}\log(1+\beta/\alpha) + \int_{\beta/\alpha}^\infty t^{-2}\ dt\right] \\
&= \alpha\log(1+\beta/\alpha) + \alpha \\
&= \alpha\log\left(e\left(1 + 2K^{1/2q}\sqrt{s}\alpha^{-1}\right)\right).
\end{aligned}$$

Taking the square root, multiplying by $\sqrt{\alpha}$ and plugging into (2.24), gives the final estimate

$$\begin{aligned}
\int_0^{K^{1/2q}\sqrt{s}} \sqrt{\log \mathcal{N}\left(T_\omega^{s,M}, \|\cdot\|_{X,q}, u\right)}\ du &\leqslant \alpha\sqrt{s\log(eM)} + \alpha\sqrt{2s}\sqrt{\log\left(e\left(1 + 2K^{1/2q}\sqrt{s}\alpha^{-1}\right)\right)} \\
&\quad + C\sqrt{sqK^{1/q}\log(4M)}\log\left(K^{1/2q}\sqrt{s}\alpha^{-1}\right).
\end{aligned}$$

In order to remove any logarithmic dependence on the number of measurements, we can then choose $\alpha = K^{1/2q}$, giving

$$\begin{aligned}\int_0^{K^{1/2q}\sqrt{s}} \sqrt{\log \mathcal{N}\left(T_\omega^{s,M}, \|\cdot\|_{X,q}, u\right)}\, du &\leqslant \sqrt{sK^{1/q}\log(eM)} + \sqrt{2sK^{1/q}\log\left(e\left(1+2\sqrt{s}\right)\right)} \\ &\quad + C\sqrt{sqK^{1/q}\log(4M)\log^2(s)} \\ &\leqslant K^{1/2q}\sqrt{sq}\left(\sqrt{\log(eM)} + \sqrt{\log(9e^2 s)}\right. \\ &\quad \left. + \sqrt{C\log(4M)\log^2(s)}\right).\end{aligned}$$

We can combine these terms making the (extremely mild) assumption (which we probably already implicitly used somewhere else) that $s, M \geqslant 2$, since then $\log(s), \log(M) \geqslant \log(2)$. Letting n represent either M or s and for an absolute constant c, we will invoke the general strategy

$$\log(cn) = \log(c) + \log(n) \leqslant \left(\frac{\log(c)}{\log(2)} + 1\right)\log(n) := c'\log(n).$$

We may also additionally add an additional multiplicative factor of $\log(n)$ to any term by paying an absolute factor of $1/\log(2)$. Thus, we find

$$\int_0^{\infty} \sqrt{\log \mathcal{N}\left(T_\omega^{s,M}, \|\cdot\|_{X,q}, u\right)}\, du \leqslant C\sqrt{K^{1/q}sq\log(M)\log^2(s)}. \tag{2.25}$$

Combining this last estimate (2.25), The fact that this was a bound (2.18) on the expectation on supremum of the Rademacher stochastic process shown in (2.15), and the fact that this expectation was a bound on the expectation of the ω-RIP constant brings us to

$$\begin{aligned}\mathbb{E}\delta_{\omega,s} &\leqslant C\frac{1}{K}s^{(p-1)/2p}\sqrt{K^{1/q}sq\log(M)\log^2(s)}\,\mathbb{E}_X \sup_{z\in T_\omega^{s,M}}\left(\sum_{k=1}^K |\langle X_k, z\rangle|^2\right)^{1/2p} \\ &= Cs^{1-1/2p}K^{1/2q-1}\sqrt{q\log(M)\log^2(s)}\,\mathbb{E}\sup_{z\in T_\omega^{s,M}}\left(\sum_{k=1}^K |\langle X_k, z\rangle|^2\right)^{1/2p}.\end{aligned}$$

If we remember where this last term originally came from, it was to represent $\tilde{A}^*\tilde{A}$ as the sum

$$\frac{1}{K}\mathbb{E}\left\|\sum_{k=1}^K (X_k X_k^* - I)\right\|_s = \mathbb{E}\left\|\tilde{A}^*\tilde{A} - I\right\|_s = \mathbb{E}\delta_{\omega,s} \tag{2.26}$$

so that we could apply symmetrization. In the reverse direction,

$$\begin{aligned}\sum_{k=1}^K |\langle X_k, z\rangle|^2 &= \sum_{k=1}^K \langle X_k X_k^* z, z\rangle \\ &= K\sum_{k=1}^K \frac{1}{K}\langle X_k X_k^* z, z\rangle - K\langle Iz, z\rangle + K\langle Iz, z\rangle \\ &= K\left[\left\langle\left(\sum_{k=1}^K \frac{1}{K}X_k X_k^* - I\right)z, z\right\rangle + \langle Iz, z\rangle\right] \\ &= K\left[\left\langle\left(\tilde{A}^*\tilde{A} - I\right)z, z\right\rangle + \langle Iz, z\rangle\right].\end{aligned}$$

Taking 1/2p powers, supremums, and expectations gives

$$\mathbb{E} \sup_{z \in T^{s,M}_{\omega}} \left(\sum_{k=1}^{K} |\langle X_k, z \rangle|^2 \right)^{1/2p} \leqslant K^{1/2p} \mathbb{E} \left(\left\| \tilde{A}^* \tilde{A} - I \right\|_s + \|I\|_s \right)^{1/2p} = K^{1/2p} \mathbb{E} \left(\delta_{\omega,s} + 1 \right)^{1/2p}.$$

By Jensen's inequality and the fact that $p \geqslant 1$ and the integrand is necessarily at least one,

$$\mathbb{E} \sup_{z \in T^{s,M}_{\omega}} \left(\sum_{k=1}^{K} |\langle X_k, z \rangle|^2 \right)^{1/2p} \leqslant K^{1/2p} \sqrt{\mathbb{E}\delta_{\omega,s} + 1},$$

and plugging into our bound for $\mathbb{E}\delta_{\omega,s}$ above,

$$\begin{aligned} \mathbb{E}\delta_{\omega,s} &\leqslant C s^{1-1/2p} K^{(1/2)(1/q+1/p)-1} \sqrt{q \log(M) \log^2(s)} \sqrt{\mathbb{E}\delta_{\omega,s} + 1} \\ &= C \sqrt{s^{2-1/p} q \frac{\log(M) \log^2(s)}{K}} \sqrt{\mathbb{E}\delta_{\omega,s} + 1}. \end{aligned}$$

We are ready now to apply a tricky choice of $p = 1 + 1/\log(s)$ and $q = 1 + \log(s) \leqslant c \log(s)$, which gives

$$1 - \frac{1}{p} = 1 - \frac{1}{1 + 1/\log(s)} = \frac{1/\log(s)}{1 + 1/\log(s)} \leqslant \frac{1}{\log(s)},$$

and

$$s^{2-1/p} \leqslant s s^{1/\log(s)} = s e^{\log(s)(1/\log(s))} = es.$$

Thus,

$$\mathbb{E}\delta_{\omega,s} \leqslant C \sqrt{\frac{s \log(M) \log^3(s)}{K}} \sqrt{\mathbb{E}\delta_{\omega,s} + 1}.$$

We now perform some algebra to solve for $\mathbb{E}\delta_{\omega,s}$. Moving C under the square root, and assuming K is large enough to make this square root term bounded by one, we must reorder an inequality of the form

$$x \leqslant a\sqrt{x+1}, \qquad a \leqslant 1.$$

Adding one to both sides, and letting $y = \sqrt{x+1}$, we have

$$\begin{aligned} & x \leqslant a\sqrt{x+1} \\ \Longrightarrow\ & y^2 - ay \leqslant 1 \\ \Longrightarrow\ & \left(y - \frac{a}{2}\right)^2 \leqslant 1 + \left(\frac{a}{2}\right)^2 \\ \Longrightarrow\ & x \leqslant (1+a)^2 - 1 \\ \Longrightarrow\ & x \leqslant a^2 + 2a \leqslant 3a, \end{aligned}$$

since $a \leqslant 1$. Then we finally obtain

$$\mathbb{E}\delta_{\omega,s} \leqslant C \sqrt{\frac{s \log(M) \log^3(s)}{K}},$$

so long as

$$K \gtrsim s \log(m) \log^3(s).$$

We have then shown a bound on the expectation of the ω-RIP constant, but we now have to show that this bound holds with high probability. One approach is to consider a modified Dudley's inequality to show that instead of just bounding expectations of the supremum of a stochastic process, we can bound L^p norms (notice in the proof that we did not make use of the strength of

Property (2) in Proposition 2.2, only the bound on expectation). The resulting argument would be more involved, but produce similar bounds with the exception of a factor of $\sqrt{p}$, which, by Proposition 2.2, proves that the ω-RIP constant is sub-gaussian, and will therefore satisfy the concentration inequality in Property (1) which can prove a "high probability" bound. This is the method used to prove a weaker probability version of Theorem 2.1 in [**22**, Theorem 8.1] (and before being too hasty, we remark that we have not verified the details to show that this method works the same way for the weighted setup). However, as also shown in [**22**, Theorem 8.4], we can strengthen this probability result by using Bernstein's inequality for empirical processes, Theorem 2.7.

In order to apply this theorem, we need to represent the ω-RIP constant as the supremum of an empirical process

$$\delta_{\omega,s} = \sup_{f \in \mathcal{F}} \sum_{k=1}^{K} f(Y_k)$$

with the bounds

$$f(Y_k) \leqslant L, \quad \mathbb{E}f(Y_k) = 0, \quad \text{and } \mathbb{E}[f(Y_k)^2] \leqslant \sigma_k^2.$$

As we recall, e.g. from (2.26), we know

$$K\delta_{\omega,s} = \left\| \sum_{k=1}^{K} (X_k X_k^* - I) \right\|_s = \sup_{z \in T_\omega^{s,M}} \left| \left\langle \left(\sum_{k=1}^{K} (X_k X_k^* - I) \right) z, z \right\rangle \right|,$$

where $\mathbb{E}(X_k X_k^* - I) = 0$. This is our starting point.

It will turn out that our class of functions should take X_k as input, but we will need to rewrite the previous quantity linearly, so that we can pull out the sum. If we instead switch this expression back to operator norm, and rewrite using the quadratic form induced by $\sum_{k=1}^{K}(X_k X_k^* - I)$, we will be able to do so. Thus, we rewrite

$$\begin{aligned}
\sup_{z \in T_\omega^{s,M}} \left| \left\langle \left(\sum_{k=1}^{K} (X_k X_k^* - I) \right) z, z \right\rangle \right| &= \sup_{\substack{S \subseteq \Lambda \\ \omega(S) \leqslant s}} \left\| \sum_{k=1}^{K} (X_k^S (X_k^S)^* - I_S) \right\| \\
&= \sup_{\substack{S \subseteq \Lambda \\ \omega(S) \leqslant s}} \sup_{z,w \in S^{|S|-1}} \operatorname{Re} \left\langle \left(\sum_{k=1}^{K} (X_k^S (X_k^S)^* - I_S) \right) z, w \right\rangle \\
&= \sup_{(z,w) \in Q_\omega^{s,M}} \operatorname{Re} \left\langle \left(\sum_{k=1}^{K} (X_k X_k^* - I) \right) z, w \right\rangle \\
&= \sup_{(z,w) \in Q_\omega^{s,M}} \sum_{k=1}^{K} \operatorname{Re} \left\langle (X_k X_k^* - I) z, w \right\rangle,
\end{aligned}$$

where

$$Q_\omega^{s,M} = \bigcup_{\substack{S \subseteq \Lambda \\ \omega(S) \leqslant s}} \{(z,w) \in \mathbb{S}^{M-1} \times \mathbb{S}^{M-1} \mid \operatorname{supp}(z), \operatorname{supp}(w) \subseteq S\}.$$

Since this bi-linear operator will be shown to be bounded and therefore continuous over $Q_\omega^{s,M}$, it suffices to consider a countable dense subset of $Q_\omega^{s,M}$, and thus, we can assume $Q_\omega^{s,M}$ is countable. This gives us the countable set of functions

$$f_{z,w}(Y) = \operatorname{Re}\langle (YY^* - I)z, w \rangle, \quad (z,w) \in Q_\omega^{s,M},$$

which has

$$K\delta_{\omega,s} = \sup_{z,w \in Q_{\omega}^{s,M}} \sum_{k=1}^{K} f_{z,w}(X_k).$$

Now, pushing expectations into the inner product, we immediately have $\mathbb{E}f_{z,w}(X_k) = 0$ for all $k \in [K]$. It remains to check the boundedness and variance conditions before we use Bernstein's inequality. For the boundedness, we first note for $\operatorname{supp}(z), \operatorname{supp}(w) \subseteq S$,

$$|f_{z,w}(X_k)| \leqslant \left|\langle (X_k^S(X_k^S)^* - I_S)z, w\rangle\right| \leqslant \left\|X_k^S(X_k^S)^* - I_S\right\|.$$

Using the characterization of the operator norm for Hermitian matrices, we obtain

$$\begin{aligned}|f_{z,w}(X_k)| &\leqslant \max_{z \in \mathbb{S}^{|S|-1}} \left|\langle (X_k^S(X_k^S)^* - I)z, z\rangle\right| \\ &\leqslant \left|\max_{x \in \mathbb{S}^{|S|-1}} \langle X_k^S(X_k^S)^* x, x\rangle - 1\right|.\end{aligned}$$

If the maximum has absolute value smaller than one, then so does $|f_{z,w}(X_k)|$. If not, we have

$$\left|\max_{x \in \mathbb{S}^{|S|-1}} \langle X_k^S(X_k^S)^* x, x\rangle - 1\right| = \max_{x \in \mathbb{S}^{|S|-1}} \langle X_k^S(X_k^S)^* x, x\rangle - 1 \leqslant \max_{x \in \mathbb{S}^{|S|-1}} \langle X_k^S(X_k^S)^* x, x\rangle = \max_{x \in \mathbb{S}^{|S|-1}} \left|\langle X_k^S, x\rangle\right|^2$$

which is bounded by s by the weighted Cauchy-Schwarz property and the fact that $\omega(S) \leqslant s$. Thus, $|f_{z,w}(X_k)| \leqslant L = s$ for all $k \in [K]$. Finally we bound the variance as

$$\begin{aligned}\mathbb{E}|f_{z,w}(X_k)|^2 &= \mathbb{E}|\mathrm{Re}\langle (X_k X_k^* - I)z, w\rangle|^2 \\ &= \mathbb{E}\left[\mathrm{Re}\langle X_k X_k^* z, w\rangle\right]^2 - 2\mathbb{E}\mathrm{Re}\langle X_k X_k^* z, w\rangle \mathrm{Re}\langle z, w\rangle + (\mathrm{Re}\langle z, w\rangle)^2 \\ &\leqslant \mathbb{E}|\langle X_k X_k^* z, w\rangle|^2 - 2\mathrm{Re}\left[\mathbb{E}\overline{\langle X_k X_k^* w, z\rangle}\langle z, w\rangle\right] + |\langle z, w\rangle|^2 \\ &\leqslant \mathbb{E}\left[|\langle X_k, z\rangle|^2 |\langle X_k, w\rangle|^2\right] - |\langle z, w\rangle|^2,\end{aligned}$$

where we have used that $\mathbb{E}X_k X_k^* = I$ to combine the cross term with $|\langle z, w\rangle|^2$. Now, using the Cauchy-Schwarz argument again on $|\langle X_k, z\rangle|^2$ gives

$$\mathbb{E}\left[|\langle X_k, z\rangle|^2 |\langle X_k, w\rangle|^2\right] \leqslant s\mathbb{E}|\langle X_k, w\rangle|^2 = s\mathbb{E}w^* X_k X_k^* w = s w^* w = s.$$

Thus, $\mathbb{E}|f_{z,w}(X_k)|^2 \leqslant \sigma_k^2 = s$.

With these these two bounds, we can now use Bernstein's inequality for the supremum of the considered empirical process $K\delta_{\omega,s}$. We first use our argument bounding the expectation of $\delta_{\omega,s}$ to say that for any $\delta \in (0,1)$ we may choose some $K \gtrsim s\delta^{-2}\log(M)\log^3(s)$ such that $\mathbb{E}\delta_{\omega,s} \leqslant \delta/2$. We then consider

$$\begin{aligned}\mathbb{P}[\delta_{\omega,s} \geqslant \delta] &\leqslant \mathbb{P}[\delta_{\omega,s} \geqslant \mathbb{E}\delta_{\omega,s} + \delta/2] \\ &\leqslant \mathbb{P}[K\delta_{\omega,s} \geqslant \mathbb{E}K\delta_{\omega,s} + K\delta/2] \\ &\leqslant \exp\left(-\frac{(K\delta/2)^2/2}{sK + 2sK\mathbb{E}\delta_{\omega,s} + s(K\delta/2)/3}\right) \\ &\leqslant \exp\left(-\frac{(K\delta/2)^2/2}{sK + sK\delta + s(K\delta/2)/3}\right) \\ &\leqslant \exp\left(-\frac{K\delta^2}{Cs}\right) \leqslant \gamma,\end{aligned}$$

so long as $K \gtrsim s\delta^{-2}\log(1/\gamma)$. Thus, $\delta_{\omega,s} \leqslant \delta$ with probability at least $1-\gamma$ so long as

$$m \gtrsim s\delta^{-2}\max\left\{\log(N)\log^3(s), \log(1/\gamma)\right\},$$

as desired. □

2.7. Tying Together and Handling Infinite Index Sets

With Theorem 2.8 (normalized sampling matrix has ω-RIP with high probability), Theorem 2.4 (ω-RIP implies ω-NSP), and Theorem 2.3 (ω-NSP provides bounds for weighted ℓ_1-minimization) in hand, we can now prove our original approximation result which we recall below.

Theorem 2.2. *For a finite index set $|\Lambda| = M$ and weights $\omega_\nu \geqslant \|\phi_\nu\|_\infty$, if $s \geqslant 2\|\omega\|_\infty^2$ and we draw*

$$K \gtrsim s\log^3(s)\log(M)$$

i.i.d. samples $\{Z^{(k)}\}_{k=1}^K$ from the orthogonalization measure $\pi\,dz$, then with probability $1 - M^{-\log^3(s)}$, we can approximately recover u from the polluted samples $y = A\hat{u} + e$ with error satisfying $\|e\|_2 \leqslant \eta$ as the solution $\hat{u}^\sharp$ of

$$\underset{z\in\mathbb{C}^M}{\text{minimize}}\,\|z\|_{\omega,1} \text{ subject to } \|Az - y\|_2 \leqslant \eta,$$

in the sense that for $u^\sharp = \sum_{\nu\in\Lambda}\hat{u}_\nu^\sharp\phi_\nu$,

$$\|u - u^\sharp\|_\infty \leqslant \|u - u^\sharp\|_{\omega,1} \leqslant B_1\sigma_s(u)_{\omega,1} + B_2\eta\sqrt{\frac{s}{K}},$$

$$\|u - u^\sharp\|_2 \leqslant \frac{C_1}{\sqrt{s}}\sigma_s(u)_{\omega,1} + C_2\frac{\eta}{\sqrt{K}}.$$

Proof. First, take the threshold for probability of failure γ in Theorem 2.8 to be $\gamma = M^{-\log^3(3s)}$, so that

$$K \gtrsim s\log^3(s)\log(M)$$

satisfies the required number of measurements for $\frac{1}{\sqrt{K}}A$ to have $\delta_{\omega,3s} \leqslant \frac{1}{3}$ with probability exceeding $1 - \gamma = 1 - M^{-\log^3(3s)} \geqslant 1 - M^{-\log^3(s)}$. By Theorem 2.4, this ω-RIP constant implies the ω-NSP.

Starting with measurements $y = A\hat{u} + e$ with $\|e\|_2 \leqslant \eta$, we convert to the normalized sampling matrix by considering $\tilde{y} = \frac{1}{\sqrt{K}}y$ and $\tilde{e} = \frac{1}{\sqrt{K}}e$ satisfying $\|\tilde{e}\|_2 \leqslant \frac{\eta}{\sqrt{K}}$. Theorem 2.3 then implies that for $\hat{u}^\sharp$ solving the weighted ℓ_1 minimization program

$$\underset{z\in\mathbb{C}^M}{\text{minimize}}\,\|z\|_{\omega,1} \text{ subject to } \|\tilde{A}z - \tilde{y}\| \leqslant \frac{\eta}{\sqrt{K}} \iff \|Az - y\| \leqslant \eta,$$

the error in the recovered solution is bounded as

$$\|\hat{u} - \hat{u}^\sharp\|_{\omega,1} \leqslant B_1\sigma_s(\hat{u})_{\omega,1} + B_2\eta\sqrt{\frac{s}{K}}$$

$$\|\hat{u} - \hat{u}^\sharp\|_2 \leqslant \frac{C_1}{\sqrt{s}}\sigma_s(\hat{u})_{\omega,1} + C_2\frac{\eta}{\sqrt{K}}.$$

Since for $u^\sharp = \sum_{\nu\in\Lambda}\hat{u}_\nu^\sharp\phi_\nu$,

$$\|u - u^\sharp\|_\infty \leqslant \sum_{\nu\in\Lambda}|\hat{u}_\nu - \hat{u}_\nu^\sharp|\|\phi_\nu\|_\infty \leqslant \sum_{\nu\in\Lambda}|\hat{u}_\nu - \hat{u}_\nu^\sharp|\omega_\nu = \|\hat{u} - \hat{u}^\sharp\|_{\omega,1},$$

we obtain the desired L^∞ bound (2.2). On the other hand, since $\{\phi_\nu\}_{\nu\in\Lambda}$ is a gPC basis and is orthonormal with respect to π dz, Parseval's identity gives

$$\|u - u^\sharp\|_{L^2_\pi} = \|\hat{u} - \hat{u}^\sharp\|_2,$$

and we find the desired L^2 bound (2.3). □

2.7.1. Non-Uniform Results for Infinite Index Sets. We continue by using Theorem 2.2 to give a result for orthonormal systems over countably infinite index sets. We start by showing that with probability comparable to that with which the ω-RIP constant holds for a finitely indexed system, the truncation of u to a certain index set has small remainder.

Lemma 2.4. *Let Λ be countably infinite such that $\Lambda_0 = \{\nu \mid \omega_\nu^2 \leqslant s/2\}$ is finite and denote $|\Lambda_0| = M$, $\Lambda_R = \Lambda \setminus \Lambda_0$. For a fixed function $u = \sum_{\nu \in \Lambda} \hat{u}_\nu \phi_\nu$, take*

$$K \gtrsim s \log(M) \log^3(s)$$

i.i.d. samples $\{Z^{(k)}\}_{k=1}^K$ from π dz and form the ensemble of measurements $y_k = u\left(Z^{(k)}\right)$. With probability exceeding $1 - M^{-\log^3(s)}$

$$\left(\sum_{k=1}^{K} \left(u_{\Lambda_R}(Z^{(k)})\right)^2\right)^{1/2} \leqslant 2\sqrt{\frac{K}{s}} \|u_{\Lambda_R}\|_{\omega,1}. \tag{2.27}$$

Proof. We wish to bound the sum of random variables $u_k^2 := u_{\Lambda_R}(Z^{(k)})^2$ with high probability. This suggests that we use a good quality concentration inequality, namely Bernstein's inequality for bounded random variables, Corollary 2.1. We first calculate the expectation of u_k^2, so that we may center. By Parseval's identity,

$$\mathbb{E}u_k^2 = \sum_{\nu \in \Lambda_R} \hat{u}_\nu^2.$$

Thus, Bernstein's inequality will bound the probability that

$$\sum_{k=1}^{K} \left(u_k^2 - \sum_{\nu \in \Lambda_R} \hat{u}_\nu^2\right)$$

exceeds some bound $t \geqslant 0$, so long as we have bounds on the variance and L^∞-norm of each summands. Indeed, we calculate

$$\|u_k^2\|_\infty \leqslant \left(\sum_{\nu \in \Lambda_R} |\hat{u}_\nu| \|\phi_\nu(Z)\|_\infty\right)^2 \leqslant \|u_{\Lambda_R}\|_{\omega,1}^2,$$

and using the definition of Λ_0, we have

$$\sum_{\nu \in \Lambda_R} \hat{u}_\nu^2 \leqslant \frac{2}{s} \sum_{\nu \in \Lambda_R} \hat{u}_\nu^2 \omega_\nu^2 \leqslant \frac{2}{s} \left(\sum_{\nu \in \Lambda_R} |u_\nu| \omega_\nu\right)^2 \leqslant \frac{2}{s} \|u_{\Lambda_R}\|_{\omega,1}^2 \leqslant \|u_{\Lambda_R}\|_{\omega,1}^2. \tag{2.28}$$

Thus, $\left\|u_k^2 - \sum_{\nu \in \Lambda_R} \hat{u}_\nu^2\right\|_\infty \leqslant 2\|u_{\Lambda_R}\|_{\omega,1}^2$. Additionally, we bound the variance as

$$\mathbb{E}\left(u_k^2 - \mathbb{E}u_k^2\right)^2 \leqslant \mathbb{E}u_k^4 \leqslant \|u_{\Lambda_R}\|_{\omega,1}^2 \mathbb{E}u_k^2 \leqslant \frac{2}{s} \|u_{\Lambda_R}\|_{\omega,1}^4.$$

Bernstein's inequality then gives

$$\mathbb{P}\left[\sum_{k=1}^{K} \left(u_k^2 - \sum_{\nu \in \Lambda_R} \hat{u}_\nu^2\right) \geqslant t\right] \leqslant \exp\left(-\frac{t^2/2}{2K\|u_{\Lambda_R}\|_{\omega,1}^4/s + 2t\|u_{\Lambda_R}\|_{\omega,1}^2/3}\right).$$

Since our goal is to bound $\sum_{k=1}^{K} u_k^2$ by $\|u_{\Lambda_R}\|_{\omega,1}^2$, we choose $t = \frac{K}{s}\|u_{\Lambda_R}\|_{\omega,1}$ (where the $\frac{K}{s}$ balances out the terms in the exponential). Thus,

$$\mathbb{P}\left[\sum_{k=1}^{K}\left(u_k^2 - \sum_{\nu\in\Lambda_R}\hat{u}_\nu^2\right) \geqslant \frac{K}{s}\|u_{\Lambda_R}\|_{\omega,1}^2\right] \leqslant \exp\left(-C\frac{K}{s}\right),$$

and taking $K \gtrsim s\log(M)\log^3(s)$, we have

$$\mathbb{P}\left[\sum_{k=1}^{K}\left(u_k^2 - \sum_{\nu\in\Lambda_R}\hat{u}_\nu^2\right) \geqslant \frac{K}{s}\|u_{\Lambda_R}\|_{\omega,1}^2\right] \leqslant M^{-\log^3(s)}.$$

So with probability exceeding $1 - M^{-\log^3(s)}$,

$$\sum_{k=1}^{K} u_k^2 \leqslant K\sum_{\nu\in\Lambda_R}\hat{u}_\nu^2 + \frac{K}{s}\|u_{\Lambda_R}\|_{\omega,1}^2 \leqslant 3\frac{K}{s}\|u_{\Lambda_R}\|_{\omega,1}^2.$$

Taking square roots gives the desired bound with the specified probability. □

By approximating this truncation using Theorem 2.2, we obtain the following result.

Corollary 2.2. *For Λ countably infinite and $\Lambda_0 = \{\nu \mid \omega_\nu^2 \leqslant s/2\}$ finite with $|\Lambda_0| = M$, and for a fixed function $\sum_{\nu\in\Lambda}\hat{u}_\nu\phi_\nu$, take*

$$K \gtrsim s\log(M)\log^3(s)$$

i.i.d. samples $\{Z^{(k)}\}_{k=1}$ from π dz, and form the measurements $y_k = u(Z^{(k)})$. With probability exceeding $1 - 2M^{-\log^3(s)}$ the following holds. Letting $\hat{u}^\sharp$ be the solution of the weighted ℓ_1 minimization program

$$\underset{z\in\mathbb{C}^M}{\text{minimize}}\ \|z\|_{\omega,1} \ \textit{subject to}\ \|Az - y\|_2 \leqslant 2\sqrt{\frac{K}{s}}\|u - u_{\Lambda_0}\|_{\omega,1}.$$

Taking $u^\sharp = \sum_{\nu\in\Lambda}\hat{u}_\nu^\sharp\phi_\nu$,

$$\|u - u^\sharp\|_\infty \leqslant B_1\sigma_{s/2}(u)_{\omega,1}$$
$$\|u - u^\sharp\|_2 \leqslant \frac{C_1}{\sqrt{s}}\sigma_{s/2}(u)_{\omega,1}.$$

Proof. We use Theorem 2.2 on the truncated function u_{Λ_0}, where the error in each sample of this truncation is $u(Z^{(k)}) - u_{\Lambda_0}(Z^{(k)}) = u_{\Lambda_R}(Z^{(k)})$. By Lemma 2.4, we know that with the specified number of measurements, the ℓ_2 norm of this error does not exceed $\eta = 2\sqrt{\frac{K}{s}}\|u - u_{\Lambda_0}\|_{\omega,1}$ with probability bounded by $M^{-\log^3(s)}$. Additionally, Theorem 2.2 says that a draw of the specified number of measurements measurements gives that solving the stated minimization program provides the bounds

$$\|u_{\Lambda_0} - u^\sharp\|_\infty \leqslant B\left[\sigma_s(u_{\Lambda_0})_{\omega,1} + \|u - u_{\Lambda_0}\|_{\omega,1}\right]$$
$$\|u_{\Lambda_0} - u^\sharp\|_{L^2} \leqslant \frac{C}{\sqrt{s}}\left[\sigma_s(u_{\Lambda_0})_{\omega,1} + \|u - u_{\Lambda_0}\|_{\omega,1}\right],$$

with probability of failure bounded by $M^{-\log^3(s)}$. Thus, the union bound tells us that the probability of either of these conditions failing is less than $2M^{-\log^3(s)}$ giving the corollary's probability estimate. In the proof of Lemma 2.4, we also found

$$\|u - u_{\Lambda_0}\|_\infty \leqslant \|u - u_{\Lambda_0}\|_{\omega,1}$$
$$\|u - u_{\Lambda_0}\|_2 \leqslant \frac{2}{\sqrt{s}}\|u - u_{\Lambda_0}\|_{\omega,1},$$

and so by the triangle inequality,

$$\left\|u - u^\sharp\right\|_\infty \leqslant B\left[\sigma_s(u_{\Lambda_0})_{\omega,1} + \|u - u_{\Lambda_0}\|_{\omega,1}\right]$$
$$\left\|u - u^\sharp\right\|_{L^2} \leqslant \frac{C}{\sqrt{s}}\left[\sigma_s(u_{\Lambda_0})_{\omega,1} + \|u - u_{\Lambda_0}\|_{\omega,1}\right].$$

The last step is to relate $\sigma_s(u_{\Lambda_0})_{\omega,1} + \|u - u_{\Lambda_0}\|_{\omega,1}$ to $\sigma_{s/2}(u)_{\omega,1}$. But since $\omega_\nu^2 > s/2$ for all $\nu \in \Lambda_R$, the support of the weighted best $s/2$-term estimate of u cannot intersect Λ_R. Thus, the weighted best $s/2$-term estimate to u_{Λ_0} is the same as that of u. Since this is a stricter sparsity condition, $\sigma_s(u_{\Lambda_0})_{\omega,1} \leqslant \sigma_{s/2}(u_{\Lambda_0})_{\omega,1}$. Since this best $s/2$-term estimate has support disjoint with u_{Λ_R}, its weighted ℓ_1 difference with u is the same as its difference with u_{Λ_0} summed with the remainder u_{Λ_R}. So $\sigma_s(u_{\Lambda_0})_{\omega,1} + \|u - u_{\Lambda_0}\|_{\omega,1} \leqslant \sigma_{s/2}(u)_{\omega,1}$, giving the desired bound. □

2.7.2. Uniform Results for Infinite Index Sets. By using Lemma 2.4 to bound the error in using a truncated version of u for sparse recovery in Corollary 2.2, we have derived a nonuniform result. However, we can take another approach, which instead of formulating $\left\|\left(u_{\Lambda_R}(Z^{(k)})\right)_{k=1}^K\right\|_{\ell_2}$ in terms of $\sqrt{K/s}\|u_{\Lambda_R}\|_{\omega,1}$, we use

$$\left\|\left(u_{\Lambda_R}(Z^{(k)})\right)_{k=1}^K\right\|_{\ell_2} \leqslant \sqrt{K}\|u_{\Lambda_R}\|_\infty \leqslant \sqrt{K}\sum_{\nu\in\Lambda_R}|\hat{u}_\nu|\omega_\nu \tag{2.29}$$

We then need some other way to balance this with the $\sigma_s(u_{\Lambda_0})_{\omega,1}$ term appearing in the error bounds of Theorem 2.2 after we use it on the truncation. The goal is to take a common upper bound using the Stechkin estimate. From Theorem 1.5, we have

$$\sigma_s(u_{\Lambda_0})_{\omega,1} \leqslant (s - \|\omega\|_\infty^2)^{1-1/p}\|u_{\Lambda_0}\|_{\omega,p}$$

for any $p < 1$, so long as $s > \|\omega\|_\infty^2$. Additionally, when $s/2 \geqslant \|\omega\|_\infty^2$, we can remove the dependence of ω in this bound by noting that $s/2 \leqslant s - \|\omega\|_\infty^2$, and so

$$\sigma_s(u_{\Lambda_0})_{\omega,1} \leqslant 2^{1/p-1}s^{1-1/p}\|u_{\Lambda_0}\|_{\omega,p}. \tag{2.30}$$

In order to balance this bound on $\sigma_s(u_{\Lambda_0})_{\omega,1}$ by $\|u_{\Lambda_0}\|_{\omega,p}$ with the ℓ_2 bound on the error (2.29) , we will introduce a set of auxiliary weights which allow us more control over $\|u_{\Lambda_R}\|_{\omega,1}$ and still allow us to make use of the property that $s/2 \geqslant \|\omega\|_\infty^2$ (without which we do not have the previously shown recovery bounds). We summarize this auxiliary weighting process in the following lemma, which we use to prove the final theorem on the topic of pure function approximation using compressive sensing.

Lemma 2.5. *Fix* $p \in (0,1)$. *Let* ξ *be a set of weights and* Λ *be countably infinite such that for* $\alpha = 2/p - 1$, $\Lambda_0^{(s,p)} = \{\nu \in \Lambda \mid \omega_\nu\xi_\nu^{-\alpha} \geqslant s^{1/2-1/p}\}$ *and* $\Lambda_R = \Lambda \setminus \Lambda_0$. *Then*

$$\left\|\left(u_{\Lambda_R}(Z^{(k)})\right)_{k=1}^K\right\|_{\ell_2} \leqslant \sqrt{\frac{K}{s}}s^{1-1/p}\|u_{\Lambda_R}\|_{\xi,p}.$$

Proof. For the given weights, estimate (2.29) gives

$$\left\|\left(u_{\Lambda_R}(Z^{(k)})\right)_{k=1}^{K}\right\|_{\ell_2} \leqslant \sqrt{K}\sup_{\eta\in\Lambda_R}(\omega_\eta\xi_\eta^{-\alpha})\sum_{\nu\in\Lambda_R}|\hat{u}_\nu|\xi_\nu^{\alpha} = \sqrt{K}\sup_{\eta\in\Lambda_R}(\omega_\eta\xi_\eta^{-\alpha})\|u_{\Lambda_R}\|_{\xi^\alpha,1}.$$

We now provide an estimate which allows us to switch from $\|u_{\Lambda_R}\|_{\xi^\alpha,1}$ to $\|u_{\Lambda_R}\|_{\xi,p}$. Indeed,

$$\begin{aligned}\|u_{\Lambda_R}\|_{\xi^\alpha,1} &= \sum_{\nu\in\Lambda_R}|\hat{u}_\nu|\xi_\nu^{\alpha}\\ &\leqslant \sup_{\eta\in\Lambda_R}|\hat{u}_\eta|^{1-p}\xi_\eta^{\alpha(1-p)}\sum_{\nu\in\Lambda_R}|\hat{u}_\nu|^{p}\xi_\nu^{\alpha p}.\end{aligned}$$

By virtue of the choice that $\alpha p = 2-p$, we find

$$\begin{aligned}\|u_{\Lambda_R}\|_{\xi^\alpha,1} &\leqslant \sup_{\eta\in\Lambda_R}|\hat{u}_\eta|^{1-p}\xi_\eta^{\alpha(1-p)}\|u_{\Lambda_R}\|_{\xi,p}^{p}\\ &\leqslant \left(\sum_{\eta\in\Lambda_R}|\hat{u}_\eta|^{p}\xi_\eta^{\alpha p}\right)^{(1-p)/p}\|u_{\Lambda_R}\|_{\xi,p}^{p} \qquad (2.31)\\ &= \|u_{\Lambda_R}\|_{\xi,p}^{1-p}\|u_{\Lambda_R}\|_{\xi,p}^{p} = \|u_{\Lambda_R}\|_{\xi,p}.\end{aligned}$$

Combining with the previous bound gives

$$\left\|\left(u_{\Lambda_R}(Z^{(k)})\right)_{k=1}^{K}\right\|_{\ell_2} \leqslant \sqrt{K}\sup_{\eta\in\Lambda_R}(\omega_\eta\xi_\eta^{-\alpha})\|u_{\Lambda_R}\|_{\xi,p}.$$

By our choice of Λ_0, for all $\nu \in \Lambda_R$, $\omega_\eta\xi_\eta^{-\alpha} \leqslant s^{1/2-1/p}$, giving the desired bound. □

Corollary 2.3. *Fix* $p \in (0,1)$. *Let* ξ *be a set of weights satisfying* $\xi_\nu \geqslant \sqrt{2}\omega_\nu^{2/(2-p)}$, *and for* $\alpha = 2/p-1$, Λ *countably infinite and* $\Lambda_0^{(s,p)} = \left\{\nu\in\Lambda \mid \omega_\nu\xi_\nu^{-\alpha} \geqslant s^{1/2-1/p}\right\}$ *finite with* $|\Lambda_0^{(s,p)}| = M^{(s,p)}$, *take*

$$K \gtrsim s\log\left(M^{(s,p)}\right)\log^3(s)$$

i.i.d. samples $\left\{Z^{(k)}\right\}_{k=1}^{K}$ *from* $\pi\,dz$. *With probability exceeding* $1-(M^{(s,p)})^{-\log^3(s)}$, *for any* $u \in \ell_{\xi,p}$, *we have the following weighted* ℓ_1 *recovery estimates. If we take the samples* $y_k = u(Z^{(k)})$, *and let* $\hat{u}^\sharp$ *be the solution of the weighted* ℓ_1 *minimization program*

$$\underset{z\in\mathbb{C}^{M^{(s,p)}}}{\text{minimize}}\ \|z\|_{\omega,1}\ \textit{subject to}\ \|Az-y\|_2 \leqslant \sqrt{\frac{K}{s}}s^{1-1/p}\|u_{\Lambda_R}\|_{\xi,p}$$

then for $u^\sharp = \sum_{\nu\in\Lambda}\hat{u}_\nu^\sharp\phi_\nu$,

$$\begin{aligned}\left\|u-u^\sharp\right\|_\infty &\leqslant B(p)s^{1-1/p}\|u\|_{\xi,p}\\ \left\|u-u^\sharp\right\|_{L^2} &\leqslant C(p)s^{1/2-1/p}\|u\|_{\xi,p}.\end{aligned}$$

Proof. Since $\xi_\nu \geqslant \sqrt{2}\omega_\nu^{2/(2-p)} > \omega_\nu$, on $\Lambda_0^{(s,p)}$,

$$\begin{aligned}s^{1/2-1/p} &\leqslant \omega_\nu\xi_\nu^{-\alpha}\\ &\leqslant \omega_\nu\left(\sqrt{2}\omega_\nu^{2/(2-p)}\right)^{1-2/p}\\ &= 2^{1/2-1/p}\omega_\nu^{1-2/p}\\ &= \left(2\omega_\nu^2\right)^{1/2-1/p}.\end{aligned}$$

Since $1/2 - 1/p < 0$, we therefore have that $s \geqslant 2\omega_\nu^2$ for all $\nu \in \Lambda_0^{(s,p)}$. Applying Theorem 2.2 on $u_{\Lambda_0^{(s,p)}}$ with measurement error vector

$$u(Z^{(k)}) - u_{\Lambda_0^{(s,p)}}(Z^{(k)}) = u_{\Lambda_R}(Z^{(k)}),$$

which, by Lemma 2.5 has ℓ_2 norm bounded by $s^{1-1/p}\sqrt{K/s}\|u_{\Lambda_R}\|_{\xi,p}$ gives the following error bounds for the specified weighted ℓ_1 minimization program:

$$\left\|u_{\Lambda_0^{(s,p)}} - u^\sharp\right\|_\infty \leqslant B\left[\sigma_s(u_{\Lambda_0^{(s,p)}})_{\omega,1} + s^{1-1/p}\|u_{\Lambda_R}\|_{\xi,p}\right]$$
$$\left\|u_{\Lambda_0^{(s,p)}} - u^\sharp\right\|_{L^2} \leqslant \frac{C}{\sqrt{s}}\left[\sigma_s(u_{\Lambda_0^{(s,p)}})_{\omega,1} + s^{1-1/p}\|u_{\Lambda_R}\|_{\xi,p}\right].$$

But from (2.30), we have

$$\left\|u_{\Lambda_0^{(s,p)}} - u^\sharp\right\|_\infty \leqslant B(p)s^{1-1/p}\left[\left\|u_{\Lambda_0^{(s,p)}}\right\|_{\omega,p} + \|u_{\Lambda_R}\|_{\xi,p}\right]$$
$$\left\|u_{\Lambda_0^{(s,p)}} - u^\sharp\right\|_{L^2} \leqslant C(p)s^{1/2-1/p}\left[\left\|u_{\Lambda_0^{(s,p)}}\right\|_{\omega,p} + \|u_{\Lambda_R}\|_{\xi,p}\right],$$

Additionally, since $\sqrt{2}\omega_\nu^{2/(2-p)} \leqslant \xi_n$,

$$\omega_\nu \leqslant \xi_\nu^{(2-p)/2} \leqslant \xi_\nu.$$

Then

$$\left\|u_{\Lambda_0^{(s,p)}}\right\|_{\omega,p} + \|u_{\Lambda_R}\|_{\xi,p} \leqslant \left\|u_{\Lambda_0^{(s,p)}}\right\|_{\xi,p} + \|u_{\Lambda_R}\|_{\xi,p} \leqslant 2\|u\|_{\xi,p}.$$

The proof of Lemma 2.5 and (2.29) additionally give

$$\left\|u - u_{\Lambda_0^{(s,p)}}\right\|_{L^2} \leqslant \left\|u - u_{\Lambda_0^{(s,p)}}\right\|_\infty = \|u_{\Lambda_R}\|_\infty \leqslant s^{1/2-1/p}\|u_{\Lambda_R}\|_{\xi,p} \leqslant s^{1-1/p}\|u\|_{\xi,p},$$

to combine with the above recovery bounds to give the total error by the triangle inequality. □

CHAPTER 3

Compressive Sensing for Solving High-Dimensional PDE

In this chapter, we will use the results from the Chapter 2 to approximate solutions to high-dimensional parametric PDE discussed in Chapter 1. We will proceed through the method found in [23], which makes heavy use of the compressive sensing techniques in [24]. But first, we recall the problem and discuss general assumptions for well-posedness.

3.1. Problem and Assumptions

We consider the special case of finding pointwise weak solutions of the standard parameterized UQ problem 1.4 where the operator $\mathcal{L}$ has affine dependence on a (now possibly infinite) parameter sequence Z as in (1.52), that is, for separable reflexive Banach spaces $\mathcal{X}$ and $\mathcal{Y}$, $Z \in \Gamma = [-1,1]^{\mathbb{N}_+}$, and

$$\mathcal{L}(Z) = \mathcal{L}_0 + \sum_{n=1}^{\infty} Z_n \mathcal{L}_n : \Gamma \to \mathcal{L}(\mathcal{X}, \mathcal{Y}^*),$$

we wish to find some $u : \Gamma \to \mathcal{X}$ such that for $f : \Gamma \to \mathcal{Y}^*$

$$\mathfrak{L}(u, w; Z) := {}_{\mathcal{Y}^*}\langle \mathcal{L}(Z)u(Z), w\rangle_{\mathcal{Y}} = {}_{\mathcal{Y}^*}\langle f(Z), w\rangle_{\mathcal{Y}} \quad \text{for all } w \in \mathcal{Y} \text{ and all } Z \in \Gamma. \tag{3.1}$$

This solution can also be expressed as saying that $\mathcal{L}(Z)u = f(Z)$ as elements of $\mathcal{Y}^*$ for all $Z \in \Gamma$. Since we will be making use of the parametric nature of the equation, we also define the bilinear form

$$\mathfrak{L}_n(u, w) := {}_{\mathcal{Y}^*}\langle \mathcal{L}_n u, w\rangle_{\mathcal{Y}} \quad \text{for all } n \in \mathbb{N}_0.$$

Recall that by the Karhunen-Loève expansion, the imposition that $\mathcal{L}$ depends on Z affinely is a mild assumption.

3.1.1. Well-Posedness of True Solution. As discussed in Section 1.3.2, when $\mathcal{X} = \mathcal{Y}$, the Lax-Milgram theorem, Theorem 1.2, gives conditions for well-posedness of the solution. However, when $\mathcal{X} \neq \mathcal{Y}$, we require more stringent versions of continuity and coercivity in the form of inf-sup conditions. These conditions are summarized in the assumption below on the mean operator, $\mathcal{L}_0$, as well as conditions to keep the fluctuation operators, $\mathcal{L}_n$ for $n \in \mathbb{N}_+$, well-behaved.

ASSUMPTION 3.1.

(1) $\mathcal{L}_0$ satisfies the inf-sup condition that there exists some $\mu_0 > 0$ with

$$\inf_{v \in \mathcal{X}\setminus\{0\}} \sup_{w \in \mathcal{Y}\setminus\{0\}} \frac{\mathfrak{L}_0(v, w)}{\|v\|_{\mathcal{X}} \|w\|_{\mathcal{Y}}} \geqslant \mu_0, \quad \inf_{\mathcal{Y}\setminus\{0\}} \sup_{v \in \mathcal{X}\setminus\{0\}} \frac{\mathfrak{L}_0(v, w)}{\|v\|_{\mathcal{X}} \|w\|_{\mathcal{Y}}} \geqslant \mu_0. \tag{3.2}$$

In particular, this implies that $\mathcal{L}_0$ is boundedly invertible (see remark below).

(2) There exists a constant $0 < \kappa < 1$ such that

$$\sum_{n \in \mathbb{N}_+} \beta_{0,n} \leqslant \kappa, \quad \textit{where } \beta_{0,n} := \left\|\mathcal{L}_0^{-1} \mathcal{L}_n\right\|_{\mathcal{L}(\mathcal{X},\mathcal{X})}, \textit{ for } n \in \mathbb{N}_+.$$

Remark 3.1. *Notice that bounded invertibility is the exact consequence of the Lax-Milgram theorem. Thus, the mean operator satisfying* (3.2) *resulting in bounded invertibility is a generalized version of the Lax-Milgram theorem. This is summarized below in Theorem 3.1. Just as for the Lax-Milgram theorem, we consider the proof outside the scope of these notes. Interested readers can see* [**21**, Section 2.2] *and references therein for a discussion of generalizations of coercivity and the Lax-Milgram theorem to non-symmetric bilinear forms.*

Theorem 3.1 ([**25**], Proposition 1, Banach-Nečas-Babuska). *Any bounded, linear operator* $\mathcal{L}_0 \in \mathcal{L}(\mathcal{X}, \mathcal{Y}^*)$ *is boundedly invertible if and only if* (3.2) *is satisfied. In particular,* (3.2) *implies that for every* $f \in \mathcal{Y}^*$*, the dual equality*

$$\mathfrak{L}_0(u, v) = {}_{\mathcal{Y}^*}\langle f, v\rangle_{\mathcal{Y}} \textit{ for all } v \in \mathcal{Y}$$

is realized with a unique $u \in \mathcal{X}$ *satisfying the stability estimate*

$$\|u\|_{\mathcal{X}} = \left\|\mathcal{L}_0^{-1} f\right\|_{\mathcal{X}} \leqslant \frac{1}{\mu_0}\|f\|_{\mathcal{Y}^*}.$$

With Assumption 3.1 in hand, we can prove an extension to the Banach-Nečas-Babuska theorem to show bounded invertibility of the *entire* affine operator, not just the mean operator.

Proposition 3.1 ([**25**], Theorem 1). *Under Assumption 3.1, for every* $Z \in \Gamma$*,* $\mathcal{L}(Z)$ *is boundedly invertible which is to say that that for every* $f \in \mathcal{Y}^*$ *and* $Z \in \Gamma$*, a unique* $u : \Gamma \to \mathcal{X}$ *exists satisfying* (3.1) *and the parametrically uniform stability estimate*

$$\sup_{Z\in\Gamma} \|u(Z)\|_{\mathcal{X}} \leqslant \frac{1}{\mu}\|f\|_{\mathcal{Y}^*},$$

where $\mu = (1-\kappa)\mu_0$*.*

Proof. Since Assumption 3.1 implies that $\mathcal{L}_0$ is boundedly invertible, we write

$$\mathcal{L}(Z) = \mathcal{L}_0 \left(I + \sum_{n\in\mathbb{N}_+} \mathcal{L}_0^{-1} Z_n \mathcal{L}_n \right).$$

Thus, it suffices to show that the second term is boundedly invertible. We prove this with a Neumann series argument. First, note that

$$\left\| - \sum_{n\in\mathbb{N}_+} \mathcal{L}_0^{-1} Z_n \mathcal{L}_n \right\| \leqslant \sum_{n\in\mathbb{N}_+} |Z_n| \|\mathcal{L}_0^{-1}\mathcal{L}_n\| \leqslant \sum_{n\in\mathbb{N}_+} \beta_{0,n} \leqslant \kappa < 1,$$

where we have used that $Z_n \in \Gamma_n = [-1, 1]$. Thus, we may write the second term as $I - A$ where for some fixed $Z \in \Gamma$, $A = -\sum_{n\in\mathbb{N}_+} Z_n \mathcal{L}_0^{-1}\mathcal{L}_n$ which has norm bounded by $\kappa < 1$.

We now show that $I - A$ is invertible and bound the norm of the inverse. For injectivity, we consider $v \in \mathcal{X} \setminus \{0\}$, and calculate

$$\|(I - A)v\|_{\mathcal{X}} \geqslant \|v\|_{\mathcal{X}} - \|Av\|_{\mathcal{X}} > (1-\kappa)\|v\|_{\mathcal{X}} > 0.$$

Thus, the nullspace $\mathcal{N}(I - A) = \{0\}$, and $I - A$ is injective. For surjectivity, for any $u \in \mathcal{X}$, we take

$$v = \sum_{k=0}^{\infty} A^k u,$$

which exists by a Cauchy sequence argument on the partial sums and the fact that

$$\left\|\sum_{k=0}^{\infty} A^k\right\| \leqslant \sum_{k=0}^{\infty} \|A^k\| \leqslant \sum_{k=0}^{\infty} \|A\|^k \leqslant \sum_{k=0}^{\infty} \kappa^k = \frac{1}{1-\kappa}. \tag{3.3}$$

Now applying $I - A$, we have

$$(I - A)v = \sum_{k=0}^{\infty} A^k u - \sum_{k=1}^{\infty} A^k u = A^0 u = u.$$

Thus, $I - A$ is surjective and invertible with $(I - A)^{-1} = \sum_{k=0}^{\infty} A^k$ which, by (3.3), has norm bounded by $(1-\kappa)^{-1}$.

Combining the bounded invertibility of $\mathcal{L}_0$ and $I - A$, we have $\mathcal{L}(Z) = \mathcal{L}_0(I - A)$ is also boundedly invertible, with

$$\left\|\mathcal{L}(Z)^{-1} f\right\|_{\mathcal{X}} \leqslant \frac{1}{1-\kappa}\left\|\mathcal{L}_0^{-1} f\right\|_{\mathcal{X}} \leqslant \frac{1}{(1-\kappa)\mu_0}\|f\|_{\mathcal{X}^*},$$

by (3.3) and the stability estimate for $\mathcal{L}_0^{-1}$ given in Theorem 3.1 when $\mathcal{L}_0$ satisfies Assumption 3.1.

□

Example 3.1. *We consider the case of the parametric diffusion equation with diffusion coefficient expanded using the Karhunen-Loève theorem, that is, when*

$$\mathcal{L}(Z)u = -D \cdot (A(x, Z)Du) = -D \cdot (\hat{A}_0(x)Du) + \sum_{n \in \mathbb{N}_+} Z_n(-D \cdot (\hat{A}_n(x)Du)).$$

Since $\mathcal{X} = \mathcal{Y} = H_0^1(\Omega)$, Assumption 3.1.1 reduces to coercivity so that the mean operator is boundedly invertible by the Lax-Milgram theorem, Theorem 1.2. So we assume the mean operator is coercive with parameter α, which is equivalent to $\hat{A}_0(x)$ being uniformly elliptic with parameter α. In order to rephrase Assumption 3.1.2, we calculate $\|\mathcal{L}_n\|$. For any $u \in H_0^1(\Omega)$ with $\|u\|_{H_0^1(\Omega)} = \|Du\|_{L^2(\Omega)} = 1$, we find

$$\begin{aligned}
\|\mathcal{L}_n u\|_{H^{-1}(\Omega)} &= \sup_{\|Dv\|_{L^2(\Omega)}=1} \int_\Omega (Dv)^T \hat{A}_n(x) Du \\
&\leqslant \sup_{\|Dv\|_{L^2(\Omega)}=1} \int_\Omega \|Dv\|_2 \left\|\hat{A}_n(x) Du\right\|_2 \\
&\leqslant \sup_{\|Dv\|_{L^2(\Omega)}=1} \|Dv\|_{L^2(\Omega)} \sqrt{\int_\Omega \left\|\hat{A}_n(x)\right\|^2 \|Du\|_2^2} \\
&\leqslant \sup_{x\in\Omega} \left\|\hat{A}_n(x)\right\| \|u\|_{H_0^1(\Omega)} \\
&= \sup_{x\in\Omega} \left\|\hat{A}_n(x)\right\|.
\end{aligned}$$

Thus, we require

$$\sum_{n\in\mathbb{N}_+} \left\|\mathcal{L}_0^{-1}\mathcal{L}_n\right\| \leqslant \sum_{n\in\mathbb{N}_+} \frac{\sup_{x\in\Omega}\left\|\hat{A}_n(x)\right\|}{\alpha} \leqslant \kappa.$$

Thus, the parametric stationary diffusion equation satisfies Assumption 3.1 when the $\hat{A}_0(x)$ is uniformly elliptic with constant α, and

$$\sum_{n\in\mathbb{N}_+} \sup_{x\in\Omega} \left\|\hat{A}_n(x)\right\| \leqslant \alpha\kappa,$$

for some $\kappa < 1$. *By Proposition 3.1, we then know that* $\mathcal{L}(Z)$ *is boundedly invertible uniformly over* Γ *with the stability estimate*

$$\|u\|_{H_0^1(\Omega)} = \left\|\mathcal{L}^{-1}f\right\|_{H_0^1(\Omega)} \leqslant \frac{1}{(1-\kappa)\alpha}\|f\|_{H^{-1}(D)}.$$

Notice as well that these two conditions are stronger than having the original affine operator $\mathcal{L}$ *be uniformly elliptic with ellipticity parameter* $\alpha(1-\kappa)$ *which provides the same well-posedness result by the Lax-Milgram theorem. Indeed, we check that for any* $\nu \in \mathbb{S}^{\ell-1}$,

$$\begin{aligned}\nu^T\left(\hat{A}_0(x) + \sum_{n\in\mathbb{N}_+}\hat{A}_n(x)\right)\nu &\geqslant \nu^T\hat{A}_0(x)\nu - \|\nu\|_2^2 \sup_{x\in\Omega}\left\|\sum_{n\in\mathbb{N}_+}\hat{A}_n(x)\right\| \\ &\geqslant \alpha - \alpha\kappa = \alpha(1-\kappa).\end{aligned}$$

The benefit of Assumption 3.1 and the Banach-Nečas-Babuska theorem is that it allows for analysis of a wider class of problems including saddle point problems where $\mathcal{X} \neq \mathcal{Y}$.

3.1.2. Well Posedness of Discrete Solution. As in the sampling methods considered in Chapter 1, we will sample values of $u(Z)$ at different parameter instances, that is, we create an ensemble of solutions to the deterministic problem fixed at certain parameter values. However, we will not be using exact solutions, but rather approximations. The method considered in [23] does this by way of Petrov-Galerkin discretization. Since we will only need the discrete solution to approximate the true solution up to some desired tolerance and any sampling methods can make use of the same solvers with the same rates of convergence, just as in Section 1.8.3.2, we defer to the finite element method theory and simply make the following assumptions

Assumption 3.2. *For every* $f \in \mathcal{Y}^*$ *and* $Z \in \Gamma$, *we consider the discrete, truncated version of* (3.1). *That is, for some finite dimensional subspaces* $\mathcal{X}_h \subset \mathcal{X}$, $\mathcal{Y}_h \subset \mathcal{Y}$ *(considered as finite element spaces with mesh parameter* h*), and dimension truncated affine operator*

$$\mathcal{L}^{(N)}(Z) = \mathcal{L}_0 + \sum_{n=1}^{N} Z_n\mathcal{L}_n, \quad \mathfrak{L}^{(N)}(u,\nu) = {}_{\mathcal{Y}^*}\langle \mathcal{L}^{(N)}u, \nu\rangle_{\mathcal{Y}},$$

we assume that the Petrov-Galerkin method produces a solution $u_{N,h} \in \mathcal{X}_h$ *satisfying*

$$\mathfrak{L}^{(N)}(u_{N,h},\nu) = {}_{\mathcal{Y}^*}\langle f,\nu\rangle_{\mathcal{Y}} \textit{ for all } \nu \in \mathcal{Y}_h.$$

Additionally, we assume that (for $u_h \in \mathcal{X}_h$ *solving the untruncated problem)* $\|u - u_h\|_{\mathcal{X}} = O(h^t)$ *with* t *and the implicit constant depending on the problem data and affine operator. Finally, we account for the dimension truncation when we apply a functional* $G \in \mathcal{X}^*$, *to the discrete solution, and assume that* $|G(u_h) - G(u_{N,h})| = O(N^{-p_0})$ *where* $p_0 > 0$ *depends on the summability of the vector of operator norms* β_0, *and the implicit constant depends on the affine data, the norm of the functional,* p_0 *the affine operator. Altogether, we have*

$$G(u_{N,h}) \to G(u) \textit{ as } N \to \infty, h \to 0.$$

Remark 3.2. *Note that when we make the finite dimensional noise assumption as in Section 1.8.3.2, we need not account for any error in the dimension truncation.*

3.1.3. Chebyshev Expansion Over Infinite Dimensional Parameter Domain. Since the recovery results from Chapter 2 are dimension independent, we will in general not make the finite dimensional noise assumption, which means that we must handle an infinite dimensional parameter domain $\Gamma = [-1,1]^{\mathbb{N}_+}$ and infinite dimensional gPC index sets. In order for a gPC basis to be well-defined we consider only indices with finite support, defining $\Lambda = \{\nu \in \mathbb{N}_0^{\mathbb{N}_+} \mid \|\nu\|_0 < \infty\}$.

In [23], the only gPC basis considered is the one of Chebyshev polynomials, as they will be shown to have L^∞ norms which play nicely with the necessary structure of the weights with respect to which the gPC coefficients of the solution to the affine parametric equation will be weighted summable (a requirement to use the compressive sensing results from Chapter 2). Additionally, prior work has shown unweighted summability of Chebyshev coefficients of solutions to affine parametric operator equations such as these, linking the ideas of analyticity and summability of the gPC expansion as hinted at previously. We make explicit the definition of the Chebyshev polynomials used as well as their orthogonalization measure.

Definition 3.1. *The one-dimensional order* j *Chebyshev polynomial is defined as*

$$\phi_j(Z) := \sqrt{2}\cos(j\arccos(Z)),$$

for all $j \in \mathbb{N}_+$ *with* $\phi_0(Z) \equiv 1$*, and the one-dimensional Chebyshev probability measure on* $[-1,1]$ *is*

$$\pi_n(z_n)\,dz_n = \frac{1}{\pi\sqrt{1-z_n^2}}\,dz_n.$$

For any $\nu \in \Lambda$*, we define the tensorized Chebyshev polynomial*

$$\phi_\nu = \prod_{n\in\mathbb{N}_+} \phi_{\nu_n}(Z_n) \tag{3.4}$$

and the tensorized Chebyshev measure

$$\pi(z)\,dz = \bigotimes_{n\in\mathbb{N}_+} \pi_n(z_n)\,dz_n = \bigotimes_{n\in\mathbb{N}_+} \frac{1}{\pi\sqrt{1-z_n^2}}\,dz_n.$$

Remark 3.3. *The tensorized Chebyshev polynomials are well-defined since the product in* (3.4) *has only finitely many factors by virtue of the fact that* $\|\nu\|_0 < \infty$*. A simple change of variables* $z_n = \cos(\theta)$ *shows that one-dimensional Chebyshev polynomials are orthonormal with respect to Chebyshev measure, and as a consequence, the tensorized Chebyshev polynomials are orthonormal with respect to the tensorized Chebyshev measure since the one-dimensional measure is a probability measure. Finally, note also that*

$$\|\phi_\nu\|_\infty = 2^{\|\nu\|_0/2}, \tag{3.5}$$

where equality is attained at the point $Z = 1 \in [-1,1]^{\mathbb{N}_+}$ *since a univariate Chebyshev polynomial of any order takes on its maximum value of* $\sqrt{2}$ *at* $Z_n = 1$.

Assumption 3.3. *For the rest of this chapter, we assume that the solution to the parametric operator equation* (3.1) $u : Z \to \mathcal{X} \in L^2_\pi(\Gamma)$ *has a gPC expansion*

$$u(Z) = \sum_{\nu\in\Lambda} \hat{u}_\nu \phi_\nu(Z)$$

in Chebyshev polynomials $\{\phi_\nu\}_{\nu\in\Lambda}$ *for coefficients* $\hat{u}_\nu \in \mathcal{X}$ *for all* $\nu \in \Lambda$.

The remainder of this chapter will focus on showing that the sequence $(\|\hat{u}_\nu\|_{\mathcal{X}})_{\nu \in \Lambda}$ is in some weighted ℓ_p space, allowing us to recover u (or rather, functionals of u) with the techniques in Chapter 2. Additionally, after determining the necessary weights, we show bounds on the size M of the truncated finite index set (and therefore a bound on the necessary number of deterministic PDE solves) $\Lambda_0 = \{\nu \in \Lambda \mid \omega_\nu^2 \leqslant s/2\}$.

3.2. Weighted Summability of the Chebyshev Coefficients

Our goal in this section is to show that $u \in S_{\omega,p}$ for some weight sequence ω and $0 < p < 1$. We will begin by first showing exponential decay of the Chebyshev coefficients of the solution in terms of its domain of analyticity. The argument follows closely the sketch of the proof of Lemma 1.1, and we will in fact prove the sketchy bound on the coefficients of the Fourier series using some classical techniques from approximation theory. Before this however, we clarify the assumption of analyticity in the parameter domain made on the solution in Section 1.8.3.1.

Proposition 3.2. *Under Assumption 3.1, for any $f \in \mathcal{Y}^*$, consider the solution to the parametric operator equation $u(Z) = \mathcal{L}^{-1}(Z)f \in \mathcal{X}$. For any $n \in \mathbb{N}_+$, if we fix $Z_n^* \in \Gamma_n^* = \prod_{j \in \mathbb{N}_+ \setminus \{n\}} \Gamma_j$, then $u(Z_n) := u(Z_n, Z_n^*)$ admits an analytic extension to the disc in the complex plane of radius $(1 - \kappa + \beta_{0,n})\beta_{0,n}^{-1}$ containing $\Gamma_n = [-1, 1]$.*

Proof. By Proposition 3.1, we know that u exists. We employ a similar argument to the one in the proof of Proposition 3.1, but single out the Z_n dimension. Indeed, by the invertibility of $\mathcal{L}_0$,

$$\begin{aligned}
\mathcal{L} &= \mathcal{L}_0 \left(I + \sum_{j \neq n} Z_j \mathcal{L}_0^{-1} \mathcal{L}_j + Z_n \mathcal{L}_0^{-1} \mathcal{L}_n \right) \\
&= \mathcal{L}_0 \left(I + \sum_{j \neq n} Z_j \mathcal{L}_0^{-1} \mathcal{L}_j \right) \left(I - \left(-I - \sum_{j \neq n} Z_j \mathcal{L}_0^{-1} \mathcal{L}_n \right)^{-1} \mathcal{L}_0^{-1} \mathcal{L}_n Z_n \right) \\
&=: AB(I - CZ_n)
\end{aligned}$$

where B is invertible since $\left\| \sum_{j \neq n} Z_j \mathcal{L}_0^{-1} \mathcal{L}_j \right\| \leqslant \sum_{j \neq n} \beta_{0,j} \leqslant \kappa - \beta_{0,n} < 1$ and the same Neumann series argument used above. We also see that $\|C\| |Z_n| \leqslant (1 - \kappa + \beta_{0,n})^{-1} \beta_{0,n} |Z_n| < 1$, so long as $|Z_n| < (1 - \kappa + \beta_{0,n})\beta_{0,n}^{-1}$ which is strictly greater than one. Thus, we employ another Neumann series argument on $I - CZ_n$ to show that

$$(I - CZ_n)^{-1} = \sum_{k=0}^{\infty} C^k Z_n^k.$$

Inverting f, we have

$$u(Z_n) = \sum_{k=0}^{\infty} Z_n^k C^k (AB)^{-1} f,$$

and thus u is analytic in $\mathcal{D}_{(1-\kappa+\beta_{0,n})\beta_{0,n}^{-1}} = \left\{ z \in \mathbb{C} \mid |z| \leqslant (1 - \kappa + \beta_{0,n})\beta_{0,n}^{-1} \right\}$. Note also that we still have the same stability estimate

$$\|u(Z_n)\|_{\mathcal{X}} \leqslant \frac{\|f\|_{\mathcal{Y}^*}}{(1 - \kappa)\mu_0}.$$

□

Remark 3.4. *The previous argument shows that discs in the complex plane exist where the restrictions of the solution are analytic whose size depends on the norms of the coefficients $\beta_{0,n}$. We can also reverse our perspective and start with some set of radii $\rho = (\rho_n)_{n\in\mathbb{N}_+}$ defining discs $\mathcal{D}_{\rho_n}$ where we want the restrictions of the solution to be analytic (cf. the ovular regions Σ_n with quasi-radial coordinate τ_n in Section 1.8.3.1 containing the elliptic regions $\mathcal{E}_{r_n}$ with quasi-radial coordinate r_n where the Chebyshev series expansion for $u(Z_n)$ is convergent and we have $\mathcal{D}_{\rho_n}$ containing Σ_n containing $\mathcal{E}_{r_n}$.). If we have this as the starting point, we can require then that*

$$\sum_{n\in\mathbb{N}_+} \rho_n\beta_{0,n} \leqslant 1-\delta, \tag{3.6}$$

which by the same argument will show that $u(Z_n)$ is analytic in $\mathcal{D}_{\rho_n}$ for any fixed $Z_n^* \in \mathcal{D}_n^* = \prod_{j\neq n}\mathcal{D}_{\rho_j}$. Additionally, we will see that $I + \sum_{n\in\mathbb{N}_+} Z_n\mathcal{L}_0^{-1}\mathcal{L}_n^{-1}$ is invertible with inverse having norm bounded by δ^{-1}. Therefore, we will have the stability estimate

$$\|u(Z)\|_{\mathcal{X}} \leqslant \frac{\|f\|_{\mathcal{Y}^*}}{\delta\mu_0}. \tag{3.7}$$

Definition 3.2. *A sequence of radii $\rho = (\rho_n)_{n\in\mathbb{N}_+}$ is called* δ-admissible *if* (3.6) *holds for some $0 < \delta < 1-\kappa$.*

Example 3.2. *For any $\delta < 1-\kappa$, we have that $\rho = \frac{1-\delta}{\kappa} > 1$, and by Assumption 3.1, the constant sequence $\rho_n = \rho$ is δ-admissible. Instead of increasing the analyticity equally across all directions, we can interpret the proof of Proposition 3.2 as giving all of the analyticity to the single dimension Z_n. The same argument will give the δ-admissible sequence of $\rho_n^* = 1$ and $\rho_n = \frac{1-\delta-\kappa+\beta_{0,n}}{\beta_{0,n}}$*

Under these assumptions, we now show the promised result of exponential decay of the Chebyshev coefficients $\hat{u}_\nu$ in terms of the size of the analyticity regions of u.

Proposition 3.3 ([23], Proposition 4.1). *For any δ-admissible sequence ρ,*

$$\|\hat{u}_\nu\|_{\mathcal{X}} \leqslant \frac{\|f\|_{\mathcal{Y}^*}}{\delta\mu_0} 2^{\|\nu\|_0/2}\rho^{-\nu},$$

where $\rho^{-\nu} = \prod_{n\in\mathbb{N}_+}\rho_n^{-\nu_n}$.

Proof. We will proceed by directly bounding the integrals defining the Chebyshev coefficients as given in (1.36). First for $\nu = 0$, the formula for $\hat{u}_0$ gives

$$\|\hat{u}_0\|_{\mathcal{X}} \leqslant \left\|\int_\Gamma u(z)\pi(z)\ dz\right\|_{\mathcal{X}} \leqslant \sup_{Z\in\Gamma}\|u(Z)\|_{\mathcal{X}} \leqslant \frac{\|f\|_{\mathcal{Y}^*}}{\delta\mu_0},$$

since Γ is a probability space, and δ-admissibility implies the stability estimate (3.7). Now fixing any $\nu \in \Lambda\setminus\{0\}$, and relabeling $\operatorname{supp}(\nu) = (n_j)_{j=1}^J$ and $\Gamma^*_{\operatorname{supp}(\nu)} = \prod_{j\notin\operatorname{supp}(\nu)}\Gamma_j$

$$\hat{u}_\nu = \int_{\Gamma^*_{\operatorname{supp}(\nu)}} \left(\int_{-1}^1 \cdots \left(\int_{-1}^1 u(z)\phi_{\nu_{n_1}}(z_{n_1})\pi_{n_1}(z_{n_1})\ dz_{n_1}\right)\cdots\phi_{\nu_{n_J}}(z_{n_J})\pi_{n_J}(z_{n_J})\ dz_{n_J}\right)\ dz^*$$

where $dz^* = \bigotimes_{j\notin\operatorname{supp}(\nu)}\pi_j(z_j)\ dz_j$. We will work with the inner integral (and to ease notation let $\nu_{n_1} = n'$), and iterate the techniques for the remaining dimensions in the support of ν.

Now, by the change of variables $z_n = \cos(\theta)$ followed by the further change of variables $\zeta = e^{i\theta}$,

$$\begin{aligned}\int_{-1}^{1} u(z_n, z_n^*)\phi_{n'}(z_n)\pi_n(z_n)\, dz_n &= \frac{\sqrt{2}}{\pi}\int_0^{\pi} u(\cos(\theta), z_n^*)\cos\big(n'\theta\big)\, d\theta \\ &= \frac{\sqrt{2}}{2\pi}\int_{-\pi}^{\pi} u(\cos(\theta), z_n^*)\cos\big(n'\theta\big)\, d\theta \\ &= \frac{\sqrt{2}}{2i\pi}\int_{|\zeta|=1} u\left(\frac{\zeta+\zeta^{-1}}{2}, z_n^*\right)\frac{\zeta^{n'}+\zeta^{-n'}}{2}\frac{d\zeta}{\zeta} \\ &= \frac{\sqrt{2}}{4i\pi}\int_{|\zeta|=1} u\left(\frac{\zeta+\zeta^{-1}}{2}, z_n^*\right)\zeta^{n'-1}\, d\zeta \\ &\quad + \frac{\sqrt{2}}{4i\pi}\int_{|\zeta|=1} u\left(\frac{\zeta+\zeta^{-1}}{2}, z_n^*\right)\zeta^{-(n'+1)}\, d\zeta.\end{aligned}$$

The final step is to use the analyticity of u and Cauchy's theorem to deform the integration region in terms of ρ_n and it will turn out that we need u to be analytic in the (elliptical) annulus

$$A_{\rho_n} = \left\{\frac{\zeta+\zeta^{-1}}{2} \mid \rho_n^{-1} \leqslant |\zeta| \leqslant \rho_n\right\}.$$

This will allow us to integrate the top integral over $|\zeta| = \rho_n$ and the bottom over $|\zeta| = \rho_n^{-1}$.

Noting that for any $1 < \sigma \leqslant \rho_n$, we have

$$\left\{\frac{\zeta+\zeta^{-1}}{2} \mid |\zeta| = \sigma\right\} = \left\{\frac{\sigma+\sigma^{-1}}{2}\cos(\theta) + i\frac{\sigma-\sigma^{-1}}{2}\sin(\theta) \mid \theta \in [0, 2\pi)\right\} = \mathcal{E}_{\log(\sigma)}$$

(that is, the "Bernstein ellipses" which define the region of convergence of the Chebyshev expansion in (1.51) with quasi-radial coordinate $\mu_n = \log(\sigma)$), we see that the semi-axes of these ellipses are both bounded by $\sigma < \rho_n$ and so $\mathcal{E}_{\log(\sigma)} \subset \mathcal{D}_{\rho_n}$, where $u(z_n, z_n^*)$ is analytic by δ-admissibilty. Additionally, for the same σ, we have the complementary equality for the reciprocal

$$\left\{\frac{\zeta+\zeta^{-1}}{2} \mid |\zeta| = \sigma^{-1}\right\} = \left\{\frac{\sigma+\sigma^{-1}}{2}\cos(\theta) - i\frac{\sigma-\sigma^{-1}}{2}\sin(\theta) \mid \theta \in [0, 2\pi)\right\} = \mathcal{E}_{\log(\sigma)},$$

just traversed in reverse. This gives that the entire elliptical annulus $A_{\rho_n} \subset \mathcal{D}_{\rho_n}$. Since the product of u with any power of ζ is also analytic in this annulus, we may apply Cauchy's theorem to the two integrands at hand, giving

$$\begin{aligned}&\int_{-1}^{1} u(z_n, z_n^*)\phi_{n'}(z_n)\pi_n(z_n)\, dz_n \\ &\qquad = \frac{\sqrt{2}}{4i\pi}\left[\rho_n^{n'-1}\int_{|\zeta|=\rho_n} u\left(\frac{\zeta+\zeta^{-1}}{2}, z_n^*\right) d\zeta + \rho_n^{n'+1}\int_{|\zeta|=\rho_n^{-1}} u\left(\frac{\zeta+\zeta^{-1}}{2}, z_n^*\right) d\zeta\right].\end{aligned}$$

Applying $\|\cdot\|_{\mathcal{X}}$ and bringing inside the integral, we find

$$\left\|\int_{-1}^{1} u(z_n, z_n^*)\phi_{n'}(z_n)\pi_n(z_n)\, dz_n\right\|_{\mathcal{X}} \leqslant \sqrt{2}\rho_n^{n'}\sup_{Z\in\Gamma}\|u(Z)\|_{\mathcal{X}} \leqslant \sqrt{2}\rho_n^{n'}\frac{\|f\|_{\mathcal{Y}^*}}{\delta\mu_0}.$$

However, if we first iterate the application of Cauchy's theorem to each nontrivial integral in the definition of $\hat{u}_\nu$ before bringing the norm inside the integral, we find (after correcting our relabeling of $\nu_n = n'$)

$$\|\hat{u}_\nu\| \leqslant 2^{\|\nu\|_0/2}\rho^\nu\frac{\|f\|_{\mathcal{Y}^*}}{\delta\mu_0},$$

as desired. □

We now work to show that $\hat{u} \in \ell_{\omega,p}(\Lambda)$ for some $0 < p \leqslant 1$ and weight sequence ω The outline will be to use Proposition 3.3 for a constructed δ-admissible sequence ρ using some further assumptions on summability of β_0 when weighted by a set of auxiliary weights b working on the dimensions of the parameters. Then based on the structure of ρ and the weighted summability of β_0, we will show weighted (with respect to the gPC indices) summability of $\hat{u}$ under a new set of weights ω relating to b in a manner which meshes well with the requirements for the recovery by weighted ℓ_1 minimization in Chapter 2.

THEOREM 3.2 ([**23**], Theorem 4.3). *Suppose that in addition to Assumption 3.1, the sequence $\beta_{0,n} = \left\|\mathcal{L}_0^{-1}\mathcal{L}_n\right\|$ satisfies summability with respect to a set of weights $b = (b_n)_{n\in\mathbb{N}_+}$ with $b_n \geqslant 1$ for all $n \in \mathbb{N}_+$ in the sense that*

$$\sum_{n\in\mathbb{N}_+} \beta_{0,n} b_n^{(2-p)/p} \leqslant \kappa_{b,p} < 1, \qquad \sum_{n\in\mathbb{N}_+} \beta_{0,n}^p b_n^{(2-p)} < \infty. \tag{3.8}$$

For any $\theta \geqslant 1$ construct the set of weights ω indexed by Λ as

$$\omega_\nu = \theta^{\|\nu\|_0} b^\nu, \quad \nu \in \Lambda.$$

Then $\hat{u} \in \ell_{\omega,p}(\Lambda)$.

PROOF. We first construct a δ-admissible set for β_0 denoted ρ for some δ to be determined. For notational convenience, for any number $a \geqslant 1$, we denote $\tilde{a} = a^{(2-p)/p}$. We can partition $\mathbb{N}_+$ into two pieces $E \sqcup F$ defined so F is the set indices of the tail sum $\sum_{n\in F} \beta_{0,n}\tilde{b}_j$ which makes this tail smaller than some $\varepsilon_F > 0$ which is possible by (3.8). Additionally, we can use another parameter $\alpha' > 0$ to shrink the finite sum $\alpha' \sum_{n\in E} \beta_{0,n}\tilde{b}_j$ under some $\varepsilon_E > 0$. We use this partition to then define the δ-admissible sequence ρ separately on E and F. Plugging into the sum concerning δ-admissibility, we have

$$\sum_{n\in\mathbb{N}_+} \rho_n\beta_{0,n} = \sum_{n\in E} \rho_n\beta_{0,n} + \sum_{n\in F} \rho_n\beta_{0,n}.$$

On our first pass, we use our choices for E and F to define

$$\rho_n' = \begin{cases} \alpha'\tilde{b}_n, & \text{if } n \in E \\ \tilde{b}_n, & \text{if } n \in F. \end{cases}$$

Thus,

$$\sum_{n\in\mathbb{N}_+} \rho_n'\beta_{0,n} \leqslant \alpha' \sum_{n\in E} \beta_{0,n}\tilde{b}_j + \sum_{n\in F} \beta_{0,n}\tilde{b}_j \leqslant \varepsilon_E + \varepsilon_F \leqslant 1-\delta.$$

for $\varepsilon_E, \varepsilon_F$ chosen suitably. Obviously, we have a significant amount of freedom in our parameter choices here which we will eventually need to use to help us achieve our goal of weighted summability. In particular, we are not sure that the coefficient decay guaranteed by Proposition 3.3 for this δ-admissible sequence,

$$\|\hat{u}_\nu\|_{\mathcal{X}} \leqslant \frac{\|f\|_{\mathcal{Y}^*}}{\delta\mu_0} 2^{\|\nu\|_0/2} \prod_{n\in E} (\alpha'\tilde{b}_n)^{-\nu_n} \prod_{n\in F} \tilde{b}_n^{-\nu_n},$$

will be useful. Making the simplest definition for ω_ν in terms of b as $\omega_\nu = \prod_{n\in\operatorname{supp}\nu} \theta b_n^{\nu_n} = \theta^{\|\nu\|_0} b^\nu$ (where θ is a parameter useful in scaling the weights to dominate L^∞ norms of Chebyshev (or possibly other) gPC polynomials), we attempt to bound $\|\hat{u}\|_{\omega,p}$. Indeed

$$
\begin{aligned}
\|\hat{u}\|_{\omega,p} &= \sum_{\nu\in\Lambda} \omega_\nu^{2-p}\|\hat{u}_\nu\|^p \\
&= \sum_{\nu\in\Lambda} \tilde{\omega}_\nu^{p}\|\hat{u}_\nu\|^p \\
&\leqslant \left(\frac{\|f\|_{\mathcal{Y}^*}}{\delta\mu_0}\right)^p \sum_{\nu\in\Lambda} \tilde{\theta}^{p\|\nu\|_0} 2^{p\|\nu\|_0/2} \prod_{n\in E} \tilde{b}_n^{p\nu_n}(\alpha'\tilde{b}_n)^{-p\nu_n} \prod_{n\in F} \tilde{b}_n^{p\nu_n}\tilde{b}_n^{-p\nu_n} \\
&= C_\delta^p \sum_{\nu\in\Lambda} (\sqrt{2}\tilde{\theta})^{p\|\nu\|_0} \prod_{n\in E} (\alpha')^{-p\nu_n} \\
&= C_\delta^p \sum_{\nu_E+\nu_F\in\Lambda} (\sqrt{2}\tilde{\theta})^{p(\|\nu_E\|_0+\|\nu_F\|_0)} \prod_{n\in E} (\alpha')^{-p\nu_n} \\
&= C_\delta^p \left(\sum_{\nu\in\Lambda_E} (\sqrt{2}\tilde{\theta})^{p\|\nu\|_0} \prod_{n\in E} (\alpha')^{-p\nu_n}\right)\left(\sum_{\nu\in\Lambda_F} (\sqrt{2}\tilde{\theta})^{p\|\nu\|_0}\right),
\end{aligned}
\tag{3.9}
$$

where Λ_E and Λ_F are the multiindices supported on E and F respectively. Since the summand in the second term is always bounded below by one and Λ_F is (very) infinite, this strategy fails at showing that the norm is finite. Thus, it seems that our choice of ρ_n' for $n \in F$ was too hasty.

Before we revise, let us consider the first term in the product to see whether our choice of ρ_n' on E was any good. We start by enumerating $E = \{e_1, e_2, \dots e_N\}$. Then

$$
\sum_{\nu\in\Lambda_E} (\sqrt{2}\tilde{\theta})^{p\|\nu\|_0} \prod_{n\in E} (\alpha')^{-p\nu_n} = \sum_{\nu_{e_1}=0}^{\infty} \cdots \sum_{\nu_{e_N}=0}^{\infty} (\sqrt{2}\tilde{\theta})^{p\|\nu\|_0} (\alpha')^{-p\nu_{e_1}} \cdots (\alpha')^{-p\nu_{e_N}}.
$$

We can also split up $(\sqrt{2}\tilde{\theta})^{p\|\nu\|_0}$ in terms of $\nu_{e_1}, \dots \nu_{e_N}$ by letting

$$
q(\nu_n) = \begin{cases} \left(\sqrt{2}\tilde{\theta}\right)^p, & \text{if } \nu \geqslant 1 \\ 1 & \text{if } \nu = 0, \end{cases}
$$

giving $(\sqrt{2}\tilde{\theta})^{p\|\nu\|_0} = q(\nu_{e_1}) \cdots q(\nu_{e_N})$. Thus,

$$
\begin{aligned}
\sum_{\nu\in\Lambda_E} (\sqrt{2}\tilde{\theta})^{p\|\nu\|_0} \prod_{n\in E} (\alpha')^{-p\nu_n} &= \left(\sum_{\nu_{e_1}=0}^{\infty} q(\nu_{e_1})(\alpha')^{-p\nu_{e_1}}\right) \cdots \left(\sum_{\nu_{e_N}=0}^{\infty} q(\nu_{e_N})(\alpha')^{-p\nu_{e_N}}\right) \\
&= \left(\sum_{j=0}^{\infty} q(j)(\alpha')^{-pj}\right)^N \\
&= \left(1 + (\sqrt{2}\tilde{\theta})^p(\alpha')^{-p} \sum_{j=0} (\alpha')^{-pj}\right)^N .
\end{aligned}
$$

Now, if $\alpha' > 1$, we find this final geometric sum is finite and thus so is the entire quantity. So when revising our original analysis, we should replace $\rho_n' = \alpha'$ with some $\rho_n > 1$ for all $n \in E$.

Let us return to our choice of ρ_n' and replace it by ρ_n which will actually produce a finite norm. As just determined, for $n \in E$, we wish to choose $\rho_n > 1$. To do this, instead of letting $\alpha' > 0$ such that

$$
\alpha' \sum_{n\in E} \beta_{0,n}\tilde{b}_j < \varepsilon_E,
$$

we write $\alpha' = \alpha - 1$ with $\alpha > 1$. Additionally, we were easily able to prove that ρ'_n was δ-admissible when $\rho'_n = \tilde{\beta}_n$ for $n \in F$, the only problem was that the $\tilde{b}_n^{-p\nu_n}$ in the coefficient decay estimate was canceled by the factor of $\tilde{b}_n^{p\nu_n}$ in the definition of ω_ν. Thus, for $n \in F$, letting $\rho_n = \max\{\tilde{b}_n, c_n\}$ will allow for us to still use this so long as c_n also plays nicely in the δ-admissibility proof as well as is able to help mitigate the infinite growth in the second term of the product bounding the norm above.

To summarize, let

$$\rho_n = \begin{cases} \alpha\tilde{b}_n, & \text{if } n \in E \\ \max\{\tilde{b}_n, c_n\} & \text{if } n \in F, \end{cases}$$

where c_n is to be determined. As before,

$$\begin{aligned} \sum_{n\in\mathbb{N}_+} \rho_n \beta_{0,n} &\leqslant \alpha \sum_{n\in E} \beta_{0,n}\tilde{b}_n + \sum_{n\in F} \beta_{0,n}\tilde{b}_n + \sum_{n\in F} \beta_{0,n}c_n \\ &\leqslant \alpha' \sum_{n\in E} \beta_{0,n} + \sum_{n\in\Lambda} \beta_{0,n}\tilde{b}_n + \sum_{n\in F} \beta_{0,n}c_n \\ &\leqslant \varepsilon_E + \kappa_{b,p} + \sum_{n\in F} \beta_{0,n}c_n. \end{aligned}$$

Now, the factor of $\kappa_{b,p}$ forces us to choose δ in terms of $\kappa_{b,p}$. Since δ must be strictly bounded by one, we can choose for example $\delta = (1 - \kappa_{b,p})/2$. Thus, we need

$$\varepsilon_E + \kappa_{b,p} + \sum_{n\in F} \beta_{0,n}c_n \leqslant 1 - \delta = \frac{1 + \kappa_{b,p}}{2} \iff \varepsilon_E + \sum_{n\in F} \beta_{0,n}c_n \leqslant \frac{1 - \kappa_{b,p}}{2} = \delta.$$

Now, recall that E (and therefore α) depends on F, and F was chosen to make the tail sum smaller ε_F. We can actually use the c terms to pick up the slack here, and save determining ε_F for later. Thus, if we specify $\varepsilon_E = \delta/2$ we simply need to specify c_n so that $\sum_{n\in F} \beta_{0,n}c_n \leqslant \delta/2$. There is again a lot of freedom here, but it turns out that for any $\nu \in \Lambda_F$ the choice

$$c_n = \frac{\delta\nu_n}{2\|\nu\|_1\beta_{0,n}}$$

works well (and has been historically successful in arguments of a similar flavor [12, 13]) in bounding the problematic second factor above.

Now using the δ-admissibility bound again with the more general ρ_n and repeating the argument to bound $\|\hat{u}\|_{\omega,p}$ in (3.9), we obtain

$$\|\hat{u}\|_{\omega,p} \leqslant C_\delta^p \left(\sum_{\nu\in\Lambda_E} \sqrt{2\tilde{\theta}}^{p\|\nu\|_0} \prod_{n\in E} \alpha^{-p\nu_n} \right) \left(\sum_{\nu\in\Lambda_F} (\sqrt{2}\tilde{\theta})^{p\|\nu\|_0} \prod_{n\in F} \tilde{b}_n^{p\nu_n} \min\left\{\tilde{b}_n^{-p\nu_n}, c_n^{-p\nu_n}\right\} \right).$$

As previously discussed, the first factor is finite, so it suffices to bound the second term. If we choose $c_n \leqslant 1$, $(\tilde{b}_n/c_n)^{p\nu_n} \geqslant 1$ allowing us to bound the second factor as

$$\sum_{\nu\in\Lambda_F} (\sqrt{2}\tilde{\theta})^{p\|\nu\|_0} \prod_{n\in F} \left(\frac{\tilde{b}_n}{c_n}\right)^{p\nu_n}.$$

Plugging in the formula for c_n (and letting $g_n = \beta_{0,n}2/\delta$), this reduces to showing that the sequence

$$\left((\sqrt{2}\tilde{\theta})^{\|\nu\|_0} \prod_{n\in F} \left(\frac{\|\nu\|_1\tilde{b}_n g_n}{\nu_n}\right)^{\nu_n} \right)_{\nu\in\Lambda_F}$$

is in $\ell_p(\Lambda_F)$.

In order to bound the terms of this sequence, we make use of the Stirling estimates

$$\frac{n!e^n}{e\sqrt{n}} \leqslant n^n \leqslant \frac{n!e^n}{\sqrt{2\pi}\sqrt{n}}$$

for all $n \geqslant 1$. Rewriting the product, we find

$$\begin{aligned}\prod_{n\in F}\left(\frac{\|\nu\|_1\tilde{b}_n g_n}{\nu_n}\right)^{\nu_n} &= \frac{\|\nu\|_1^{\|\nu\|_1}}{\prod_{n\in F}\nu_n^{\nu_n}}\tilde{b}^\nu g^\nu \\ &\leqslant \frac{\|\nu\|_1!e^{\|\nu\|_1}\prod_{n\in F}\max\{1, e\sqrt{\nu_n}\}}{\prod_{n\in F}\nu_n!e^{\nu_n}}\tilde{b}^\nu g^\nu \\ &= \frac{\|\nu\|_1!}{\nu!}\prod_{n\in F}\max\{1, e\sqrt{\nu_n}(\tilde{b}_n g_n)^{\nu_n}\},\end{aligned}$$

where we employ the maximum to account for the case where $\nu_n = 0$. Bounding $(\sqrt{2}\tilde{\theta}) \leqslant \prod_{n\in F}(\sqrt{2}\tilde{\theta})^{\nu_n}$, we have

$$(\sqrt{2}\tilde{\theta})^{\|\nu\|_0}\prod_{n\in F}\left(\frac{\|\nu\|_1\tilde{b}_n g_n}{\nu_n}\right)^{\nu_n} \leqslant \frac{\|\nu\|_1!}{\nu!}\prod_{n\in F}\max\{1, e\sqrt{\nu_n}(\sqrt{2}\tilde{\theta}\tilde{b}_n g_n)^{\nu_n}\}.$$

Noting that $e\sqrt{\nu_n} \leqslant e^{\nu_n}$ for $\nu_n \geqslant 1$ (which can be seen by comparing derivatives), we replace the maximum with

$$\max\{1, e\sqrt{\nu_n}(\sqrt{2}\tilde{\theta}\tilde{b}_n g_n)^{\nu_n}\} \leqslant (\sqrt{2}\tilde{\theta}e\tilde{b}_n g_n)^{\nu_n} =: h_n^{\nu_n}.$$

Noting that

$$h_n \leqslant \frac{2e\sqrt{2}\tilde{\theta}}{\delta}\beta_{0,n}\tilde{b}_n.$$

By the assumption (3.8), $h \in \ell_p(F)$, and using our original choice of F,

$$\|h\|_{\ell_1(F)} \leqslant \frac{2e\sqrt{2}\tilde{\theta}}{\delta}\varepsilon_F.$$

Thus choosing $\varepsilon_F < \delta/(2e\sqrt{2}\tilde{\theta})$ gives that $\|h\|_{\ell_1(F)} < 1$.

But where does this get us? We have shown that the sequence we wish to show is in $\ell_p(\Lambda_F)$ is bounded by a sequence

$$\frac{\|\nu\|_1!}{\nu!}h^\nu$$

where $h \in \ell_p(F)$, and $\|h\|_{\ell_1(F)} < 1$. The proof is then finished by employing the following lemma.

□

Lemma 3.1 ([**12**], Theorem 7.2). *For* $0 < p \leqslant 1$, *the sequence* $\left(\frac{\|\nu\|_1!}{\nu!}h^\nu\right)_{\nu\in\Lambda} \in \ell_p(\Lambda)$ *if (and only if)* $h \in \ell_p(\mathbb{N}_+)$ *and* $\|h\|_{\ell_1(\mathbb{N}_+)} < 1$.

Proof. We prove only the if direction which is the only direction needed to finish the proof of Theorem 3.2. We start by proving a bound on the simpler sequences of the form $\{\alpha^\nu\}_{\nu\in\Lambda}$ [**12**, Lemma 7.1]. By the factoring argument used to prove the finiteness of the first factor bounding $\|\hat{u}\|_{\omega,p}$ in the proof of Theorem 3.2 and summing the resulting geometric series, we find

$$\|\alpha^\nu\|_{\ell_p(\Lambda)}^p = \prod_{n\in\mathbb{N}_+}\frac{1}{1-\alpha_n^p},$$

for $\alpha_n < 1$ for all n. Additionally, we have

$$1 \leqslant 1-\alpha_n^p + \frac{1-\alpha_n^p}{1-\|\alpha\|_\infty^p}\alpha_n^p \implies \frac{1}{1-\alpha_n^p} \leqslant 1 + \frac{1}{1-\|\alpha\|_\infty^p}\alpha_n^p,$$

and thus

$$\log\left(\|\alpha^\nu\|^p_{\ell_p(\Lambda)}\right) \leqslant \sum_{n\in\mathbb{N}_+} \log\left(1 + \frac{1}{1-\|\alpha\|^p_\infty}\alpha^p_n\right) \leqslant \frac{1}{1-\|\alpha\|^p_\infty}\sum_{n\in\mathbb{N}_+}\alpha^p_n = \frac{\|\alpha\|^p_{\ell_p(\mathbb{N}_+)}}{1-\|\alpha\|^p_\infty},$$

giving the bound

$$\|\alpha^\nu\|_{\ell_p(\Lambda)} \leqslant \exp\left(\frac{\|\alpha\|_{\ell_p(\mathbb{N}_+)}}{p(1-\|\alpha\|^p_\infty)}\right). \tag{3.10}$$

Additionally, for any $\|\gamma\|_{\ell_1(\mathbb{N}_+)}$, by the multinomial theorem,

$$\left\|\frac{\|\nu\|_1!\gamma^\nu}{\nu!}\right\|_{\ell_1(\Lambda)} = \sum_{k=0}^\infty \sum_{\|\nu\|_1=k}\frac{k!}{\nu!}\gamma^\nu = \sum_{k=0}^\infty\left(\sum_{n\in\mathbb{N}_+}\gamma_n\right)^k = \frac{1}{1-\|\gamma\|_{\ell_1(\mathbb{N}_+)}}. \tag{3.11}$$

Now, we assume we can factor $h_n = \gamma_n\alpha_n$ where

$$\|\gamma\|_{\ell_1(\mathbb{N}_+)} < 1, \quad \|\alpha\|_\infty < 1, \quad \|\alpha\|_{\ell_{p'}(\mathbb{N}_+)} < \infty,$$

for $p' = p/(1-p)$. Separating into these factors, applying Hölder's inequality, and the bounds (3.10), and (3.11),

$$\begin{aligned}\sum_{\nu\in\Lambda}\left(\frac{\|\nu\|_1!}{\nu!}h^\nu\right)^p &= \sum_{\nu\in\Lambda}\left(\frac{\|\nu\|_1!}{\nu!}\gamma^\nu\right)^p\alpha^{\nu p} \\ &\leqslant \left(\sum_{\nu\in\Lambda}\frac{\|\nu\|_1!}{\nu!}\gamma^\nu\right)^p\left(\sum_{\nu\in\Lambda}\alpha^{\nu\frac{p}{1-p}}\right)^{1-p} \\ &\leqslant \left(\frac{1}{1-\|\gamma\|_{\ell_1(\mathbb{N}_+)}}\right)^p \exp\left(\frac{(1-p)\|\alpha\|_{\ell_{p'}(\mathbb{N}_+)}}{1-\|\alpha\|^{p'}_\infty}\right) < \infty.\end{aligned}$$

In order to show that the factors γ, α exist, we start in a similar fashion to the previous proof, fixing some cutoff point N such that the tail is bounded by

$$\sum_{n>N} h^p_n < \delta$$

for some $\delta > 0$ to be determined which is possible since $h \in \ell_p(\mathbb{N}_+)$. We then define the factors

$$\gamma_n = \begin{cases}(1+\delta)h_n, & \text{if } n \leqslant N\\ h^p_n, & \text{if } n > N,\end{cases} \qquad \alpha_n = \begin{cases}\frac{1}{1+\delta}, & \text{if } n\leqslant N\\ h^{1-p}_n, & \text{if } n > N.\end{cases}$$

Since $\|h\|_\infty < 1$, we also have $\|\delta\|_\infty \leqslant 1$. Additionally, we calculate

$$\|\gamma\|_{\ell_1(\mathbb{N}_+)} = (1+\delta)\sum_{n\leqslant N}h_n + \sum_{n>N}h^p_n < (1+\delta)\|h\|_{\ell_1(\mathbb{N}_+)} + \delta.$$

Choosing $\delta = (1-\|h\|_{\ell_1(\mathbb{N}_+)})(1+\|h\|_{\ell_1(\mathbb{N}_+)})^{-1}$ makes this sum strictly less than one. Finally,

$$\sum_{n\in\mathbb{N}_+}\alpha_n^{p/(1-p)} = N\left(\frac{1}{1+\delta}\right) + \sum_{n>N}h^p_n < \infty,$$

since $h\in\ell_p(\mathbb{N}_+)$. Thus, the necessary factors to show that $\left(\frac{\|\nu\|_1!}{\nu!}h^\nu\right)_{\nu\in\Lambda} \in \ell_p(\Lambda)$ exist, and the proof is complete. □

Example 3.3. *Depending on the structure of the sequence* β_0 *satisfying just summability*

$$\sum_{n\in\mathbb{N}_+} \beta_{0,n} \leqslant \kappa < 1 \tag{3.12}$$

as in Assumption 3.1, we can derive a set of weights b *so that the properties in* (3.8) *hold and therefore the corresponding solution has gPC coefficients* $\hat{u} \in \ell_{\omega,p}(\Lambda)$.

Trivial weights: *Taking* $b = 1$, (3.12) *satisfies the first condition of* (3.8). *To satisfy the second part, we can simply require* $\beta_0 \in \ell_p(\mathbb{N}_+)$. *Then the weights that we apply to the function have the form* $\omega_\nu = \theta^{\|\nu\|_0}$. *Then* $\hat{u} \in \ell_{\omega,p}(\Lambda)$ *tells us that* $\hat{u}$ *must have smaller terms which involve large numbers of dimensions* $\|\nu\|_0$ *to counteract the exponential growth of* $\theta^{\|\nu\|_0}$ *in the number of dimensions.*

Constant weights: *Instead of taking* $b = 1$, *we can try and push* (3.12) *to its limits, and take* $b_n = 1 + \tau$ *so that*

$$\sum_{n\in\mathbb{N}_+} \beta_{0,n} b_j^{(2-p)/p} \leqslant (1+\tau)^{(2-p)/p}\kappa = \kappa_{\nu,p} < 1,$$

where $(1+\tau)^{(2-p)/p}$ *makes up some of the multiplicative slack* κ^{-1} *between* κ *and 1. Again,* $\beta_0 \in \ell_p(\mathbb{N}_+)$ *also implies* $\beta_0 \in \ell_{b,p}(\mathbb{N}_+)$, *so* (3.8) *is satisfied. Then the gPC weights are* $\omega_\nu = \theta^{\|\nu\|_0}(1+\tau)^{\|\nu\|_1}$. *Now these weights will grow exponentially in the number of nontrivial dimensions* $\|\nu\|_0$ *as well as in the total degree of the gPC polynomials. Since the coefficients of the solution satisfy* $\hat{u} \in \ell_{\omega,p}(\Lambda)$, *as before, they will not be large when a large number of dimensions are involved as well as not be large when corresponding to high degree gPC polynomials. This hints at our intuition of* u *being well represented by a sparse gPC expansion.*

Polynomial weights: *Now, if we assume some further structure on the summability* (3.12), *such as having the norms decay polynomially, we can derive more interesting weights characterizing anisotropic behavior. In particular, we assume* $\beta_{0,n} \leqslant cn^{-t}$ *for* $t > 1$ *and* $c > 0$ *small enough to have*

$$\sum_{n\in\mathbb{N}_+} \beta_{0,n} \leqslant c \sum_{n\in\mathbb{N}_+} n^{-t} < 1,$$

so that (3.12) *is satisfied. For the first condition of* (3.8), *we let* $b_n^{(2-p)/p} = \gamma n^\tau$ *for some* $g > 1$ *so that*

$$\sum_{n\in\mathbb{N}_+} \beta_{0,n} b_n^{(2-p)/p} \leqslant c\gamma \sum_{n\in\mathbb{N}_+} n^{-(t-\tau)}.$$

Choosing $\tau < t - 1$ *small enough will ensure that*

$$\sum_{n\in\mathbb{N}_+} n^{-(t-\tau)} \leqslant \sum_{n\in\mathbb{N}_+} n^{-t} + \varepsilon \leqslant \frac{1}{c} + \varepsilon,$$

by continuity of the Riemann-zeta function for $t > 1$. *Choosing* $\varepsilon = (\gamma - 1)/c$ *gives the first condition of* (3.8). *To show that* $\beta_0 \in \ell_{b,p}(\mathbb{N}_+)$, *we consider*

$$\sum_{n\in\mathbb{N}_+} \beta_{0,n}^p b_n^{2-p} \leqslant c^p\gamma^p \sum_{n\in\mathbb{N}_+} n^{-p(t-\tau)},$$

which is finite when in addition to $\tau < t - 1$*, we also have* $\tau < t - 1/p$ *which is only valid for* $p > 1/t$*. Now, under this setup, our gPC weights take the form*

$$\omega_\nu = \theta^{\|\nu\|_0} \prod_{n \in \operatorname{supp} \nu} \gamma^{\tilde{p}\nu_n} n^{\tilde{p}\tau\nu_n},$$

for $\tilde{p} = p/(2-p)$*. As always, these weights grow in the number of dimensions. However, for a fixed dimension, we see that the growth is exponential as* ν_n *increases, and thus, the gPC coefficients of the solution should decay exponentially in fixed dimensions. For a fixed multiindex* ν *taken as a truncation of the constant multiindex* $j \cdot 1$ *for* $j \in \mathbb{N}_+$ *some constant, we see that the weights will grow polynomially in the number of dimensions. Thus, in general, for large dimensions, the coefficients will have to be polynomially decaying.*

Exponential weights: *Finally, we consider the case where the operator norms are bounded exponentially, as* $\beta_{0,n} \leqslant c\alpha^n$*. When* $c > 0$ *is small enough and* $\alpha < 1$*, we have* $\beta_0 \in \ell_1(\mathbb{N}_+)$ *as*

$$\sum_{n \in \mathbb{N}_+} \beta_{0,n} \leqslant c \sum_{n \in \mathbb{N}_+} \alpha^n \leqslant \frac{c\alpha}{1-\alpha} < 1,$$

For weights, we proceed analogously to the polynomial case. To ease notation, choose some $b_n^{(2-p)/p} = \sigma^n$ *for some* $\sigma > 1$*, which gives*

$$\sum_{n \in \mathbb{N}_+} \beta_{0,n} b_n^{(2-p)/p} \leqslant c \sum_{n \in \mathbb{N}_+} (\alpha\sigma)^n = \frac{c\alpha\sigma}{1-\alpha\sigma}$$

so long as $\sigma < 1/\alpha$*. As before, continuity of* $cx/(1-x)$ *gives that for some* σ *close enough to one depending on* c *and* α*, we can have*

$$\frac{c\alpha\sigma}{1-\alpha\sigma} \leqslant \frac{c\alpha}{1-\alpha} + \varepsilon,$$

where we can choose ε *to be contained in the slack* $1 - c\alpha/(1-\alpha)$*, thus giving the first condition of*(3.8)*. For* $\beta \in \ell_{b,p}(\mathbb{N}_+)$*, we calculate*

$$\sum_{n \in \mathbb{N}_+} \beta_{0,n}^p b_n^{2-p} \leqslant c^p \sum_{n \in \mathbb{N}_+} (\alpha\sigma)^{pn},$$

which is still convergent no matter the value of $p < 1$ *when* $\sigma < 1/\alpha$*. We find the gPC weights are*

$$\omega_\nu = \theta^{\|\nu\|_0} \prod_{n \in \operatorname{supp}(\nu)} \sigma^{\tilde{p} n \nu_n}$$

where $\tilde{p} = p/(2-p)$ *as before. Here for fixed* n *, we again have exponential increase in* ν_n*. However, for* ν_n *a fixed constant vector as before, increase in* n *is exponential. Thus, we have that the corresponding solution coefficients should also decay exponentially for fixed gPC polynomial orders as the dimension increases.*

3.3. The Compressive Sensing Petrov-Galerkin Algorithm

We have achieved our goal of being able to apply the results of Chapter 2, in particular, Corollary 2.2. Instead of applying the compressive sensing reconstruction to our solution u which is in general $\mathcal{X}$-valued, we apply a functional $G \in \mathcal{X}^*$ (e.g. some spatial QoI of the solution of a UQ problem) to produce a new function $F : \Gamma \to \mathbb{C}$ defined by $F(Z) = G(u(Z))$. We additionally have the Chebyshev-gPC expansion for F

$$F = \sum_{\nu \in \Lambda} \hat{F}_\nu \phi_\nu := \sum_{\nu \in \Lambda} G(\hat{u}_\nu)\phi_\nu,$$

which by the boundedness of G inherits the result of $F \in S_{\omega,p}(\Lambda)$ from Theorem 3.2 under the conditions defined there (i.e. weighted ℓ_p summability of β_0 with respect to a set of weights $b \geqslant 1$ and $\|\beta\|_{b^{(2-p)/p},1} < 1$) with the weight sequence

$$\omega_\nu = \theta^{\|\nu\|_0} b^\nu.$$

With this setup, we define Algorithm 10 to produce an approximate solution to (3.1) and prove an associated convergence result Theorem 3.3.

Algorithm 3.1: Calculating the approximate solution to a parametric PDE using compressive sensing and Petrov-Galerkin (CSPG) discretization.

```
Input:
    • Weights $(b_n)_{n \in \mathbb{N}_+} \geqslant 1$ satisfying (3.8) for some $0 < p < 1$.
    • Accuracy $\varepsilon$ of the Petrov-Galerkin approximation and sparsity parameter $s$.
    • Index set $\Lambda_0 = \{\nu \in \Lambda \mid \omega_\nu^2 \leqslant s/2\}$ where $\omega_\nu = 2^{\|\nu\|_0/2} b^\nu$ such that the cardinality $M := |\Lambda_0| < \infty$ is finite with maximum dimension $N = \max_{\nu \in \Lambda_0, \nu_n \neq 0} n$.
    • Number of samples $K \asymp s \log(M) \log^3(s)$
Result: An approximation $F^\sharp(Z)$ to the functional $G$ applied to the solution $u(Z)$ of (3.1).
Draw samples $Z^{(1)}, \dots, Z^{(K)} \in \Gamma = [-1,1]^N$ independently from the tensorized Chebyshev measure (3.1).
for $k = 1, \dots, K$ do
    Obtain the Petrov-Galerkin discretization $u_{N,h}(Z^{(k)})$ of the dimension-truncated problem for some $h$ depending on $\varepsilon > 0$ such that $|G(u(Z^{(k)})) - G(u_{N,h}(Z^{(k)}))| < \varepsilon$.
    $y_k \leftarrow G(u_{N,h}(Z^{(k)}))$.
    for $\nu \in \Lambda_0$ do Calculate the corresponding sample matrix row.
        $A_{k,\nu} \leftarrow \phi_\nu(Z^{(k)})$.
    end
end
Compute the solution $\hat{F}^\sharp$ to the weighted $\ell_1$-minimization program
    $\underset{z \in \mathbb{C}}{\text{minimize}} \|z\|_{\omega,1}$ subject to $\|Az - y\|_2 \leqslant 2\sqrt{K}\varepsilon$.
$F^\sharp \leftarrow \sum_{\nu \in \Lambda_0} \hat{F}^\sharp_\nu \phi_\nu$.
```

Theorem 3.3 ([23], Theorem 5.1). *With Assumptions 3.1, 3.2, and 3.3, let* $u = \sum_{\nu \in \Lambda} \hat{u}_\nu \phi_\nu$ *be the true solution to the affine parametric PDE* (3.1), *and* $F = G(u) = \sum_{\nu \in \Lambda} \hat{F}_\nu \phi_\nu$ *with* $\hat{F}_\nu = G(u_\nu)$. *Running Algorithm 10 with accuracy parameter* $\varepsilon > 0$ *and sparsity parameter* s *satisfying*

$$2^{1/p} s^{1/2-1/p} \|F_{\Lambda_R}\|_{\omega,p} \leqslant \varepsilon \leqslant E s^{1/2-1/p} \|F_{\Lambda_R}\|_{\omega,p}, \tag{3.13}$$

for $E > 2^{1/p}$ *a constant independent of* s*, produces a solution* $F^\sharp$ *that for some universal constants* $B, C > 0$*, with probability exceeding* $1 - 2M^{-\log^3(s)}$*, satisfies*

$$\|F^\sharp - F\|_\infty \leqslant B\|F\|_{\omega,p} s^{1-1/p} \leqslant B'_F \left(\frac{\log(M)\log^3(K)}{K}\right)^{1/p-1},$$

$$\|F^\sharp - F\|_2 \leqslant C\|F\|_{\omega,p} s^{1/2-1/p} \leqslant C'_F \left(\frac{\log(M)\log^3(K)}{K}\right)^{1/p-1/2},$$

where B, C *depend only on* E *and* B'_F, C'_F *depend only on* B*,* C*, and* $\|F\|_{\omega,p}$*.*

Proof. The argument is a copy of the proof of Corollary 2.2 where instead of concluding bounds in terms of $\sigma_{s/2}(F)_{\omega,1}$, we use factors of s and $\|F\|_{\omega,p}$ arising from the Stechkin estimate Theorem 1.5 and the definition of Λ_0. Indeed, to handle the resulting factor of $\sigma_s(F_{\Lambda_0})_{\omega,1}$ in the weighted minimization error bounds, since $F \in \ell_{\omega,p}(\Lambda_0)$ by Theorem 3.2, the weighted Stechkin estimate implies

$$\sigma_s(F_{\Lambda_0})_{\omega,1} \leqslant \left(s - \max_{\nu\in\Lambda_0} \omega_\nu^2\right)^{1-1/p} \|F_{\Lambda_0}\|_{\omega,p} \leqslant \left(\frac{s}{2}\right)^{1-1/p} \|F\|_{\omega,p}, \tag{3.14}$$

since $\omega_\nu^2 \leqslant \frac{s}{2}$ on Λ_0, and $1 - 1/p < 0$.

As for the ℓ_2 norm of the truncated measurements (that is, the values of $F - F_{\Lambda_0} = F_{\Lambda_R}$ with $\Lambda_R = \Lambda \setminus \Lambda_0$ at the measurement points), Lemma 2.4 gives that

$$\left(\sum_{k=1}^K F_{\Lambda_R}^2(Z^{(k)})\right)^{1/2} \leqslant 2\sqrt{\frac{K}{s}}\|F_{\Lambda_R}\|_{\omega,1}.$$

However, on Λ_R, we know that $\left(\frac{s}{2}\right)^{1/p-1} \leqslant \left(\omega_\nu^2\right)^{1/p-1}$ (where the exponents were chosen to match those in (3.14)) which gives

$$\|F_{\Lambda_R}\|_{\omega,1} = \sum_{\nu\in\Lambda_R} \hat{F}_\nu \omega_\nu \leqslant \left(\frac{s}{2}\right)^{1-1/p} \sum_{\nu\in\Lambda_R} \hat{F}_\nu \omega_\nu^{2/p-1} = \left(\frac{s}{2}\right)^{1-1/p} \|F_{\Lambda_R}\|_{\omega^\alpha,1},$$

for $\alpha = 2/p - 1$. By (2.31) in the proof of Lemma 2.5, we can convert the norm in terms of ω^α into the weighted p-norm with respect to ω, giving

$$\|F_{\Lambda_R}\|_{\omega,1} \leqslant \left(\frac{s}{2}\right)^{1-1/p} \|F_{\Lambda_R}\|_{\omega,p}. \tag{3.15}$$

Thus

$$\left(\sum_{k=1}^K F_{\Lambda_R}^2(Z^{(k)})\right)^{1/2} \leqslant 2^{1/p}\sqrt{K}s^{1/2-1/p}\|F_{\Lambda_R}\|_{\omega,p} \leqslant 2^{1/p}\sqrt{K}s^{1/2-1/p}\|F\|_{\omega,p}. \tag{3.16}$$

Since

$$2^{1/p}s^{1/2-1/p}\|F_{\Lambda_R}\|_{\omega,p} \leqslant \varepsilon$$

by assumption, we can bound the ℓ_2 norm of the noisy measurements as

$$\left\|y_k - F_{\Lambda_0}(Z^{(k)})\right\|_{\ell_2([K])} \leqslant \left\|y_k - F(Z^{(k)})\right\|_{\ell_2([K])} + \left\|F_{\Lambda_R}(Z^{(k)})\right\|_{\ell_2([K])} \leqslant 2\sqrt{K}\varepsilon,$$

since each discrete measurement is within ε of the true value $F(Z^{(k)})$.

The error bounds from applying the weighted ℓ_1 minimization program on the finitely indexed F_{Λ_0} from Theorem 2.2 return

$$\|F_{\Lambda_0} - F^\sharp\|_\infty \leqslant B_1 \sigma_s(F_{\Lambda_0})_{\omega,1} + B_2(2\sqrt{K}\varepsilon)\sqrt{\frac{s}{K}}$$
$$\|F_{\Lambda_0} - F^\sharp\|_2 \leqslant \frac{C_1}{\sqrt{s}}\sigma_s(F_{\Lambda_0}) + C_2\frac{2\sqrt{K}\varepsilon}{\sqrt{K}}.$$

An application of (3.14) and (3.13) give

$$\|F_{\Lambda_0} - F^\sharp\|_\infty \leqslant B'\|F\|_{\omega,p} s^{1-1/p}$$
$$\|F_{\Lambda_0} - F^\sharp\|_2 \leqslant C'\|F\|_{\omega,p} s^{1/2-1/p},$$

for B', C' depending only on E and p (which, since E and p are in one-to-one correspondence, can be thought to be depending only on E). On the other hand,

$$\|F - F_{\Lambda_0}\|_\infty \leqslant \|F_{\Lambda_R}\|_{\omega,1} \leqslant \left(\frac{s}{2}\right)^{1-1/p} \|F\|_{\omega,p}$$
$$\|F - F_{\Lambda_0}\|_2 \leqslant \sqrt{\frac{2}{s}}\|F_{\Lambda_R}\|_{\omega,1} \leqslant \left(\frac{s}{2}\right)^{1/2-1/p} \|F_{\Lambda_R}\|_{\omega,p}$$

by (3.15) and (2.28) from the proof of Lemma 2.4. The triangle inequality and these two sets of estimates give the final error

$$\|F - F^\sharp\|_\infty \leqslant B\|F\|_{\omega,p} s^{1-1/p}$$
$$\|F - F^\sharp\|_2 \leqslant C\|F\|_{\omega,p} s^{1/2-1/p},$$

again for $B, C > 0$ depending only on E. Since $K \asymp s\log(M)\log^3(s)$, we have in particular that

$$\frac{1}{s} \lesssim \frac{\log^3(s)\log(M)}{K} \leqslant \frac{\log^3(K)\log(M)}{K},$$

giving the error bounds in terms of only K and M as desired. □

Remark 3.5. *Note that in order to even run the constrained minimization program in Algorithm 10, we need to know the constraint ahead of time, that is, we need an approximation of ε. The only necessity on ε is that given in* (3.13)*, which requires knowledge of $\|F_{\Lambda_R}\|_{\omega,p}$ (or we can also rephrase* (3.13) *to use the possibly easier to compute $\|F\|_{\omega,p}$ with no change to the proof). However, our only insight into this quantity is through the proofs of Theorem 3.2 and Lemma 3.1 which are complex and provide loose bounds. The next chapter will explore alternatives to this constrained weighted ℓ_1 minimization problem which do not rely on bounds for $\|F_{\Lambda_R}\|_{\omega,p}$ or even ε and are still convergent under additional, more exotic errors.*

3.4. Approximating the Size of the Truncated Index Set

In order to quantify the choice $K \asymp s\log(M)\log^3(s)$, we recall that s is a parameter chosen to be in correspondence with ε (or vice versa, we can consider s to be free/fixed and choose ε accordingly) by (3.13). Thus, we need to know (or have bounds on)

$$M = |\Lambda_0| = \left|\left\{\nu \in \Lambda \mid \omega_n^2 \leqslant s/2\right\}\right| = \left|\left\{\nu \in \Lambda \mid 2^{\|\nu\|_0} b^{2\nu} \leqslant s/2\right\}\right|. \tag{3.17}$$

We give a general bound for this quantity below and then proceed with the (very) gory details for some examples of weights as discussed in Example 3.3. We use the convention that $\lg(x) = \log_2(x)$.

Theorem 3.4 ([23], Theorem 5.3). *For a finite weight sequence* $b \in \mathbb{R}^N_{>1}$*, define* $a_n := 2\lg(b_n)$ *and* $A = \lg(s/2)$*. Then*

$$M \leqslant 1 + \sum_{k=1}^{\min\{N,\lfloor A \rfloor\}} \frac{(A-k)^k}{k!} \sum_{\substack{S \subseteq [N] \\ |S|=k \\ \|a_S\|_1 \leqslant A-k}} \prod_{n \in S} a_n^{-1}. \tag{3.18}$$

Proof. Taking the base-2 logarithm in the definition of Λ_0 in (3.17) and decomposing Λ into all possible supports for ν, we find

$$\Lambda_0 = \{\nu \in \Lambda \mid \|\nu\|_0 + \nu \cdot a \leqslant A\} = \{0\} \sqcup \bigsqcup_{k=1}^{N} \bigsqcup_{S \subseteq [N], |S|=k} \{\nu \in \mathbb{N}_+^k \mid \nu \cdot a_S \leqslant A - k\} \tag{3.19}$$

where $\nu \cdot a$ is the dot product on $\mathbb{R}^N$. Thus, we bound the size of the smaller sets of the form $\{\nu \in \mathbb{N}_+^k \mid \nu \cdot c \leqslant B\}$ and sum them. Notice also that these sets are nothing but the anisotropic total degree index sets.

We proceed by induction on k to show that

$$\left|\left\{\nu \in \mathbb{N}_+^k \mid \nu \cdot c \leqslant B\right\}\right| \leqslant \frac{B^k}{k! \prod_{n=1}^k c_n}. \tag{3.20}$$

For $k = 1$, this is true since the largest $\nu_1 \in \mathbb{N}_+$ satisfying $\nu_1 c_1 \leqslant B$ is $\nu_1 = \lfloor B/c_1 \rfloor \leqslant B/c_1$. For general $k \geqslant 1$ and $\nu \in \mathbb{N}_+^{k+1}$, we split the dot product into $\nu \cdot c = \bar{\nu} \cdot \bar{c} + \nu_{k+1} c_{k+1}$ where $\bar{\nu}$ is the first k-components of the vector ν. We then have

$$\begin{aligned} \left|\left\{\nu \in \mathbb{N}_+^{k+1} \mid \nu \cdot c \leqslant B\right\}\right| &= \left| \bigsqcup_{B - \nu_{k+1} c_{k+1} \geqslant 0} \left\{\bar{\nu} \in \mathbb{N}_+^k \mid \bar{\nu} \cdot \bar{c} \leqslant B - \nu_{k+1} c_{k+1}\right\} \right| \\ &= \sum_{\nu_{k+1}=1}^{\lfloor B/c_{k+1} \rfloor} \frac{(B - \nu_{k+1} c_{k+1})^k}{k! \prod_{n=1}^k c_n}. \end{aligned} \tag{3.21}$$

Bounding the sum with an integral, we have

$$\sum_{\nu_{k+1}=1}^{\lfloor B/c_{k+1} \rfloor} (B - \nu_{k+1} c_{k+1})^k \leqslant \int_0^{B/c_{k+1}} (B - \nu_{k+1} c_{k+1})^k \, d\nu_{k+1} = \frac{1}{c_{k+1}} \int_0^B u^k \, du = \frac{B^{k+1}}{(k+1) c_{k+1}}.$$

Plugging this into (3.21) provides the desired inequality. .

Note that in (3.19), we have that $A - k \leqslant 0$ for any $k > \lfloor A \rfloor$, and so there are no ν satisfying the condition $\nu \cdot a_S \leqslant A - k$ by virtue of the fact that $a > 0$. Thus, it suffices to sum k up to $\lfloor A \rfloor$. Additionally, when $\sum_{n \in S} a_n > A - k$, $\nu \in \mathbb{N}_+^k$ implies that $\nu \cdot a_S \geqslant A - k$. So it suffices to only sum over S with $\|a_S\|_1 \leqslant A - k$. Inserting (3.20) into (3.19), taking cardinalities and applying these bounds gives

$$M \leqslant 1 + \sum_{k=1}^{\min\{N,\lfloor A \rfloor\}} \sum_{\substack{S \subseteq [N], |S|=k \\ \|a_S\|_1 \leqslant A-k}} \frac{(A-k)^k}{k! \prod_{n \in S} a_n}$$

as desired. □

Corollary 3.1. *For $s \geqslant 2$, we consider three different cases of weights on β_0 (cf. Example 3.3).*

Constant weights: *Under the finite dimensional noise assumption with* $\Gamma = [-1,1]^N$*, let* $b_n = \sigma = 1+\tau$*,* $\tau > 0$*,* $n \in [N]$ *be constant weights (alternatively, consider an infinite-dimensional parameter space with* $b_n = \infty$ *for* $n > N$*). Then*

$$M \leqslant \begin{cases} (\log_{\sigma^2}(\sigma^2 s/2))^N, & \textit{if } N \leqslant \log_{2\sigma^2}(s/2), \\ \left(\left(1+\frac{1}{\lg(\sigma^2)}\right) eN\right)^{\log_{2\sigma^2}(s/2)} & \textit{if } N > \log_{2\sigma^2}(s/2). \end{cases} \tag{3.22}$$

Polynomial weights: *Suppose that* $b_n = cn^\alpha$ *for some* $c > 1$ *and* $\alpha > 0$*. Then there exist* $C_{\alpha,c} > 0$ *and* $\gamma_{\alpha,c} > 0$ *such that*

$$M \leqslant C_{\alpha,c} s^{\gamma_{\alpha,c} \log(s)}. \tag{3.23}$$

Exponential weights: *Let* $b_n = \sigma^n$ *for some* $\sigma > 1$*. Then*

$$M \leqslant 1 + \frac{C_\sigma}{2\pi\sqrt{\log_\sigma(s/2)}} \left(e^3 \sqrt{\log_\sigma(s/2)}\right)^{\sqrt{\log_\sigma(s/2)}}. \tag{3.24}$$

Proof.

Constant weights: As before, we let $a_n := 2\lg(b_n) = \lg(\sigma^2)$ and $A = \lg(s/2)$. With these constant weights, the assumption $\|a_S\|_1 \leqslant A-k$ for $|S| = k$ becomes $k \leqslant A/(1+a) = A/(1+\lg(\sigma^2))$. We split the analysis into two different cases for $N \geqslant k$, the first being $N \leqslant A/(1+\lg(\sigma^2))$ so that $\|a_S\|_1 \leqslant A-k$ is always satisfied. Notice that this split occurs when

$$N \leqslant \frac{\lg(s/2)}{1+\lg(\sigma^2)} = \frac{\lg(s/2)}{\lg(2\sigma^2)} = \log_{2\sigma^2}(s/2),$$

by the change of base formula. In this case, we also have here that $N \leqslant A$ so that (3.18) becomes

$$\begin{aligned} M &\leqslant 1 + \sum_{k=1}^{N} \frac{(A-k)^k}{k!} \sum_{S \subseteq N, |S|=k} \lg(\sigma^2)^{-k} \\ &= 1 + \sum_{k=1}^{N} \frac{(A-k)^k}{k!} \binom{N}{k} \lg(\sigma^2)^{-k} \\ &\leqslant \sum_{k=0}^{N} \binom{N}{k} (A/\lg(\sigma^2))^k \\ &= (A/\lg(\sigma^2)+1)^N \end{aligned}$$

by the binomial theorem. Substituting $A = \lg(s/2)$ and using the same change of base trick,

$$M \leqslant (\log_{\sigma^2}(\sigma^2 s/2))^N,$$

as desired.

Now for $N > \log_{2\sigma^2}(s/2) = A/\lg(2\sigma^2)$ when $k > \lfloor A/\lg(2\sigma^2) \rfloor$ as above, we have that the set of all $S \subseteq N$ with $|S| = k$ and $\|a_S\|_1 \leqslant A-k$ is empty. Thus, the upper bound in the first sum can be taken to be $\lfloor A/\lg(2\sigma^2) \rfloor$. Keeping the rest of the previous argument the same, we have

$$M \leqslant \sum_{k=0}^{\lfloor A/\lg(2\sigma^2) \rfloor} \binom{N}{k} (A/\lg(\sigma^2))^k \leqslant (A/\lg(\sigma^2))^{\lfloor A/\lg(2\sigma^2) \rfloor} \sum_{k=0}^{\lfloor A/\lg(2\sigma^2) \rfloor} \binom{N}{k}.$$

We bound the sum of the binomial coefficients with a tightening of (2.20). Using part of this bound and the fact that $(N/k)^k$ is increasing in k,

(3.25)
$$\sum_{k=1}^{A'} \binom{N}{k} \leqslant \sum_{k=1}^{A'} \left(\frac{N}{k}\right)^k \frac{k^k}{k!} \leqslant \left(\frac{N}{A'}\right)^{A'} \sum_{k=1}^{A'} \frac{(A')^k}{k!} \leqslant \left(\frac{eN}{A'}\right)^{A'} = \left(\frac{eN}{\lfloor A/\lg(2\sigma^2)\rfloor}\right)^{\lfloor A/\lg(2\sigma^2)\rfloor}.$$

If we again use the fact that $(eN/A')^{A'}$ is increasing in A', we see

$$\sum_{k=1}^{A'} \binom{N}{k} \leqslant \left(\frac{eN}{A/\lg(2\sigma^2)}\right)^{A/\lg(2\sigma^2)}.$$

Using this bound in the one for M above gives

$$M \leqslant \left(eN\frac{\lg(2\sigma^2)}{\lg(\sigma^2)}\right)^{A/\lg(2\sigma^2)} = \left(\left(1+\frac{1}{\lg(\sigma^2)}\right)eN\right)^{\log_{2\sigma^2}(s/2)},$$

finishing the bounds for constant weights.

Polynomial weights: Again, we let $a_n = 2\lg(b_n) = 2\lg c + 2\alpha \lg n$ and $A = \lg(s/2)$. In general, all constants in the following calculations will be positive and depend on c and α. First note that since the weights ω increase polynomially in n and ν, we know that M must be finite and can contain only ν with support contained in some $[N]$. In particular, the we can calculate the largest N by all ν with $\omega_\nu = 2^{\|\nu\|_0} b^{2\nu} \leqslant s/2$. The one allowing for the largest N with $\nu_N \neq 0$ should be fully supported on $\{N\}$ and should be as small as possible, that is, $\nu = e_N$. This gives that $b_N \leqslant \sqrt{s/4}$, and so $N = \left\lfloor (s/(4c))^{1/2\alpha} \right\rfloor$.

Now in (3.18), we will need to bound the product $\prod_{n\in S} a_n^{-1}$ where $S \subseteq [N]$ and $|S| = k$ We can uniformly lower bound any a_n by $2\lg c + 2\alpha$ except for $n = 1$ where we must choose $a_1 \geqslant 2\lg c$. Of course $2\lg c$ is also a lower bound for any other a_n, so we incur no harm by upper bounding the product as

$$2^{-k}(\lg c)^{-1}(\lg c + \alpha)^{-(k-1)} = (2\lg c)^{-1}(2\lg c + 2\alpha)^{-(k-1)} \geqslant \prod_{k\in S} a_n^{-1},$$

just in case $1 \in A$.

We now have to figure out how many of these products we sum up. We are assigned to take only index sets with $|S| = k$ and $\|a_S\|_1 \leqslant A - k$. Expanding this second condition requires that

$$2\alpha \lg\left(\prod_{n\in S} n\right) = 2\alpha \sum_{n\in S} \lg n \leqslant A - k(1 + 2\lg c).$$

The smallest value of the left hand side is when $S = [k]$ giving that $2\alpha\lg(k!) \leqslant A - k(1 + 2\lg c)$. Thus, we instead impose this restriction on k, allowing for S to range over all $S \subseteq [N]$ with $|S| = k$ (of which there are $\binom{N}{s}$ index sets). This then only makes the sum larger, upper bounding our value for M.

Putting these bounds together and summing k up to $\lfloor A \rfloor \geqslant \min\{N, \lfloor A \rfloor\}$, we have

$$\begin{aligned} M &\leqslant 1 + \sum_{\substack{k \in [\lfloor A \rfloor] \\ 2\alpha \lg(k!) \leqslant A - k(1+2\lg c)}} \binom{N}{k} \frac{1}{k!} \left(\frac{A-k}{2} \right)^k \frac{1}{\lg c (\lg c + \alpha)^{k-1}} \\ &\leqslant 1 + \left(1 + \frac{\alpha}{\lg c}\right) \sum_{\substack{k \in [\lfloor A \rfloor] \\ 2\alpha \lg(k!) \leqslant A - k(1+2\lg c)}} \binom{N}{k} \frac{1}{k!} \left(\frac{A}{2\lg c + 2\alpha} \right)^k, \end{aligned}$$

where in the last step, we simply factored out $(\lg c + \alpha)/\lg c$ and removed the subtraction by k on A.

We now quantify the limits of this summation. To do this, we consider summing up to the maximum number L such that

$$2\alpha \lg(L!) \leqslant A - L(1 + 2\lg c). \tag{3.26}$$

Since we already know how to bound quantities of the form $\sum_{k=1}^{L} \binom{N}{k}$ as in the polynomial weight case, we attempt to find a uniform upper bound for

$$\frac{1}{k!} \left(\frac{A}{2\lg c + 2\alpha} \right)^k =: \frac{1}{k!} B^k$$

over k so that we can factor this common upper bound from the sum. This exponential will grow faster than the logarithm so long as $B > k$. We can see this by rewriting

$$\frac{B^k}{k!} = \exp\left(k \log B - \sum_{\ell=1}^{k} \log \ell \right).$$

Under the condition that $B > L \geqslant \ell$, we have $\log B \geqslant \log \ell$, and so adding $L - k$ factors of $\log B - \log \ell \geqslant 0$ for ℓ ranging from $k+1$ to L gives the upper bound

$$\frac{B^k}{k!} \leqslant \exp\left(L \log B - \sum_{\ell=1}^{L} \log \ell \right) = \frac{B^L}{L!}. \tag{3.27}$$

Now this bound is valid when $B = \frac{A}{2\lg c + 2\alpha}$ is in fact larger than L. When $L \geqslant 4$,

$$\lg(L!) = \sum_{\ell=1}^{L} \lg \ell \geqslant 0 + 1 + \lg 3 + \sum_{\ell=4}^{L} \lg(4) \geqslant 1 + 1 + 1 + 1 + \sum_{k=5}^{L} 1 = L.$$

Thus, (3.26) gives that

$$L \leqslant \frac{A}{1 + 2\lg c + 2\alpha} \leqslant \frac{A}{2\lg c + 2\alpha} = B \tag{3.28}$$

as desired. From (3.26), increasing A allows L to be larger, so it is only for A below some threshold depending on c and α that $L < 4$. Since A is in direct correspondence with s, this means that there are finitely many value s of s for which this argument does not hold. By increasing the constants in the bound we obtain $M \leqslant C_{\alpha,c} s^{\gamma_{\alpha,c} \log(s)}$, we can ensure that this bound still holds for the finitely many index sets generated by the s which force $L < 4$. Thus, it suffices to consider only $L \geqslant 4$, where we additionally have (3.28) and (3.27).

Using (3.27) and applying (3.25) from the constant weight case above,

$$\begin{aligned} M &\leqslant 1+\left(1+\frac{\alpha}{\lg c}\right)\frac{1}{L!}\left(\frac{A}{2\lg c+2\alpha}\right)^{L}\sum_{k=1}^{L}\binom{N}{k} \\ &\leqslant 1+\left(1+\frac{\alpha}{\lg c}\right)\frac{1}{L!}\left(\frac{A}{2\lg c+2\alpha}\right)^{L}\left(\frac{eN}{L}\right)^{L} \\ &\leqslant 1+\left(1+\frac{\alpha}{\lg c}\right)\frac{1}{L!}\left(\frac{AeN}{L(2\lg c+2\alpha)}\right)^{L}. \end{aligned}$$

We're now left to bound L! from below, which, by Stirling's formula (2.11), we obtain

$$L! = \Gamma(L+1) = \sqrt{2\pi L}L^{L}e^{-L}\exp\left(\frac{\theta(L)}{12(L+1)}-1\right) \geqslant \frac{\sqrt{2\pi L}}{e}\left(\frac{L}{e}\right)^{L}$$

since $0 \leqslant \theta(L) \leqslant 1$. Thus

$$\begin{aligned} M &\leqslant 1+\left(1+\frac{\alpha}{\lg c}\right)\frac{e}{\sqrt{2\pi L}}\left(\frac{Ae^{2}N}{L^{2}(2\lg c+2\alpha)}\right)^{L} \\ &\leqslant 1+\left(1+\frac{\alpha}{\lg c}\right)\frac{e}{\sqrt{2\pi L}}\left(\frac{Ae^{2}N}{L^{2}(2\lg c+2\alpha)}\right)^{A/(1+2\lg(c2^{\alpha}))} \end{aligned}$$

where the second bound is a consequence of (3.28).

All we have left is to remove our dependence on L by finding a lower bound. We start by noting that since L is the maximal integer such that (3.26) holds, we must have

$$A \leqslant 2\alpha\lg((L+1)!) + (L+1)(1+2\lg c) = 2\alpha\lg(L+1) + 2\alpha\lg(L!) + (L+1)(1+2\lg c).$$

Since $L \geqslant 4$, $L+1 \leqslant \frac{5}{4}L$ and so

$$\lg(L+1) \leqslant \lg\left(\frac{5}{4}\right) + \lg L \leqslant C_1\lg L \leqslant C_1 L.$$

Therefore

$$A \leqslant 2\alpha\lg(L!) + C_2 L.$$

An application of Stirling's formula to now upper bound L! now gives

$$\lg(L!) \leqslant \lg\left(\sqrt{2\pi L}(L/e)^{L}e^{1/12L}\right) \leqslant \frac{1}{2}\lg(2\pi Le^{1/24}) + L\lg(L/e) \leqslant L\lg(C_3L).$$

Thus,

$$A \leqslant L(2\alpha\lg(C_3L)+C_2) \leqslant L(C_4\lg(C_5A)+C_2),$$

by (3.28). Finally then

$$L \geqslant \frac{A}{C_4\lg(C_5A)+C_2} = \frac{A}{C_6\lg(C_7A)}.$$

Applying in our most recent bound for M gives

$$\begin{aligned} M &\leqslant 1+\left(1+\frac{\alpha}{\lg c}\right)\sqrt{\frac{C_8\lg(C_7A)}{A}}\left(\frac{C_9N\lg^2(C_7A)}{A}\right)^{C_{10}A} \\ &\leqslant 1+C_{11}\sqrt{\frac{C_8\lg(C_7\lg(s/2))}{\lg(s/2)}}\left(\frac{C_9N\lg^2(C_7\lg(s/2))}{\lg(s/2)}\right)^{C_{10}\lg(s/2)} \\ &\leqslant 1+C_{11}\sqrt{\frac{C_8\lg(C_7\lg(s/2))}{\lg(s/2)}}\left(\frac{C_{12}s\lg^2(C_7\lg(s/2))}{\lg(s/2)}\right)^{C_{13}\lg(s/2)}, \end{aligned}$$

since $N=\lfloor(s/(4c))^{1/2\alpha}\rfloor$.

Now we consider

$$\lim_{s\to\infty}\frac{\lg^2(\lg(s))}{s\lg(s)}=\lim_{t\to\infty}\frac{\lg^2(t)}{t2^t}.$$

Two applications of L'Hosptial's rule give show that this limit is zero, and therefore

$$\frac{C_{12}s\lg^2(C_7\lg(s/2))}{\lg(s/2)}^{C_{13}\lg(s/2)}\leqslant C_{c,\alpha}s^{\gamma_{c,\alpha}\log(s)},$$

for $C_{c,\alpha},\gamma_{c,\alpha}$ large enough. The same holds for the constant 1 and the square root term, and therefore

$$M\leqslant C_{c,\alpha}s^{\gamma_{c,\alpha}\log(s)}$$

as desired.

Exponential weights: Again, since exponential weights force monotonically growing ω_ν in ν and n, we will have that M is finite and therefore all $\nu\in\Lambda_0$ are supported in $[N]$. Calculating with the same strategy as before leads us to $\sigma^N=b_N\leqslant\sqrt{s/4}$ and so $N=\left\lfloor\frac{\lg(s/4)}{2\lg(\sigma)}\right\rfloor$. For exponential weights, $a_n=2\lg(b_n)=2n\lg(\sigma)$. Now, we again note that for any $|S|=k$,

$$2\lg\sigma\sum_{n\in S}n=\|a_S\|_1\leqslant A-k\implies k\leqslant\sqrt{\frac{A}{\lg\sigma}}$$

by using that $\sum_{n\in S}n\geqslant\sum_{n=1}^{k}n=(n^2+n)/2\geqslant n^2/2$. Additionally,

$$\prod_{n\in S}a_n^{-1}\leqslant(2\lg\sigma)^{-k}(k!)^{-1}.$$

Thus, by a similar reasoning to the previous argument,

$$M\leqslant 1+\sum_{k=1}^{\left\lfloor\sqrt{\frac{A}{\lg\sigma}}\right\rfloor}\left(\frac{A}{\lg\sigma}\right)^k\frac{1}{(k!)^2}\binom{N}{k}.$$

Applying the reasoning that $k^2 \leqslant A/\lg\sigma$ for all k being summed over allows us to bound the exponential and factorial terms as above, giving

$$\begin{aligned} M &\leqslant 1 + \left(\frac{A}{\lg\sigma}\right)^{\sqrt{A/\lg\sigma}} \frac{1}{(\lfloor A/\lg\sigma\rfloor)^2} \sum_{k=1}^{\left\lfloor\sqrt{\frac{A}{\lg\sigma}}\right\rfloor} \binom{N}{k} \\ &\leqslant 1 + \left(\frac{A}{\lg\sigma}\right)^{\sqrt{A/\lg\sigma}} \frac{C_\sigma}{2\pi\sqrt{A/\lg(\sigma)}(e^2 A/\lg(\sigma))^{\sqrt{A/\lg(\sigma)}}} \left(\frac{eN}{\sqrt{A/\lg(\sigma)}}\right)^{\sqrt{A/\lg(\sigma)}} \\ &= 1 + \frac{C_\sigma}{2\pi\sqrt{A/\lg(\sigma)}} \left(\frac{e^3 N}{\sqrt{A/\lg(\sigma)}}\right)^{\sqrt{A/\lg(\sigma)}}, \end{aligned}$$

by using Stirling's formula asymptotically. Noting that $A/\lg(\sigma) = \log_\sigma(s/2)$ and

$$\frac{N}{\sqrt{\log_\sigma(s/2)}} \leqslant \frac{\log_\sigma(s/4)}{\sqrt{\log_\sigma(s/2)}} \leqslant \sqrt{\log_\sigma(s/2)},$$

we have

$$M \leqslant 1 + \frac{C_\sigma}{2\pi\sqrt{\log_\sigma(s/2)}} \left(e^3\sqrt{\log_\sigma(s/2)}\right)^{\sqrt{\log_\sigma(s/2)}},$$

as desired.

□

Note that $K \gtrsim s\log(M)\log^3(s)$ is satisfied when instead of $\log(M)$, we use the logarithm of the upper bounds in Corollary 3.1. Additionally, with this logarithm substituted for $\log(M)$, taking K asymptotically equal to this value allows us to similarly rephrase the recovery guarantees in Theorem 3.3 to remove dependence on M. This final example explores these new bounds for the cases of weights given in Corollary 3.1.

Example 3.4.

Constant weights: *Rewriting the conditions in* (3.22) *as conditions on* s, *we see that*

$$M \leqslant \begin{cases} \left(\log_{\sigma^2}(\sigma^2 s/2)\right)^N, & \textit{if } s \geqslant 2(2\sigma^2)^N \\ \left(\left(1 + \frac{1}{\lg(\sigma^2)}\right) eN\right)^{\log_{2\sigma^2}(s/2)} & \textit{if } s < 2(2\sigma^2)^N. \end{cases}$$

Thus, when $s \geqslant 2(2\sigma^2)^N$,

$$\begin{aligned} \log(M) &\leqslant N\log\left(\log_{\sigma^2}(s/2) + 1\right) \\ &\leqslant N\log(C_\sigma \log(s)) \\ &\leqslant C_\sigma N \log\log(s). \end{aligned}$$

When $s < 2(2\sigma^2)^N$,

$$\begin{aligned} \log(M) &\leqslant \log_{2\sigma^2}(s/2)\log\left(1 + \frac{1}{\lg(\sigma^2)} eN\right) \\ &\leqslant C_\sigma \log(s)\log(N). \end{aligned}$$

Thus, taking a number of measurements asymptotic to

$$K \asymp \begin{cases} s\log^4(s)\log(N), & \textit{if } s < 2(2\sigma^2)^N \\ sN\log^3(s)\log\log(s) & \textit{if } s \geqslant 2(2\sigma^2)^N, \end{cases}$$

will allow for the use of Theorem 3.3.

Polynomial weights: *For polynomial weights,* (3.23) *gives*

$$\log(M) \leqslant C_{\alpha,c}\log^2(s).$$

Thus, a number of measurements asymptotic to

$$K \asymp s\log^5(s)$$

suffices to apply Theorem 3.3. In this case, we can also solve for s *to rewrite the error bounds, since this number of measurements implies*

$$\frac{1}{s} \leqslant \frac{\log^5(s)}{K} \leqslant \frac{\log^5(K)}{K}.$$

Thus, the approximation $F^\sharp$ *to the functional applied to the true solution satisfies*

$$\left\|F^\sharp - F\right\|_\infty \leqslant B_F' \left(\frac{\log^5(K)}{K}\right)^{1/p-1}$$

$$\left\|F^\sharp - F\right\|_2 \leqslant C_F' \left(\frac{\log^5(K)}{K}\right)^{1/p-1/2}.$$

Exponential weights: *In the case of exponential weights,* (3.24) *shows that the number of points in the truncated index set* M *actually grows slower than* s*, and will therefore be overtaken for some* s *large enough. Indeed, we can bound* M *in this case by*

$$M \leqslant C_1 \log(s)^{C_2\sqrt{\log(s)}}.$$

Considering the limit

$$\lim_{s\to\infty} \frac{\log(s)^{C\sqrt{\log(s)}}}{s} = \lim_{t\to\infty} \frac{t^{C\sqrt{t}}}{\exp(t)} = \exp\Big(\lim_{t\to\infty} \sqrt{t}(C\log(t) - \sqrt{t})\Big) = 1,$$

since $\log(t) = o(\sqrt{t})$*, we thus see that* M *grows slower than* s *asymptotically. In particular, this means that the number of measurements* $K \geqslant s$ *overtakes the number of degrees of freedom in our approximation, leading to an over-determined system, in which case there may be more efficient (e.g. least squares) reconstruction algorithms.*

Remark 3.6. *Recall that in Theorem 1.8, we saw that the error in an anisotropic sparse grid approximation of the true solution decays algebraically in the number of measurements at a rate depending on the sum of the sizes of the regions of analyticity of the true solution. As shown above, the decay in error for polynomial weights in the compressive sensing Petrov-Galerkin algorithm satisfies a similar structure, up to logarithmic factors, where the exponent only depends on* p *rather than the entire sum of the regions of analyticity. However, by the discussion in Example 3.3, when the norms* $\beta_{0,n}$ *of the nominal operators decay polynomially, the value of* p *is directly determined by the rate of this polynomial decay which also determines the* δ*-admissible sequences and sizes of analyticity regions. Thus, the rate in both cases are a direct consequence of the sums of the sizes of the analyticity regions (or equivalently the summability of the operator norms of the original affine parametric operator).*

CHAPTER 4

Alternative Reconstruction Methods

4.1. A Summary of Errors

In previous chapters, we have focused our attention on using a small number (relative to the length of the gPC expansion) of parametric measurements of some solution $u(Z)$ to an affine parameterized PDE (or more generally, any compressible function with a valid gPC expansion) to produce an accurate approximation. For simplicity, in the remainder of these notes, we consider u to be a functional applied to the solution so that samples are complex numbers rather than elements in $\mathcal{X}(\Omega)$. The general strategy, e.g., the one outlined in Algorithm 10, is to use some numerical approximation of (a functional of) the solution (e.g., a Petrov-Galerkin discretization) at fixed parameter values $u_{N,h}(Z^{(k)})$ as noisy samples of the solution $u(Z^{(k)})$ which we further regard as noisy samples of a truncated version of the solution $u_{\Lambda_0}(Z^{(k)})$.

If we denote y_k the value of the measurement $u_{N,h}(Z^{(k)})$, we have

$$y_k = u(Z^{(k)}) + e^{\text{disc}} = u_{\Lambda_0}(Z^{(k)}) + e^{\text{trunc}} + e^{\text{disc}} = A\hat{u}_{\Lambda_0} + e^{\text{trunc}} + e^{\text{disc}}.$$

We then use a constrained weighted ℓ_1 minimization problem

$$\underset{z\in\mathbb{C}^M}{\text{minimize}} \, \|z\|_{\omega,1} \text{ subject to } \|Az - y\|_2 \leqslant \left\|e^{\text{trunc}} + e^{\text{disc}}\right\|_2 \leqslant \left\|e^{\text{trunc}}\right\|_2 + \left\|e^{\text{disc}}\right\|_2 \tag{4.1}$$

to reconstruct an accurate approximation of the truncated coefficient vector of the solution depending on the size of these errors and sparsity parameters.

Though we can choose $\left\|e^{\text{disc}}\right\|_2$ to be as small as desired (assuming a convergent numerical PDE solution scheme), as mentioned in Remark 3.5, estimating $\left\|e^{\text{trunc}}\right\|_2$ requires a priori knowledge of the solution and its behavior on $\Lambda_R := \Lambda \setminus \Lambda_0$. Quantitatively, for $u \in \ell_{\omega,p}$, by virtue of the Stechkin estimate, we have shown in (3.16) that

$$\left\|e^{\text{trunc}}\right\|_2 \leqslant 2^{1/p}\sqrt{K}s^{1/2-1/p}\|u_{\Lambda_R}\|_{\omega,p}.$$

Thus, choosing either s to be large or the weights so that Λ_0 captures the majority of the behavior of u, we can suppose that this error is small. However, without more detailed information on the parametric operator or the solution itself, we cannot provide an accurate estimate of $\|u_{\Lambda_R}\|_{\omega,p}$, and therefore cannot effectively constrain the weighted ℓ_1 minimization problem.

Before considering alternative unconstrained methods which allow for minimization with respect to unknown noise, we introduce a further type of noise. Since we can group $e^{\text{disc}} + e^{\text{trunc}} =: e^{\text{bounded}}$ as errors which are pervasive throughout each sample but are known to be bounded (by the previous discussion), we now consider further error which may be arbitrarily unbounded but sparse. This type of error is referred to as *corruption error*. As an example, we see that the task of obtaining the samples $u_{N,h}(Z^{(k)})$ is entirely parallelizable. Performing these calculations in a distributed setting, we could consider the case of a faulty node which returns nonsense (or perhaps nothing at all) for some small number of these parallel computations. Thus, we expect arbitrarily large errors in the sparse subset of measurements which were dispatched to this faulty node. Of course, we may not know ahead of time which computations were sent to this node, and so the support of this error vector is unknown. We will henceforth denote this error as e^{sparse}, giving the total measurement error of the truncated solution as $e = e^{\text{bounded}} + e^{\text{sparse}}$. For the remainder of these notes, to simplify our presentation of alternative minimization programs and recovery bounds, we assume our measurement error in the truncated solution is normalized, that is $e \leftarrow \frac{1}{\sqrt{K}}e$. With this new definition of e, letting $\tilde{y} = \frac{1}{\sqrt{K}}y$ and $\tilde{A} = \frac{1}{\sqrt{K}}A$ be the normalized samples of u and sampling matrix, e is defined to satisfy

$$e := \tilde{y} - \tilde{A}\hat{u}_{\Lambda_0}. \tag{4.2}$$

4.2. Lower Set Sparsity

In Chapter 2, we truncated our infinitely indexed gPC expansions to sets only of the form $\Lambda_0 = \{\nu \in \Lambda \mid 2\omega_\nu^2 \leqslant s\}$ in order to maintain generality and ensure that weighted sparse recovery results hold. In Chapter 3, we found explicit weights in terms of the affine operator and discussed the resulting truncation index sets. However, in Section 4.3, we will be introducing alternatives to the constrained ℓ_1 minimization routine we have previously considered for reconstruction of the coefficients from measurements. In order to provide a consistent and better defined setting in which to compare these methods, we narrow our perspective with regard to truncation as in [**1**] by supposing that the support of the gPC coefficients has further structure. For this, we introduce lower sets, and return to the hyperbolic cross index set.

Definition 4.1. *Any index set* $S \subseteq \mathbb{N}_0^N$ *is a* lower set *if it is downward closed, that is, for all* $\nu \in S$, *and* $\eta \in \mathbb{N}_0^N$, *if* $\eta \leqslant \nu$, *then* $\eta \in S$ *as well. We define the* error in the best s-term approximation to a vector x in lower sets *as*

$$\sigma_{L,s}(x)_{\omega,p} = \inf_{\substack{z:\|z\|_0\leqslant s\\ \text{supp}(z)\ \textit{lower}}} \|x - z\|_{\omega,p}.$$

Note here that the infimum is taken over traditionally sparse vectors rather than weighted sparse vectors, whereas the norm is still weighted.

Proposition 4.1. *Let* Λ_0 *be the hyperbolic cross index set* (1.32) *indexed at the sparsity level* s, *that is*

$$\Lambda_0 = \Lambda_{HC}(s) = \left\{\nu \in \mathbb{N}_0^N \mid \prod_{n=1}^{N}(\nu_n + 1) \leqslant s\right\}.$$

Then

$$\Lambda_0 = \bigcup_{\substack{|S|\leqslant s\\ S\ \textit{lower}}} S.$$

Proof. For any $\nu \in \mathbb{N}_0^N$, we define the rectangular block

$$\mathcal{R}_\nu := \{\eta \in \mathbb{N}_0^N \mid \eta \leqslant \nu\}.$$

Note that $\mathcal{R}_\nu$ is a lower set and

$$|\mathcal{R}_\nu| = \prod_{n=1}^{N}(\nu_n + 1).$$

Then for any $\nu \in \Lambda_0$, we know that $|\mathcal{R}_\nu| \leqslant s$, and therefore $\nu \in \mathcal{R}_\nu$ which is a subset of the union of lower sets of cardinality bounded by s. Now, if we have any element $\nu \in S$ where S is a lower set with $|S| \leqslant s$, we must have $\mathcal{R}_\nu \subseteq S$ by the fact that S is lower. Thus, $|\mathcal{R}_\nu| \leqslant |S| \leqslant s$, and so $\nu \in \Lambda_0$ completing the proof. □

For the remainder of the notes, we will restrict our attention to coefficient vectors indexed on the hyperbolic cross so as to account for all sparse vectors indexed on lower sets, letting $|\Lambda_{\mathrm{HC}}(s)| = M$, and reindexing $\Lambda_{\mathrm{HC}}(s) = [M]$ when convenient. There is strong motivation for the assumption of lower set sparsity in the context of solving parametric PDE, as the Legendre series for solutions of problems similar and more extensive (e.g., operators exhibiting nonlinearities in the parameters or random boundary conditions) to those considered in these notes were shown to have this sparsity in [**10**] using arguments similar to those in Chapter 3. We now briefly outline the flow of Chapter 2 in terms of compressive sensing on the hyperbolic cross to sense vectors with lower set sparsity so as to be able to link to alternative reconstruction algorithms. Additionally, we will only consider gPC expansions with Chebyshev and Legendre bases. As such, in the following, our weight sequence ω will always be taken to be $\omega_\nu = \|\phi_\nu\|_\infty$ for ϕ_ν the ν-tensor product Chebyshev or Legendre polynomial.

Definition 4.2. *The* one-dimensional order j (normalized) Legendre polynomial *is the* j*th orthonormal polynomial basis function calculated from applying Gram-Schmidt orthonormalization to the polynomial basis of univariate Taylor monomials* $\{z_n^j\}_{j=0}^{\infty}$ *with the under the inner product* $L^2_\pi([-1,1])$ *with uniform probability density function* $\pi_n(z_n) = \frac{1}{2}$*. For any* $\nu \in \mathbb{N}_0^N$*, the* tensorized Legendre polynomial *is defined as*

$$\phi_\nu(Z) = \prod_{n=1}^{N} \phi_{\nu_n}(Z_n)$$

where ϕ_{ν_n} *is the order* ν_n *univariate Legendre polynomial.*

Proposition 4.2 (Properties of Legendre Polynomials). *The tensorized Legendre polynomials are orthonormal with respect to uniform measure on* $\Gamma = [-1,1]^N$*. Additionally, the tensorized Legendre polynomials have* $\|\phi_\nu\|_\infty = \prod_{n=1}^{N} \sqrt{2\nu_n + 1}$ *with equality attained at* $Z = 1$*, that is* $\phi_\nu(1) = \prod_{n=1}^{N} \sqrt{2\nu_n + 1}$*.*

Proof. Any reference on orthogonal polynomials e.g., [**7**, Appendix, A.4]. □

Definition 4.3 ([**1**], Definition 5.2). *The* intrinsic lower sparsity of order s *is defined as the maximum weighted cardinality of lower sets with cardinality bounded by* s*, that is,*

$$\mathcal{S}(s) := \max\{\omega(S) \mid S \subseteq \mathbb{N}_0^N,\ |S| \leqslant s,\ S \textit{ is lower}\}.$$

Note that by Proposition 4.1, it suffices to consider just the maximum over subsets of the hyperbolic cross.

Definition 4.4 ([**1**], Definition 5.3, cf. Definition 2.1). *A matrix* $A \in \mathbb{C}^{K,M}$ *is said to satisfy the* lower robust null space property of order s with constants $\rho \in (0,1)$ and $\tau > 0$ *if*

$$\|v_S\|_2 \leqslant \frac{\rho}{\sqrt{\mathcal{S}(s)}}\|v_{S^c}\|_{\omega,1} + \tau\|Av\|_2 \textit{ for all } v \in \mathbb{C}^M \textit{ with } S \subseteq [M] \textit{ with } \omega(S) \leqslant \mathcal{S}(s)\ ,$$

that is, it satisfies the weighted robust null space property of order $\mathcal{S}(s)$*.*

Corollary 4.1 ([**1**], Theorem 5.6, cf. Lemma 2.1). *If* $A \in \mathbb{C}^{K,M}$ *satisfies the lower robust null space property of order* s *with constants* $\rho \in (0,1)$ *and* $\tau > 0$*, then for all* $x, z \in \mathbb{C}^M$*, we have*

$$\|x - z\|_{\omega,1} \leqslant \frac{1+\rho}{1-\rho}\left(\|z\|_{\omega,1} - \|x\|_{\omega,1} + 2\sigma_{L,s}(x)_{\omega,1}\right) + \frac{2\tau\sqrt{\mathcal{S}(s)}}{1-\rho}\|A(x-z)\|_2. \tag{4.3}$$

Proof. Since A satisfies the lower robust null space property of order s, it satisfies the weighted robust null space property of order $\mathcal{S}(s)$. Additionally, we must have $\sigma_{\mathcal{S}(s)}(x)_{\omega,1} \leqslant \sigma_{L,s}(x)_{\omega,1}$, since being lower is a stricter feasibility condition on the index sets considered in these quantities. An application of Lemma 2.1 finishes the $\ell_{\omega,1}$ estimate. □

We cannot directly use the ℓ_2 bound in Lemma 2.1 as $\mathcal{S}(s)$ is not necessarily larger than $2\|\omega_{\Lambda_0}\|_\infty^2$. However, we can use the following bounds to prove a similar result. Additionally, this bound will be able to relate our notion of the weighted restricted isometry property implying the weighted null space property to lower sets with less stringent requirements on the sparsity.

Lemma 4.1 ([**11**], Lemma 4.1). *For $s \geqslant 2$, for Chebyshev polynomials, the intrinsic sparsity is bounded below by*

$$\mathcal{S}_T(s) \geqslant \frac{3}{2}\|\omega_{\Lambda_0}\|_\infty^2, \tag{4.4}$$

and for Legendre polynomials, the intrinsic sparsity is bounded below by

$$\mathcal{S}_L(s) \geqslant \frac{4}{3}\|\omega_{\Lambda_0}\|_\infty^2. \tag{4.5}$$

In particular, for either basis, $\mathcal{S}(s) - \|\omega_{\Lambda_0}\|_\infty^2 \geqslant \frac{1}{4}\mathcal{S}(s)$. Additionally, we have the bounds

$$\mathcal{S}_T(3s) \geqslant 3\mathcal{S}_T(s), \qquad \mathcal{S}_L(2s) \geqslant 3\mathcal{S}_L(s). \tag{4.6}$$

Proof. First note that for any weight, since any degree univariate Legendre and Chebyshev polynomials attain their maximum at the same point, we have $\omega_\nu = \|\phi_\nu\|_\infty = \prod_{n=1}^N \|\phi_{\nu_n}\|_\infty$. Now for any index $\nu \in \Lambda_0 \setminus \{0\}$ (since we already know $\frac{3}{2}\omega_0^2 = \frac{3}{2} \leqslant 2 = s \leqslant \mathcal{S}(s)$ since all weights are larger than one), we have $\mathcal{R}_\nu \subseteq \Lambda_0$ and $|\mathcal{R}_\nu| \leqslant s$ since Λ_0 is the degree s hyperbolic cross. Thus, $\mathcal{S}_T(s) \geqslant \omega(\mathcal{R}_\nu)$. We can directly calculate $\omega(\mathcal{R}_\nu)$ with Chebyshev weights by the factoring argument, letting $q(\eta_n) = 2 - \delta_{\eta_n,0}$,

$$\begin{aligned}
\omega(\mathcal{R}_\nu) &= \sum_{\eta \leqslant \nu} \omega_\eta^2 \\
&= \sum_{\eta_1=0}^{\nu_1} \cdots \sum_{\eta_N=0}^{\nu_N} 2^{\|\eta\|_0} \\
&= \sum_{\eta_1=0}^{\nu_1} \cdots \sum_{\eta_N=0}^{\nu_N} q(\eta_1)\cdots q(\eta_N) \\
&= \sum_{\eta_1=0}^{\nu_1} q(\eta_1) \cdots \sum_{\eta_N=0}^{\nu_N} q(\eta_N) \\
&= \prod_{n=1}^N (1 + 2\nu_n) \\
&\geqslant \prod_{n=1}^N (3 - 2\delta_{\nu_n,0}) \\
&= 3^{\|\nu\|_0}.
\end{aligned}$$

Thus, we have the bound

$$\mathcal{S}_T(s) \geqslant 3^{\|\nu\|_0} = \left(\frac{3}{2}\right)^{\|\nu\|_0} \omega_\nu^2 \geqslant \frac{3}{2}\omega_\nu^2.$$

Since this applies to all ω_ν^2 with $\nu \in \Lambda_0$, (4.4) follows.

For Legendre polynomials, the only difference is the choice of $q(\eta_n) = 2\nu_n + 1$. The above argument gives

$$\mathcal{S}_L(s) \geqslant \omega(\mathcal{R}_\nu) = \prod_{n=1}^{N} \sum_{\eta_n=0}^{\nu_n} (2\eta_n + 1) = \prod_{n=1}^{N} (\nu_n + 1)^2$$

by expanding the sum and simplifying. Additionally, since $(\nu_n + 1)^2 \geqslant \frac{4}{3}(2\nu_n + 1)$ for all $\nu_n \geqslant 1$ (which can be checked by expanding and calculating derivatives) and $(\nu_n + 1)^2 \geqslant (2\nu_n + 1)$ for $\nu_n = 0$, we have

$$\mathcal{S}_L(s) \geqslant \left(\frac{4}{3}\right)^{\|\nu_0\|} \prod_{n=1}^{N} (2\nu_n + 1) = \left(\frac{4}{3}\right)^{\|\nu\|_0} \omega_\nu^2 \geqslant \frac{4}{3}\omega_\nu^2.$$

Again this uniform upper bound over Λ_0 implies (4.5).

In order to prove (4.6), we start by considering any lower $S \subseteq \Lambda_0$ with $|S| \leqslant s$. Letting any $\nu \in S$ be rewritten $\nu = (\nu_1, \tilde{\nu})$, we expand S to a lower set S' of cardinality $2|S|$ by defining $S' = \{(2\nu_1, \tilde{\nu}), (2\nu_1 + 1, \tilde{\nu}) \mid \nu \in S\}$. This one-to-two mapping of S to S' directly gives that $|S'| = 2|S|$. Additionally, if $\nu \in S'$ and $\eta \leqslant \nu$, we have that $\tilde{\eta} \leqslant \tilde{\nu}$. Additionally, ν_1 can be written as either $2\nu_1'$ or $2\nu_1' + 1$ such that $(\nu_1', \tilde{\nu}) \in S$. Thus, taking $\eta_1' = \lfloor \eta_1/2 \rfloor$, we must have $\eta_1' \leqslant \nu_1'$ and therefore $(\eta_1', \eta) \in S$. Thus, $\eta \in S'$, and S' is lower.

Now, for Legendre polynomials, consider

$$\omega(S') = \sum_{\nu \in S} \sqrt{2(2\nu_1) + 1}^2 \omega_{\tilde{\nu}}^2 + \sum_{\nu \in S} \sqrt{2(2\nu_1 + 1) + 1}^2 \omega_{\tilde{\nu}}^2 = \sum_{\nu \in S} 4\sqrt{2\nu_1 + 1}^2 \omega_{\tilde{\nu}}^2 = 4 \sum_{\nu \in S} \omega_\nu^2.$$

Thus,

$$\mathcal{S}_L(2s) \geqslant \omega(S') = 4\omega(S) \geqslant 3\omega(S)$$

for any $S \subseteq \Lambda$ with $|S| \leqslant s$, giving $\mathcal{S}_L(2s) \geqslant 3\mathcal{S}_L(s)$.

For Chebyshev polynomials, we repeat the process with the expanded set $S' = \{(3\nu_1, \tilde{\nu}), (3\nu_1 + 1, \tilde{\nu}), (3\nu_1 + 2, \tilde{\nu}) \mid \nu \in S\}$ so that $|S'| = 3|S|$. A similar analysis shows that S' is lower and

$$\omega(S') = \sum_{\nu \in S} (\omega_{3\nu_1}^2 + \omega_{3\nu_1+1}^2 + \omega_{3\nu_1+2}^2)\omega_{\tilde{\nu}}^2 \geqslant 3 \sum_{\nu \in S} \omega_{\nu}^2 = 3\omega(S),$$

since a univariate Chebyshev polynomial satisfies $\|\phi_j\|_\infty = \sqrt{2}$ if $j \neq 0$ and $\|\phi_j\|_\infty = 1$ if $j = 0$. Thus $\mathcal{S}_T(3s) \geqslant 3\mathcal{S}_T(s)$. □

Lemma 4.2 ([1], Lemma 5.8). *If $A \in \mathbb{C}^{K,M}$ satisfies the lower robust null space property of order s,*

$$\|x - z\|_2 \leqslant \frac{2 + \rho}{\sqrt{\mathcal{S}(s)}} \|x - z\|_{\omega,1} + \tau \|A(x - z)\|_2.$$

Proof. Letting $v = x - z$, fix S with $\omega(S) \leqslant \mathcal{S}(s)$ so that $\|v - v_S\|_2 = \|v_{S^c}\|_2 = \sigma_{\mathcal{S}(s)}(v)_{\omega,2}$. Then by the lower robust null space property,

$$\|v_S\|_2 \leqslant \frac{\rho}{\sqrt{\mathcal{S}(s)}} \|v_{S^c}\|_{\omega,1} + \tau \|Av\|_2 \leqslant \frac{\rho}{\sqrt{\mathcal{S}(s)}} \|v\|_{\omega,1} + \tau \|Av\|_2.$$

By the weighted Stechkin estimate, Theorem 1.5, we also have

$$\|v_{S^c}\|_2 = \sigma_{\mathcal{S}(s)}(v)_{\omega,2} \leqslant \frac{1}{\sqrt{\mathcal{S}(s) - \|\omega_{\Lambda_0}\|_\infty^2}} \|v\|_{\omega,1} \leqslant \frac{2}{\sqrt{\mathcal{S}(s)}} \|v\|_{\omega,1},$$

where the second inequality follows from Lemma 4.1. Summing these two bounds gives the desired inequality. □

As mentioned previously, the weighted null space property (and therefore, the lower null space property) is difficult to check for a matrix. Thus, we move up the chain, and define a lower restricted isometry property which implies the lower null space property.

Definition 4.5 ([1], Definition 5.2, cf. Definition 2.2). *For $A \in \mathbb{C}^{K,M}$, $s \geqslant 2$, the* lower restricted isometry constant $\delta_{L,s}$ *for* A *is the smallest number for which*

$$(1-\delta_{L,s})\|x\|_2^2 \leqslant \|Ax\|_2^2 \leqslant (1+\delta_{L,s})\|x\|_2^2$$

for all $x \in \mathbb{C}^M$ *with* $\|x\|_{\omega,0} \leqslant \mathcal{S}(s)$. *We say that* A *satisfies the* lower restricted isometry property *(lower RIP) if* $\delta_{L,s}$ *is sufficiently small. Note that this is equivalent to* A *satisfying the weighted restricted isometry property of order* $\mathcal{S}(s)$.

Theorem 4.1 ([1], Lemma 5.4, cf. Theorem 2.4). *Let* $A \in \mathbb{C}^{K,M}$ *have lower RIP constant*

$$\delta_{L,\alpha s} \leqslant \frac{1}{5},$$

where $\alpha = 2$ *for Legendre intrinsic weights and* $\alpha = 3$ *for Chebyshev intrinsic weights. Then* A *has the robust lower null space property of order* s *with constants* $\tau = \sqrt{1+\delta_{L,\alpha s}}/(1-\delta_{L,\alpha s})$ *and* $\rho = 4\delta_{L,\alpha s}/(1-\delta_{L,\alpha s})$.

Proof. It suffices to show that assuming A has the ω-RIP of order $\mathcal{S}(\alpha s)$, it has the ω-NSP of order $\mathcal{S}(s)$. First, by (4.6), $3\mathcal{S}(s) \leqslant \mathcal{S}(\alpha s)$, and so $\delta_{\omega,3\mathcal{S}(s)} \leqslant \delta_{\omega,\mathcal{S}(\alpha s)} = \delta_{L,\alpha s}$. By Lemma 4.1, we only have that $\mathcal{S}(s) \geqslant \frac{4}{3}\|\omega_{\Lambda_0}\|_\infty^2$, so we cannot directly use Theorem 2.4. However, a repetition of the proof with this bound replaces the factor of two in (2.9) by four. Everything else carries through with the new NSP constant $\rho = 4\delta_{\omega,3\mathcal{S}(s)}/(1-\delta_{\omega,3\mathcal{S}(s)})$. Replacing $\delta_{\omega,3\mathcal{S}(s)}$ by $\delta_{L,\alpha s}$ everywhere and noting that the resulting ρ is bounded by one when $\delta_{L,\alpha s} \leqslant 1/5$ finishes the theorem. □

We now report the necessary number of samples for the sampling matrix to have the lower RIP with high probability. Due to its complexity, we do not consider the proof.

Theorem 4.2 ([1], Theorem 5.5, cf. Theorem 2.8). *Fix* $\delta, \gamma \in (0,1)$, *and let* $(\phi_\nu)_{\nu \in \Lambda_0}$ *be the tensorized Chebyshev or Legendre basis on the hyperbolic cross index set. Take the intrinsic weight sequence* $\omega_\nu = \|\phi_\nu\|_\infty$ *and*

$$K \gtrsim \mathcal{S}(s) L(s, M, \delta, \gamma)$$

i.i.d. sampling points $\{Z^{(k)}\}_{k=1}^K$ *drawn from the orthogonalization measure* π *of the basis, where* $L(s,M,\delta,\gamma)$ *is the polylogarithmic factor*

$$L = \frac{1}{\delta^2}\log\left(\frac{\mathcal{S}(s)}{\delta^2}\right)\max\left\{\frac{1}{\delta^4}\log\left(\frac{\mathcal{S}(s)}{\delta^2}\log\left(\frac{\mathcal{S}(s)}{\delta^2}\right)\right)\log(M), \frac{1}{\delta}\log\left(\frac{1}{\delta\gamma}\log\left(\frac{\mathcal{S}(s)}{\delta^2}\right)\right)\right\}. \tag{4.7}$$

With probability exceeding $1-\gamma$, *the normalized sampling matrix* $\tilde{A} \in \mathbb{C}^{K,M}$ *with entries* $\tilde{A}_{k,\nu} = \frac{1}{\sqrt{K}}\phi_\nu(Z^{(k)})$ *has lower RIP constant of order* s *satisfying* $\delta_{L,s} \leqslant \delta$.

We finish this section with some asymptotically tight bounds on the intrinsic sparsities associated to intrinsic Chebyshev and Legendre weights which will be useful in providing better quantitative bounds on the number of measurements as well as recovery estimates.

Lemma 4.3 ([1], Lemma 2.2). *Let* $2 \leqslant s \leqslant 2^{N+1}$. *We have*

$$s^\kappa/4 \leqslant \mathcal{S}(s) \leqslant s^\kappa,$$

where

$$\kappa = \frac{\log(3)}{\log(2)}, \qquad \kappa = 2 \tag{4.8}$$

for Chebyshev and Legendre weights respectively. The upper bound holds for all $s \geqslant 2$.

Proof. We prove the upper bound following the argument in [**9**, Lemma 3.1]. We begin with Legendre weights. It suffices to prove the upper bound

$$\omega(S) \leqslant |S|^2 \leqslant s^2$$

for all lower S with $|S| \leqslant s$. We proceed by induction on s. When $s = 1$, the bound is trivially true since the lower set S must equal $\{0\}$, and for Legendre polynomials, $\omega_0^2 = 1$. Now supposing the bound is true for all lower S' with $|S'| \leqslant s$, we consider S lower with $|S| = s + 1$. Since $|S| > 1$, we know that there must be some nonzero multiindex in S. We assume without loss of generality that there is some $\nu \in S$ with $\nu_1 \neq 0$. Now, we define $0 < K = \max_{\nu \in S} \nu_1 \leqslant |S|$, and we decompose S into its slices in ν_1, that is, we write

$$S = \bigsqcup_{\nu_1=0}^{K} \{(\nu_1, \tilde{\nu}) \mid (\nu_1, \tilde{\nu}) \in S\}. \tag{4.9}$$

We shift each slice down to zero in the first dimension, and analyze these slices, defining

$$S_{\nu_1} = \{(0, \tilde{\nu}) \mid (\nu_1, \tilde{\nu}) \in S\}.$$

By shifting down to zero, we ensure that S_{ν_1} is a lower set. Additionally, $|S_{\nu_1}| < |S|$ by (4.9) and the fact that $K > 0$. And finally, since S is lower, each of the slices must be contained in the slice one lower, that is, $S_{\nu_1} \subseteq S_{\nu_1 - 1}$ and therefore $|S_{\nu_1}| \leqslant |S_{\nu_1 - 1}|$. We now rewrite

$$\omega(S) = \sum_{\nu_1=0}^{K} \sum_{\tilde{\nu}:(\nu_1,\tilde{\nu})\in S} \omega_{\nu_1}^2 \omega_{\tilde{\nu}}^2 = \sum_{\nu_1=0}^{K} (2\nu_1 + 1)\omega(S_{\nu_1})^2 \leqslant \sum_{\nu_1=0}^{K} (2\nu_1 + 1)|S_{\nu_1}|^2, \tag{4.10}$$

where the last inequality is by the inductive hypothesis. By the fact that

$$\nu_1 |S_{\nu_1}| = \sum_{j=0}^{\nu_1 - 1} |S_{\nu_1}| \leqslant \sum_{j=0}^{\nu_1 - 1} |S_j|$$

by the downwards nestedness of the S_{ν_1}, we have

$$\begin{aligned}
\omega(S) &\leqslant \sum_{\nu_1=0}^{K} |S_{\nu_1}|^2 + 2 \sum_{\nu_1=1}^{K} \sum_{j=0}^{\nu_1-1} |S_{\nu_1}||S_j| \\
&= \sum_{\nu_1=0}^{K} \sum_{j=\nu_1} |S_{\nu_1}||S_j| + \sum_{\nu_1=1}^{K} \sum_{j=0}^{\nu_1-1} |S_{\nu_1}||S_j| + \sum_{\nu_1=0}^{K} \sum_{j=\nu_1+1}^{K} |S_{\nu_1}||S_j| \\
&= \sum_{\nu_1=0}^{K} \sum_{j=0}^{K} |S_{\nu_1}||S_j| \\
&= \left(\sum_{\nu_1=0}^{K} |S_{\nu_1}| \right)^2 .
\end{aligned}$$

Since $|S_{\nu_1}|$ is the same cardinality as the slices in (4.9), we know that $\sum_{\nu_1}^{K} |S_{\nu_1}| = |S|$, and so $\omega(S) \leqslant |S|^2$ as desired.

For Chebyshev weights, we again define $q(\nu_1) = 2 - \delta_{\nu_1,0}$ so that $\omega_\nu^2 = \|\phi_\nu\|_\infty^2 = \prod_{n=1}^N q(\nu_n)$. The same analysis holds (with the modified inductive hypothesis) up until (4.10), where we obtain

$$\begin{aligned}\omega(S) &\leqslant \sum_{\nu_1=0}^K q(\nu_1)\omega(S_{\nu_1})^2 \\ &\leqslant \sum_{\nu_1=0}^K q(\nu_1)|S_{\nu_1}|^{\log(3)/\log(2)} \\ &= |S_0|^{\log(3)/\log(2)} + 2\sum_{\nu_1=1}^K |S_{\nu_1}|^{\log(3)/\log(2)}.\end{aligned}$$

We now prove the claim [**9**, Proposition 3.2], that for any $a_0 \geqslant a_1 \geqslant \ldots \geqslant a_K$ and any $\alpha \geqslant \log(3)/\log(2)$, we have

$$a_0^\alpha + 2(a_1^\alpha + \ldots + a_K^\alpha) \leqslant (a_0 + \ldots + a_K)^\alpha,$$

which finishes the proof of the upper bound. Indeed, we proceed inductively on K with the case $K = 0$ holding trivially. The trickiest case is showing that it is true for $K = 1$, that is, showing that

$$a_0^\alpha + 2a_1^\alpha \leqslant (a_0 + a_1)^\alpha.$$

The insight comes from rearranging, where we instead try and show

$$2a_1^\alpha \leqslant (a_0 + a_1)^\alpha - a_0^\alpha,$$

and treating the right hand size like a function of a_0. Taking derivatives shows that this function is increasing in a_0. Attempting relate back to $2a_1^\alpha$, we plug in a_1 to find

$$(a_1 + a_1)^\alpha - a_1^\alpha \leqslant (a_0 + a_1)^\alpha - a_0^\alpha.$$

Now the left hand side gives $(2^\alpha - 1)a_1^\alpha$, which is exactly greater than $2a_1^\alpha$ when $2^\alpha \geqslant 3$, that is, $\alpha \geqslant \log(3)/\log(2)$. Now assuming the bound holds for values less than some arbitrary K, we find

$$\begin{aligned}(a_0 + \ldots + a_K)^\alpha &\geqslant (a_0 + \ldots + a_{K-1})^\alpha + 2a_K^\alpha \\ &\geqslant a_0^\alpha + 2(a_1^\alpha + \ldots + a_{K-1}^\alpha) + 2a_K^\alpha \\ &= a_0^\alpha + 2(a_1^\alpha + \ldots + a_K^\alpha),\end{aligned}$$

completing the proof of the claim.

For the lower bounds we proceed as in [**11**, Lemma 3.6]. In the proof of Lemma 4.1, our factoring argument showed that for any $\nu \in \Lambda_0$,

$$\omega(\mathcal{R}_\nu) = \prod_{n=1}^N (1 + 2\nu_n), \quad \omega(\mathcal{R}_\nu) = \prod_{n=1}^N (\nu_n + 1)^2$$

for Chebyshev and Legendre weights respectively. If $s \leqslant 2^{N+1}$, we can fit s in a dyadic interval $2^{N'} \leqslant s \leqslant 2^{N'+1}$ with $N' \leqslant N$, and choose the multiindex $\nu = e_1 + \ldots + e_N'$ which has $|\mathcal{R}(\nu)| = 2^{N'} \leqslant s$. Thus, for Chebyshev weights

$$\mathcal{S}_T(s) \geqslant \omega(\mathcal{R}_\nu) = 3^{N'} = 2^{N'\log(3)/\log(2)} \geqslant \left(\frac{s}{2}\right)^{\log(3)/\log(2)} = \frac{s^{\log(3)/\log(2)}}{3} \geqslant \frac{s^\kappa}{4}.$$

For Legendre weights,

$$\mathcal{S}_L(s) \geqslant \omega(\mathcal{R}_\nu) = 2^{2N'} \geqslant \left(\frac{s}{2}\right)^2 = \frac{s^\kappa}{4},$$

as desired. □

4.3. Overview of Alternative Minimization Algorithms

We now outline some alternatives to the weighted ℓ_1 minimization problem constrained by bounds on the truncation, discretization, or corruption error as considered in [1]. In the following, $\eta \geqslant 0$ and $\lambda > 0$ are tuning parameters which are not directly determined by the problem. To simplify the recovery bounds and minimization program, we account for normalization in our samples, that is we let $\tilde{A} \in \mathbb{C}^{K,M}$ be the normalized sampling matrix $\tilde{A}_{k,\nu} = \frac{1}{\sqrt{K}}\phi_\nu(Z^{(k)})$ and $\tilde{y}$ be the normalized samples of u, $\tilde{y}_k = \frac{1}{\sqrt{K}}u(Z^{(k)})$.

Weighted Quadratically-Constrained Basis Pursuit (WQCBP):

$$\hat{u}^\sharp = \arg\min_{z\in\mathbb{C}^M} \|z\|_{\omega,1} \text{ such that } \left\|\tilde{A}z - \tilde{y}\right\|_2 \leqslant \eta. \tag{4.11}$$

This algorithm is the same as constrained weighted ℓ_1 minimization. However, note that $\eta \geqslant 0$ is chosen to be arbitrary rather than determined by the errors. Thus, we expect bounds to depend on the decoupled η and $\|e\|_2$.

Weighted LASSO (WLASSO):

$$\hat{u}^\sharp = \arg\min_{z\in\mathbb{C}^M} \|z\|_{\omega,1} + \lambda\left\|\tilde{A}z - \tilde{y}\right\|_2^2. \tag{4.12}$$

WLASSO removes the constraint in WQCBP by instead using it as a penalty term.

Weighted Square-Root LASSO (WSR-LASSO):

$$\hat{u}^\sharp = \arg\min_{z\in\mathbb{C}^M} \|z\|_{\omega,1} + \lambda\left\|\tilde{A}z - \tilde{y}\right\|_2. \tag{4.13}$$

The only difference here from WLASSO is that the penalty term is the standard ℓ_2-norm (which may result in more difficult theoretical analysis) rather than its square.

Weighted LAD-LASSO (WLAD-LASSO):

$$\hat{u}^\sharp = \arg\min_{z\in\mathbb{C}^M} \|z\|_{\omega,1} + \lambda\left\|\tilde{A}z - \tilde{y}\right\|_1. \tag{4.14}$$

Here, the penalty is considered in the ℓ_1-norm. Since basis pursuit has been shown to be effective for sparse recovery, we anticipate that this method will be oriented towards the sparse error case.

4.4. Analysis of Alternative Minimization Algorithms

We now show that the algorithms in Section 4.4 provide accurate reconstructions of our solution u assuming the tools from Section 4.2. These proofs will take the form of the proof of Theorem 2.2, however, since we are not using basis pursuit constrained by the error for reconstruction, the recovery bounds for weighted ℓ_1 minimization given the ω-NSP in Theorem 2.3 do not apply. We will instead refine the proof of this theorem in the case of each of the recovery algorithms, with Corollary 4.1 taking the place of Lemma 2.1.

4.4.1. Weighted Quadratically-Constrained Basis Pursuit.

Theorem 4.3 ([1], Theorem 5.10). *Let $0 < \gamma < 1$, $0 < \delta \leqslant 1/5$, $2 \leqslant s \leqslant 2^{N+1}$, and $\Lambda_0 = \Lambda_{HC}(s)$ with $\{\phi_\nu\}_{\nu\in\Lambda_0}$ tensor Legendre or Chebyshev polynomial bases. If we draw*

$$K \gtrsim s^\kappa L(s, n, \delta, \gamma),$$

i.i.d. measurements $\{Z^{(k)}\}_{k=1}^K$ from the orthogonalization measure $\pi\,dz$ for κ as in (4.8) and L as in (4.7), then with probability $1-\gamma$, the following holds. For any $\eta \geqslant 0$, letting $\hat{u}^\sharp$ be the *solution of the WQCBP problem* (4.11)*, defining* $u^\sharp = \sum_{\nu \in \Lambda_0} \hat{u}^\sharp_\nu \phi_\nu$*,*

$$\begin{aligned} \|u - u^\sharp\|_\infty &\lesssim \sigma_{L,s}(u)_{\omega,1} + s^{\kappa/2}\left(\eta + \|e\|_2 + \mathcal{T}\right) \\ \|u - u^\sharp\|_2 &\lesssim s^{-\kappa/2}\sigma_{L,s}(u)_{\omega,1} + \eta + \|e\|_2 + \mathcal{T} + \|u_{\Lambda_R}\|_2, \end{aligned}$$

where

$$\mathcal{T} = \mathcal{T}(A, \Lambda_0, e, \omega, \eta) := \min\left\{ \frac{\|z\|_{\omega,1}}{s^{\kappa/2}} \mid z \in \mathbb{C}^M,\ \|\tilde{A}z - e\|_2 \leqslant \eta \right\},$$

and all implicit constants depend on δ*.*

Proof. For the assumed number of samples, by our upper bounds on intrinsic sparsity in Lemma 4.1, we know that the measurements satisfy the criterion for $\tilde{A}$ to have lower RIP constant bounded by δ with probability at least $1 - \gamma$ by Theorem 4.2. Since $\delta \leqslant 1/5$ the lower NSP holds by Theorem 4.1, and therefore the distance bound (4.3) from Corollary 4.1 holds.

In this distance bound, we let $x = \hat{u}_{\Lambda_0}$ and $z = \hat{u}^\sharp$ and work with three separate pieces. For the first, we note

$$\sigma_{L,s}(\hat{u}_{\Lambda_0})_{\omega,1} = \sigma_{L,s}(u)_{\omega,1},$$

since minimizing over cardinality s lower sets is equivalent to minimizing over cardinality s lower sets contained in the hyperbolic cross by Proposition 4.1. For the second, we find

$$\sqrt{\mathcal{S}(s)}\|\tilde{A}(\hat{u}_{\Lambda_0} - \hat{u}^\sharp)\|_2 \leqslant s^{\kappa/2}\left(\|e\|_2 + \|\tilde{y} - \tilde{A}\hat{u}^\sharp\|_2\right) \leqslant s^{\kappa/2}\left(\|e\|_2 + \eta\right),$$

by the upper bounds on intrinsic sparsity in Lemma 4.3 and the fact that $\hat{u}^\sharp$ is in the feasible set of the optimization problem defined by the constraint $\|\tilde{y} - \tilde{A}z\|_2 \leqslant \eta$. Finally, we are left to bound

$$\|\hat{u}^\sharp\|_{\omega,1} - \|\hat{u}_{\Lambda_0}\|_{\omega,1}.$$

But by the fact that $\hat{u}^\sharp$ solves (4.11), we may consider $\|\hat{u}_{\Lambda_0}\|_{\omega,1}$ subtracted from the objective function without changing the minimum. By the reverse triangle inequality, we may rewrite

$$\begin{aligned} \|\hat{u}^\sharp\|_{\omega,1} - \|\hat{u}_{\Lambda_0}\|_{\omega,1} &\leqslant \min\{\|z - \hat{u}_{\Lambda_0}\|_{\omega,1} \mid z \in \mathbb{C}^M\ \|\tilde{A}z - \tilde{y}\|_2 \leqslant \eta\} \\ &= \min\{\|z\|_{\omega,1} \mid z \in \mathbb{C}^M\ \|\tilde{A}z - e\|_2 \leqslant \eta\} \\ &= s^{\kappa/2}\mathcal{T}. \end{aligned}$$

Using our trick of bounding the infinity norm of gPC expansions by weighted ℓ_1 norms, we have

$$\begin{aligned} \|u - u^\sharp\|_\infty &\leqslant \|\hat{u} - \hat{u}^\sharp\|_{\omega,1} \\ &= \|\hat{u}_{\Lambda_0} - \hat{u}^\sharp\|_{\omega,1} + \|\hat{u}_{\Lambda_R}\|_{\omega,1} \\ &\lesssim \sigma_{L,s}(\hat{u}_{\Lambda_0})_{\omega,1} + \|\hat{u}^\sharp\|_{\omega,1} - \|\hat{u}_{\Lambda_0}\|_{\omega,1} + \sqrt{\mathcal{S}(s)}\|\tilde{A}(\hat{u}_{\Lambda_0} - \hat{u}^\sharp)\|_2 + \|u_{\Lambda_R}\|_{\omega,1} \\ &\leqslant \sigma_{L,s}(u)_{\omega,1} + s^{\kappa/2}\left(\mathcal{T} + \|e\|_2 + \eta\right) + \|u_{\Lambda_R}\|_{\omega,1}. \end{aligned} \tag{4.15}$$

Noting that the best lower set approximation to u of cardinality s must be contained within the hyperbolic cross, we know that $\|u_{\Lambda_R}\|_{\omega,1} \leqslant \sigma_{L,s}(u)_{\omega,1}$ finishing the L^∞ bound.

For the L^2 bound, we use Lemma 4.2 and Parseval's identity giving

$$\|u - u^\sharp\|_2 \leqslant \|u_{\Lambda_0} - u^\sharp\|_2 + \|u_{\Lambda_R}\|_2 \lesssim s^{-\kappa/2}\|\hat{u}_{\Lambda_0} - \hat{u}^\sharp\|_{\omega,1} + \|\tilde{A}(\hat{u}_{\Lambda_0} - \hat{u}^\sharp)\|_2 + \|u_{\Lambda_R}\|_2.$$

The inequalities used to prove the L^∞ bound in (4.15) then give

$$\|u - u^\sharp\|_2 \leqslant s^{-\kappa/2}\sigma_{L,s}(u)_{\omega,1} + \|e\|_2 + \eta + \mathcal{T} + \|u_{\Lambda_R}\|_2,$$

as desired. □

We now provide a bound for $\mathcal{T}$ in terms of properties of the sampling matrix, the parameters involved, and the relationship between the true error and our "proxy error" in η.

THEOREM 4.4 ([1], Theorem 5.11). *For $K \asymp s^{\kappa}L$ as in Theorem 4.3, the bound for the constant*

$$\mathcal{T} \lesssim \frac{s^{\alpha/2}\sqrt{L}}{\sigma_K\left(\sqrt{\frac{K}{M}}\tilde{A}^*\right)}\max\{\|e\|_2 - \eta, 0\}$$

holds, where $\alpha = 1$ or $\alpha = 2$ for Chebyshev and Legendre weights respectively.

PROOF. An upper bound for $\mathcal{T}$ follows by bounding $s^{-\kappa/2}\|z\|_{\omega,1}$ for some feasible $z \in \mathbb{C}^M$ satisfying $\left\|\tilde{A}z - e\right\|_2 \leqslant \eta$. If we first consider the case where $\tilde{A}$ is full rank, we can make use of the pseudoinverse $\tilde{A}^\dagger = \tilde{A}^*(\tilde{A}\tilde{A}^*)^{-1}$. Choosing $z = \tilde{A}^\dagger\left(1 - \frac{\eta}{\|e\|_2}\right)e$ ensures that $\left\|\tilde{A}z - e\right\|_2 = \eta$. Assuming that our number of measurements K is asymptotically equal to $s^\kappa L$, we can bound $s^{-\kappa/2} \lesssim \frac{\sqrt{L}}{\sqrt{K}}$. Thus, $\mathcal{T} \lesssim \frac{\sqrt{L}}{\sqrt{K}}\|z\|_{\omega,1}$ for our choice of z. Instead of directly working with the weighted ℓ_1 norm, we use Cauchy-Schwarz to bound

$$\|z\|_{\omega,1} \leqslant \sqrt{\sum_{\nu\in\Lambda_0}\omega_\nu^2}\|z\|_2 \leqslant \sqrt{\omega(\Lambda_0)}\left\|\tilde{A}^\dagger\right\|\left|\|e\|_2 - \eta\right|.$$

We can bound the operator norm of the pseudoinverse by a singular value decomposition of $\tilde{A}^* = U\Sigma V^*$, a giving

$$\tilde{A}^\dagger = U\Sigma V^*\left(V\Sigma^* U^* U\Sigma V^*\right)^{-1} = U\Sigma^{-1}V^*,$$

where if $\Sigma = \operatorname{diag}(\sigma_k(\tilde{A}^*))_{k=1}^K$, the largest diagonal element of Σ^{-1} and therefore the largest singular value and operator norm of $\tilde{A}^\dagger$ is $\sigma_K(\tilde{A}^*)^{-1}$.

It remains to bound $\omega(\Lambda_0)$ which we will do in terms of the size of the hyperbolic cross $|\Lambda_0| = M$. For Chebyshev polynomials,

$$\omega(\Lambda_0) = \sum_{\nu\in\Lambda_0} 2^{\|\nu\|_0} \leqslant \sum_{\nu\in\Lambda_0}\prod_{n=1}^N(\nu_n + 1) \leqslant \sum_{\nu\in\Lambda_0} s = Ms,$$

where we use that $\prod_{n=1}^N(\nu_n+1) \leqslant s$ for all $\nu \in \Lambda_0$ by the definition of the hyperbolic cross. For Legendre polynomials,

$$\omega(\Lambda_0) = \sum_{\nu\in\Lambda_0}\prod_{n=1}^N(2\nu_n+1) \leqslant \sum_{\nu\in\Lambda_0}\prod_{n=1}^N(\nu_n+1)^2 \leqslant \sum_{\nu\in\Lambda_0} s^2 = Ms^2.$$

Thus, $\sqrt{\omega(\Lambda_0)} \leqslant \sqrt{M}s^{\alpha/2}$ with $\alpha = 1$ or $\alpha = 2$ for Chebyshev and Legendre weights respectively. Putting it all together, we find

$$\mathcal{T} \lesssim \frac{\sqrt{L}}{\sqrt{K}}\sqrt{M}s^{\alpha/2}\sigma_K(\tilde{A}^*)^{-1}\left|\|e\|_2 - \eta\right| = \frac{s^{\alpha/2}\sqrt{L}}{\sigma_K\left(\sqrt{\frac{K}{M}}\tilde{A}^*\right)}\left|\|e\|_2 - \eta\right|$$

But of course, when $\|e\|_2 \leqslant \eta$, the trivial choice of $z = 0$ satisfies $\left\|\tilde{A}z - e\right\|_2 \leqslant \eta$, and so $\mathcal{T} = 0$. Additionally, when $\tilde{A}$ is not full rank, $\sigma_K(\tilde{A}^*) = 0$, and the same bound holds from the full rank case making the definition $1/0 = \infty$ (though this bound is obviously not very informative; however, the rank zero case is probabilistically rare, see Lemma 4.4 below). Thus, we can say that in general,

$$\mathcal{T} \lesssim \frac{s^{\alpha/2}\sqrt{L}}{\sigma_K\left(\sqrt{\frac{K}{M}}\tilde{A}^*\right)}\max\{\|e\|_2 - \eta, 0\}.$$

□

Thus, recovery bounds act as in the case where the constraint is determined by the error with an added term depending on how close η is chosen to $\|e\|_2$ with a factor depending polynomially on the sparsity, the polylogarithmic factor in the number of measurements, and the smallest singular value of $\sqrt{\frac{K}{M}}\tilde{A}^*$. For $N = 1$, it can be shown that this singular value stays bounded away from zero with high probability [8]. In the following lemma, we show at least that the smallest singular value of the expectation of the sampling matrix has a nice closed form which heuristically supports this assertion.

Lemma 4.4 ([3], Lemma 3). *For a gPC basis,*

$$\sigma_K\left(\sqrt{\frac{K}{M}}\mathbb{E}\tilde{A}^*\right) = \sqrt{1 - \frac{1}{M}}.$$

Proof. We determine $\lambda_{\min}(\frac{K}{M}\mathbb{E}\tilde{A}\tilde{A}^*)$ and the equality for the singular value results by taking square roots. Calculating the entries of the mean matrix

$$\left(\mathbb{E}\left[\frac{K}{M}\tilde{A}\tilde{A}^*\right]\right)_{k,j} = \frac{1}{M}\sum_{\nu\in\Lambda_0}\mathbb{E}\phi_\nu(Z^{(k)})\overline{\phi_\nu(Z^{(j)})} = \begin{cases} 1, & \text{if } k = j \\ \frac{1}{M} & \text{if } k \neq j \end{cases}$$

where the latter case holds by noting that $\mathbb{E}\phi_\nu = \mathbb{E}\phi_\nu\phi_0 = \delta_{\nu,0}$. This circulant matrix is diagonalized by the discrete Fourier transform matrix which provides the closed form for the eigenvalues

$$\lambda_k = 1 + \frac{1}{M}\sum_{j=1}^{K-1}\exp\left(i\frac{2\pi jk}{K}\right), \quad k = 1, \dots K.$$

When $k = K$, $\lambda_K = 1 + \frac{1}{M}(K-1)$. When $k \neq K$, a standard geometric series allows us to evaluate the sum to be -1. Thus, $\lambda_k = 1 - \frac{1}{M}$ for all $k \neq K$, and therefore $\lambda_{\min}(\mathbb{E}\frac{K}{M}\tilde{A}\tilde{A}^*) = 1 - \frac{1}{M}$ as desired. □

4.4.2. Weighted LASSO. The results for WLASSO are similar under a certain class of parameter values λ. However, the parameter value still depends on the actual value of the error.

Theorem 4.5 ([1], Theorem 5.13). *Let* $0 < \gamma < 1$, $0 < \delta \leqslant 1/5$, $2 \leqslant s \leqslant 2^{N+1}$, *and* $\Lambda_0 = \Lambda_{HC}(s)$ *with* $\{\phi_\nu\}_{\nu\in\Lambda_0}$ *tensor Legendre or Chebyshev polynomial bases. If we draw*

$$K \gtrsim s^\kappa L(s, n, \delta, \gamma)$$

i.i.d. measurements $\{Z^{(k)}\}_{k=1}^K$ *from the orthogonalization measure* $\pi\,\mathrm{d}z$ *for* κ *as in* (4.8) *and* L *as in* (4.7)*, then with probability* $1-\gamma$*, the following holds. For any*

$$\lambda = \theta\frac{\sqrt{\mathcal{S}(s)}}{\|e\|_2}, \quad \theta > 0,$$

letting $\hat{u}^\sharp$ *be the solution of the WLASSO problem* (4.12)*, defining* $u^\sharp = \sum_{\nu\in\Lambda_0}\hat{u}^\sharp_\nu\phi_\nu$*,*

$$\|u - u^\sharp\|_\infty \lesssim \sigma_{L,s}(u)_{\omega,1} + s^{\kappa/2}\|e\|_2$$
$$\|u - u^\sharp\|_2 \lesssim s^{-\kappa/2}\sigma_{L,s}(u)_{\omega,1} + \|e\|_2 + \|u_{\Lambda_R}\|_2,$$

where all implicit constants depend on δ *and* θ.

Proof. As in the proof of the recovery results for WQCBP, we know that the distance bound (4.3) holds with probability exceeding $1-\gamma$. As before then,

$$\|u-u^\sharp\|_\infty \lesssim \sigma_{L,s}(u)_{\omega,1} + \|\hat{u}^\sharp\|_{\omega,1} - \|\hat{u}_{\Lambda_0}\|_{\omega,1} + \sqrt{\mathcal{S}(s)}\|\tilde{A}(\hat{u}_{\Lambda_0}-\hat{u}^\sharp)\|_2. \tag{4.16}$$

A more appropriate accounting of the last term by inserting $\tilde{y}$ and applying the triangle inequality gives

$$\|\tilde{A}(\hat{u}_{\Lambda_0}-\hat{u}^\sharp)\|_2 \leqslant \|e\|_2 + \|\tilde{y}-\tilde{A}\hat{u}^\sharp\|_2.$$

Now, we rewrite the solution to the WLASSO minimization problem as the constrained problem

$$(\hat{u}^\sharp, e^\sharp) = \arg \min_{(z,d)\in\mathbb{C}^M\times\mathbb{C}^K} \|z\|_{\omega,1} + \lambda\|d\|_2^2 \text{ such that } \tilde{A}z + d = \tilde{y}, \tag{4.17}$$

which then gives

$$\sqrt{\mathcal{S}(s)}\|\tilde{A}(\hat{u}_{\Lambda_0}-\hat{u}^\sharp)\|_2 \leqslant \sqrt{\mathcal{S}(s)}\|e\|_2 + \sqrt{\mathcal{S}(s)}\|e^\sharp\|_2. \tag{4.18}$$

We will want to use the fact that $\hat{u}^\sharp$ and $e^\sharp$ solve (4.17) to get (4.16) in terms of only the error and the error in the best lower s-sparse estimate, and so we attempt to write (4.16) in terms of the objective function applied to this minimizer.

In light of this approach, we try to separate $\sqrt{\mathcal{S}(s)}$ from $\|e^\sharp\|_2$ in (4.18) while introducing λ. Cauchy's inequality with constants does the job, with

$$\sqrt{\mathcal{S}(s)}\|e^\sharp\|_2 = 2\frac{\sqrt{\mathcal{S}(s)}}{2\sqrt{\lambda}}\sqrt{\lambda}\|e^\sharp\|_2 \leqslant \frac{\mathcal{S}(s)}{4\lambda} + \lambda\|e^\sharp\|_2^2. \tag{4.19}$$

Thus

$$\begin{aligned}\|u-u^\sharp\|_\infty &\lesssim \sigma_{L,s}(u)_{\omega,1} + \left(\|\hat{u}^\sharp\|_{\omega,1} + \lambda\|e^\sharp\|_2^2\right) - \|\hat{u}_{\Lambda_0}\|_{\omega,1} + \lambda^{-1}\mathcal{S}(s) + \sqrt{\mathcal{S}(s)}\|e\|_2\\ &\lesssim \sigma_{L,s}(u)_{\omega,1} + \lambda\|e\|_2^2 + \lambda^{-1}\mathcal{S}(s) + \sqrt{\mathcal{S}(s)}\|e\|_2.\end{aligned}$$

Choosing $\lambda \propto \frac{\sqrt{\mathcal{S}(s)}}{\|e\|}$ followed by the upper bounds on intrinsic sparsity in Lemma 4.3 give the desired L^∞ bound.

For the L^2 bound, making use of Lemma 4.2, as in WQCBP case and (4.19),

$$\begin{aligned}\|u-u^\sharp\|_2 &\lesssim \frac{1}{\sqrt{\mathcal{S}(s)}}\sigma_{L,s}(u)_{\omega,1} + \frac{1}{\sqrt{\mathcal{S}(s)}}\left(\|\hat{u}^\sharp\|_{\omega,1} - \|\hat{u}_{\Lambda_0}\|_{\omega,1}\right) + \|\tilde{A}(\hat{u}_{\Lambda_0}-\hat{u}^\sharp)\|_2 + \|u_{\Lambda_R}\|_2\\ &\lesssim \frac{1}{\sqrt{\mathcal{S}(s)}}\sigma_{L,s}(u)_{\omega,1} + \frac{1}{\sqrt{\mathcal{S}(s)}}\left(\|\hat{u}^\sharp\|_{\omega,1} - \|\hat{u}_{\Lambda_0}\|_{\omega,1}\right) + \|e\|_2 + \frac{\sqrt{\mathcal{S}(s)}\|e^\sharp\|_2}{\sqrt{\mathcal{S}(s)}} + \|u_{\Lambda_R}\|_2\\ &\lesssim \frac{1}{\sqrt{\mathcal{S}(s)}}\sigma_{L,s}(u)_{\omega,1} + \frac{1}{\sqrt{\mathcal{S}(s)}}\left(\|\hat{u}^\sharp\|_{\omega,1} + \lambda\|e^\sharp\|_2^2 - \|\hat{u}_{\Lambda_0}\|_{\omega,1}\right) + \lambda^{-1}\sqrt{\mathcal{S}(s)} + \|u_{\Lambda_R}\|_2\\ &\lesssim s^{-\lambda/2}\sigma_{L,s}(u)_{\omega,1} + \|e\|_2 + \|u_{\Lambda_R}\|_2,\end{aligned}$$

as desired. □

4.4.3. Weighted Square-root LASSO. The proof for WSR-LASSO proceeds much the same as WLASSO with the exception of the requirements for λ. Here, we note no dependence necessary on the measurement error in order to have similar recovery results to the previous cases. This lack of dependence on e in choosing λ is the main benefit of the WSR-LASSO methods over the previous two considered when attempting to reconstruct from measurements with completely unknown error.

Theorem 4.6 ([1], Theorem 5.14). *Let $0 < \gamma < 1$, $0 < \delta \leqslant 1/5$, $2 \leqslant s \leqslant 2^{N+1}$, and $\Lambda_0 = \Lambda_{HC}(s)$ with $\{\phi_\nu\}_{\nu \in \Lambda_0}$ tensor Legendre or Chebyshev polynomial bases. If we draw*

$$K \gtrsim s^\kappa L(s, n, \delta, \gamma),$$

i.i.d. measurements $\{Z^{(k)}\}_{k=1}^K$ from the orthogonalization measure $\pi\, dz$ for κ as in (4.8) *and* L *as in* (4.7)*, then with probability $1-\gamma$, the following holds. For any*

$$\lambda = \theta\sqrt{\mathcal{S}(s)}, \quad \theta \geqslant \frac{(5+\rho)\tau}{(1+\rho)(2+\rho)} = \max\left\{\frac{(5+\rho)\tau}{(1+\rho)(2+\rho)}, \frac{2\tau}{1+\rho}\right\},$$

(where $\rho < 1$ and τ are the constants in the lower NSP depending on δ), letting $\hat{u}^\sharp$ be the solution of the WSR-LASSO problem (4.13)*, defining $u^\sharp = \sum_{\nu \in \Lambda_0} \hat{u}^\sharp_\nu \phi_\nu$,*

$$\left\|u - u^\sharp\right\|_\infty \lesssim \sigma_{L,s}(u)_{\omega,1} + s^{\kappa/2}\|e\|_2$$
$$\left\|u - u^\sharp\right\|_2 \lesssim s^{-\kappa/2}\sigma_{L,s}(u)_{\omega,1} + \|e\|_2 + \|u_{\Lambda_R}\|_2,$$

where all implicit constants depend on δ and θ.

Proof. We now rewrite the WSR-LASSO minimization as the constrained problem

$$(\hat{u}^\sharp, e^\sharp) = \arg \min_{(z,d)\in\mathbb{C}^M\times\mathbb{C}^K} \|z\|_{\omega,1} + \lambda\|d\|_2 \text{ such that } \tilde{A}z + d = \tilde{y}.$$

In order to choose λ properly we more carefully consider the ℓ_1 error bound implied by the lower NSP (4.3) and split $\tilde{A}(\hat{u}_{\Lambda_0} - \hat{u}^\sharp)$ as in the previous proof, giving

$$\begin{aligned}\left\|\hat{u}_{\Lambda_0} - \hat{u}^\sharp\right\|_{\omega,1} &\leqslant C\sigma_{L,s}(u)_{\omega,1} + \frac{1+\rho}{1-\rho}\left(\left\|\hat{u}^\sharp\right\|_{\omega,1} - \|\hat{u}_{\Lambda_0}\|_{\omega,1}\right) + \frac{2\tau\sqrt{\mathcal{S}(s)}}{1-\rho}\left\|e^\sharp\right\|_2 + C\sqrt{\mathcal{S}(s)}\|e\|_2 \\ &\leqslant C\sigma_{L,s}(u)_{\omega,1} + \frac{1+\rho}{1-\rho}\left(\left\|\hat{u}^\sharp\right\|_{\omega,1} + \lambda\left\|e^\sharp\right\|_2 - \|\hat{u}_{\Lambda_0}\|_{\omega,1}\right) + C\sqrt{\mathcal{S}(s)}\|e\|_2 \\ &\lesssim \sigma_{L,s}(u)_{\omega,1} + s^{-\kappa/2}\|e\|_2,\end{aligned}$$

where the second inequality follows from the fact that $\lambda \geqslant \frac{2\tau\sqrt{\mathcal{S}(s)}}{1+\rho}$. Thus, the same bound holds (with a slightly different constant after accounting for the truncation) for $\left\|u - u^\sharp\right\|_\infty$.

For the L^2 case, Lemma 4.2 gives

$$\begin{aligned}&\left\|\hat{u}_{\Lambda_0} - \hat{u}^\sharp\right\|_2 \\ &\leqslant \frac{C}{\sqrt{\mathcal{S}(s)}}\sigma_{L,s}(u)_{\sigma,1} + \frac{(1+\rho)(2+\rho)}{(1-\rho)\sqrt{\mathcal{S}(s)}}\left(\left\|\hat{u}^\sharp\right\|_{\omega,1} - \|\hat{u}_{\Lambda_0}\|_{\omega,1}\right) + \frac{(5+\rho)\tau}{1-\rho}\left\|e^\sharp\right\|_2 + C\|e\|_2 \\ &\leqslant \frac{C}{\sqrt{\mathcal{S}(s)}}\sigma_{L,s}(u)_{\sigma,1} + \frac{(1+\rho)(2+\rho)}{(1-\rho)\sqrt{\mathcal{S}(s)}}\left(\left\|\hat{u}^\sharp\right\|_{\omega,1} + \lambda\left\|e^\sharp\right\|_2 - \|\hat{u}_{\Lambda_0}\|_{\omega,1}\right) + C\|e\|_2 \\ &\lesssim s^{-\kappa/2}\sigma_{L,s}(u)_{\sigma,1} + \|e\|_2,\end{aligned}$$

where again, the second inequality is due to the fact that

$$\lambda \geqslant \frac{(5+\rho)\tau\sqrt{\mathcal{S}(s)}}{(1+\rho)(2+\rho)}.$$

After combining with Parseval's identity on $\left\|u - u^\sharp\right\|_2 \leqslant \left\|u_{\Lambda_0} - u^\sharp\right\|_2 + \|u_{\Lambda_R}\|_2$, we recover the L^2 bound. □

4.4.4. Weighted LAD-LASSO. Since the ℓ_2 norm of the error bounds the error in approximating the solution using the previous three algorithms, these three algorithms should work well with $e = e^{\text{bounded}}$, that is, pervasive, but small errors in each measurement. However, for $e = e^{\text{bounded}} + e^{\text{sparse}}$, where $\|e^{\text{sparse}}\|_2$ is potentially large, these recovery bounds are not informative. Thus, we introduce the WLAD-LASSO method to perform basis pursuit on both the sparse sequence of coefficients of the original function, as well as on the (potentially) sparse error.

In this chapter, we have made extensive use of the standard compressed sensing workflow of showing that with high probability the sampling matrix has the lower RIP which implies the lower NSP which shows a weighted ℓ_1 bound (and associated ℓ_2 bound) in the distance between two vectors. These distance bounds are responsible for showing that weighted basis pursuit (in some sense) then gives acceptable error bounds. Now that we will be performing weighted basis pursuit in both the coefficient sequence and the error, we will need to introduce these same compressive sensing notions working on two vectors disjointly but simultaneously.

Definition 4.6 ([1], Definition 5.15). *Let* $M, m \in \mathbb{N}_+$*, and let* $x \in \mathbb{C}^{M+m}$ *with weight sequence* $\omega \in \mathbb{R}^{M+m}$ *partitioned as*

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad \omega = \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix},$$

with $x_1 \in \mathbb{C}^M$, $x_2 \in \mathbb{C}^m$, $\omega_1 \in \mathbb{R}^M$*, and* $\omega_2 \in R^m$*. Given a sparsity pair* $s = (s_1, s_2) \in \mathbb{R}^2$*, we say that* x is weighted s-sparse *if* $\|x_i\|_{\omega_i,0} \leqslant s_i$ *for* $i = 1, 2$.

Definition 4.7 ([1], Definition 5.16). *Let* $\lambda > 0$ *be a* scale*, and for* $\omega \in \mathbb{C}^{M+m}$ *partitioned as above,* $s \in \mathbb{R}^2$*, we define the* scaled weights ω_λ *and* scaled sparsity s_λ *as*

$$\omega_\lambda = \begin{bmatrix} \omega_1 \\ \lambda\omega_2 \end{bmatrix}, \quad s_\lambda = s_1 + \lambda^2 s_2.$$

The $\ell_{\omega_\lambda,1}$ error in the weighted best s-term approximation to x *is*

$$\sigma_s(x)_{\omega_\lambda,1} := \inf_{z:\|x_i\|_{\omega_i,0} \leqslant s_i} \|x - z\|_{\omega_\lambda,1}.$$

Remark 4.1. *The notion of scaled sparsity allows us to switch between two- and standard one-level scaled sparsity. As an example, for a sparsity pair* s*, if* x *is weighted* s*-sparse, then we can use our standard Cauchy-Schwarz argument on the weighted* ℓ_1 *norm summed only over the support to show that*

$$\begin{aligned} \|x\|_{\omega_\lambda,1} &= \sum_{\nu\in\text{supp}(x)} \omega_{\lambda,\nu}|x_\nu| \\ &\leqslant \sqrt{\sum_{\nu\in\text{supp}(x)} \omega_{\lambda,\nu}^2}\|x\|_2 \\ &\leqslant \sqrt{\sum_{\nu\in\text{supp}(x_1)} \omega_{1,\nu}^2 + \sum_{\nu\in\text{supp}(x_2)} \lambda^2\omega_{2,\nu}^2}\|x\|_2 \\ &\leqslant \sqrt{s_1 + \lambda^2 s_2}\|x\|_2 \\ &= \sqrt{s_\lambda}\|x\|_2. \end{aligned} \tag{4.20}$$

Definition 4.8 ([1], Definition 5.15, cf. Definition 2.1). *For a 2-level weight sequence* ω*, a matrix* $B \in \mathbb{C}^{K,M+m}$ *is said to satisfy the* 2-level weighted robust null space property of scale

$\lambda > 0$ and order $s = (s_1, s_2)$ *with constants* $\rho \in (0,1)$ *and* $\tau > 0$ *if*

$$\|v_{S_1 \cup S_2}\|_2 \leqslant \frac{\rho}{\sqrt{s_\lambda}} \left\| v_{(S_1 \cup S_2)^c} \right\|_{\omega_\lambda, 1} + \tau \|Bv\|_2$$

for all $v \in \mathbb{C}^{M+m}$ *and all* $S_1 \subseteq [M], S_2 \subseteq M + [m]$ *with* $\omega_1(S_1) \leqslant s_1$ *and* $\omega_2(S_2) \leqslant s_2$.

Lemma 4.5 ([1], Theorem 5.18, cf. Lemma 2.1). *If* $B \in \mathbb{C}^{K, M+m}$ *satisfies the 2-level weighted robust null space property of scale* $\lambda > 0$ *and order* $s = (s_1, s_2)$ *with constants* $\rho \in (0,1)$ *and* $\tau > 0$*, then for all* $x, z \in \mathbb{C}^{M+m}$*, we have*

$$\|x - z\|_{\omega_\lambda, 1} \leqslant \frac{1+\rho}{1-\rho} \left(\|z\|_{\omega_\lambda, 1} - \|x\|_{\omega_\lambda, 1} + 2\sigma_s(x)_{\omega_\lambda, 1} \right) + \frac{2\tau\sqrt{s_\lambda}}{1-\rho} \|B(x-z)\|_2, \tag{4.21}$$

and given $B(x-z) = 0$, $\|\omega_1\|_\infty^2 \leqslant \frac{3}{4}s_1$, *and* $\|\omega_2\|_\infty^2 \leqslant \frac{1}{2}s_2$, *we have*

$$\|x - z\|_2 \leqslant C_1 \frac{1+\sqrt{\Theta}}{\sqrt{s_\lambda}} \left(\|z\|_{\omega_\lambda, 1} - \|x\|_{\omega_\lambda, 1} + 2\sigma_s(x)_{\omega_\lambda, 1} \right), \tag{4.22}$$

with $C_1 = \frac{\max\{2\sqrt{\rho}, \rho\}(1+\rho)}{1-\rho}$ *and* $\Theta = \frac{\sqrt{s_\lambda}}{\min\{\sqrt{s_1}, \lambda\sqrt{s_2}\}}$.

Proof. The proof of (4.21) is exactly the same as the weighted ℓ_1 part of the 1-level Lemma 2.1 where we use (4.20) to make use of the 2-level weighted NSP.

In order to prove (4.22), we would like to mimic the proof in Lemma 2.1. However, when $Bv = 0$, we can use a small adjustment to the proof of the weighted Stechkin estimate, Theorem 1.5 (at the sacrifice of some generality) and the 2-level NSP to provide a tighter bound. We first show that for any vector y and weight sequence ξ with $\|\xi\|_\infty < s$, letting y_S realize the quasi-best s-term approximation to y (recall Definition 1.8), we have

$$\tilde{\sigma}_s(y)_{\xi,2}^2 = \|y_{S^c}\|_{\xi,2}^2 \leqslant \frac{1}{\sqrt{s - \|\xi\|_\infty^2}} \|y_S\|_2 \|y_{S^c}\|_{\xi,1}. \tag{4.23}$$

Following the proof of Theorem 1.5, we have

$$\begin{aligned} \|y_{S^c}\|_{\xi,2}^2 &= \sum_{\nu \notin S} |y_\nu|^2 \\ &\leqslant \sup_{\nu \notin S} |y_\nu| \xi_\nu^{-1} \sum_{\nu \notin S} |y_\nu| \xi_\nu \\ &= \sup_{\nu \notin S} |y_\nu| \xi_\nu^{-1} \|y_{S^c}\|_{\xi,1}. \end{aligned}$$

The improvement comes from taking a square root from the first term before introducing the weighted cardinality of S. Indeed,

$$\begin{aligned} \sqrt{\sup_{\nu \notin S} |y_\nu|^2 \xi_\nu^{-2}} &= \sqrt{\frac{1}{\xi(S)} \sum_{\eta \in S} \xi_\eta^2 \sup_{\nu \notin S} |y_\nu|^2 \xi_\nu^{-2}} \\ &\leqslant \frac{1}{\sqrt{\xi(S)}} \sqrt{\sum_{\eta \in S} |y_\eta|^2} \\ &= \frac{1}{\sqrt{\xi(S)}} \|y_S\|_2 \end{aligned}$$

by the fact that S contains the largest elements of the sequence of $|y_\nu|^2 |\xi_\nu|^{-2}$. The lower bound $s - \|\xi\|_\infty^2 \leqslant \xi(S)$ holds due to the fact that S was chosen to be the largest index set with $\xi(S) \leqslant s$ as in the original proof, giving (4.23).

Returning to the 2-level setup, define $\nu = x - z$. Letting S_1 with $\omega_1(S_1) \leqslant s_1$ and S_2 with $\omega_2(S_2) \leqslant s_2$ realize the quasi-best s_1- and s_2-term approximations to ν_1 and ν_2 respectively, we have

$$\|\nu\|_2 \leqslant \|\nu_{S_1 \cup S_2}\|_2 + \left\|\nu_{(S_1 \cup S_2)^C}\right\|_2. \tag{4.24}$$

Twice using (4.23) on the square of the second term, as well as the assumptions that $\|\omega_1\|_\infty^2 \leqslant 3s_1/4$ and $\|\omega_1\|_\infty^2 \leqslant s_2/2$ followed by the 2-level NSP gives

$$\begin{aligned}
\left\|\nu_{(S_1 \cup S_2)^C}\right\|_2^2 &= \left\|(\nu_1)_{S_1^C}\right\|_2^2 + \left\|(\nu_2)_{S_2^C}\right\|_2^2 \\
&\leqslant \frac{2}{\sqrt{s_1}} \|(\nu_1)_{S_1}\|_2 \left\|(\nu_1)_{S_1^C}\right\|_{\omega_1,1} + \frac{\sqrt{2}}{\sqrt{s_2}} \|(\nu_2)_{S_2}\|_2 \left\|(\nu_2)_{S_2^C}\right\|_{\omega_2,1} \\
&\leqslant \frac{2\left\|\nu_{(S_1 \cup S_2)^C}\right\|_{\omega_\lambda,1}}{\min\left\{\sqrt{s_1}, \lambda\sqrt{s_2}\right\}} \left(\|(\nu_1)_{S_1}\|_2 + \|(\nu_2)_{S_2}\|_2\right) \\
&\leqslant \frac{2\sqrt{2}\left\|\nu_{(S_1 \cup S_2)^C}\right\|_{\omega_\lambda,1}}{\min\left\{\sqrt{s_1}, \lambda\sqrt{s_2}\right\}} \|\nu_{S_1 \cup S_2}\|_2 \\
&\leqslant \frac{2\sqrt{2}}{\min\left\{\sqrt{s_1}, \lambda\sqrt{s_2}\right\}\sqrt{s_\lambda}} \left\|\nu_{(S_1 \cup S_2)^C}\right\|_{\omega_\lambda,1}^2,
\end{aligned}$$

where we have also used that $B\nu = 0$. Combining this bound with the 2-level NSP applied to the first part of (4.24) and defining the parameter $\Theta := \sqrt{s_\lambda}/\min\left\{\sqrt{s_1}, \lambda\sqrt{s_2}\right\}$, we obtain

$$\begin{aligned}
\|\nu\|_2 &\leqslant \frac{2\sqrt{\rho}\sqrt{\Theta} + \rho}{\sqrt{s_\lambda}} \left\|\nu_{(S_1 \cup S_2)^C}\right\|_{\omega_\lambda,1} \\
&\leqslant \frac{\max\left\{2\sqrt{\rho}, \rho\right\}(1 + \sqrt{\Theta})}{\sqrt{s_\lambda}} \|\nu\|_{\omega_\lambda,1}.
\end{aligned}$$

Applying (4.21) finally gives

$$\|x - z\|_2 \leqslant C_1 \frac{1 + \sqrt{\Theta}}{\sqrt{s_\lambda}} \left(\|z\|_{\omega_\lambda,1} - \|x\|_{\omega_\lambda,1} + 2\sigma_s(x)_{\omega_\lambda,1}\right),$$

with $C_1 = \frac{\max\{2\sqrt{\rho},\rho\}(1+\rho)}{1-\rho}$, as desired.

□

Definition 4.9 ([1], Definition 5.19, cf. Definition 2.2). *For* $B \in \mathbb{C}^{K,M+m}$*, a sparsity pair* $s = (s_1, s_2)$*, and a 2-level weight sequence* ω*, the* 2-level ω-RIP constant $\delta_{\omega,s}$ *for* A *is the smallest number for which*

$$(1 - \delta_{\omega,s})\|x\|_2^2 \leqslant \|Bx\|_2^2 \leqslant (1 + \delta_{\omega,s})\|x\|_2^2$$

for all $x \in \mathbb{C}^{M+m}$ *with* $\|x_i\|_{\omega_i,0} \leqslant s_i$ *for* $i = 1, 2$*. We say that* B *satisfies the* 2-level weighted restricted isometry property *if* $\delta_{\omega,s}$ *is sufficiently small.*

Theorem 4.7 ([1], Theorem 5.21, cf. Theorem 2.4). *Let* $B \in \mathbb{C}^{K,M+m}$ *have 2-level* ω*-RIP constant*

$$\delta_{\omega,3s} < \frac{1}{1 + 4\Theta}, \quad \Theta := \frac{\sqrt{s_\lambda}}{\min\left\{\sqrt{s_1}, \lambda\sqrt{s_2}\right\}}$$

(where Θ is as in Lemma 4.5) for $\frac{3}{4}s_1 \geqslant \|\omega_1\|_\infty^2$ and $\frac{1}{2}s_2 \geqslant \|\omega_2\|_\infty^2$. Then B has the 2-level weighted robust null space property of scale λ and order s with constants

$$\rho = \frac{4\delta_{\omega,3s}\Theta}{1-\delta_{\omega,3s}}, \quad \tau = \frac{\sqrt{1+\delta_{\omega,3s}}}{1-\delta_{\omega,3s}}.$$

Proof. The proof is again a repetition of the proof of Theorem 2.4, where we work over both levels simultaneously. We work out the details to clarify the introduction of Θ.

We first take some $v \in \mathbb{C}^{M+m}$ decomposed as $v = [x; e]$ and fix index sets $S \subseteq [M]$, $T \supseteq M + [m]$ with $\omega_1(S) \leqslant s_1$ and $\omega_2(T) \leqslant s_2$ In the vein of the proof of Theorem 2.4, we derive an ℓ_2 bound of v restricted to $S \cup T$ in terms of the weighted ℓ_1 norm by way of the non-increasing rearrangements of the sequences $|x_v|\omega_{1,v}^{-1}$ and $|e_v|\omega_{2,v}^{-1}$ respectively. We partition S^C and T^C reordered into the indices of these rearrangements into blocks $S_1, S_2, \ldots$ and $T_1, T_2, \ldots$ respectively, where all blocks (except possibly the first) are assumed to be largest possible satisfying $s_1 - \|\omega_1\|_\infty^2 \leqslant \omega_1(S_\ell) \leqslant s_1$ and $s_2 - \|\omega_2\|_\infty^2 \leqslant \omega_2(T_\ell) \leqslant s_2$. Under this setup, the same argument to derive (2.9) in the original theorem gives

$$\begin{aligned} \|x_{S_\ell}\|_2 &\leqslant \frac{\sqrt{s_1}}{s_1 - \|\omega_1\|_\infty^2}\|x_{S_{\ell-1}}\|_{\omega_1,1} \leqslant \frac{4}{\sqrt{s_1}}\|x_{S_{\ell-1}}\|_{\omega_1,1} \\ \|e_{T_\ell}\|_2 &\leqslant \frac{\sqrt{s_2}}{s_2 - \|\omega_2\|_\infty^2}\|e_{T_{\ell-1}}\|_{\omega_2,1} \leqslant \frac{2}{\sqrt{s_2}}\|e_{T_{\ell-1}}\|_{\omega_1,1}, \end{aligned} \tag{4.25}$$

where we have used our assumptions relating $\|\omega_i\|_\infty^2$ and s_i.

Applying the 2-level ω-RIP on $v_{S\cup T} + v_{S_1 \cup T_1}$ gives

$$\begin{aligned} \|v_{S\cup T} + v_{S_1\cup T_1}\|_2^2 &\leqslant \frac{1}{1-\delta_{\omega,2s}}\|B(v_{S\cup T} + v_{S_1\cup T_1})\|_2^2 \\ &= \frac{1}{1-\delta_{\omega,2s}}\langle B(v_{S\cup T} + v_{S_1\cup T_1}), Bv - \sum_{\ell\geqslant 2} Bv_{S_\ell\cup T_\ell}\rangle \\ &\leqslant \frac{\sqrt{1+\delta_{\omega,2s}}}{1-\delta_{\omega,2s}}\|v_{S\cup T} + v_{S_1\cup T_1}\|_2\|Bv\|_2 \\ &\quad + \frac{1}{1-\delta_{\omega,2s}}\sum_{\ell\geqslant 2}|\langle B(v_{S\cup T} + v_{S_1\cup T_1}), Bv_{S_\ell\cup T_\ell}\rangle|, \end{aligned} \tag{4.26}$$

where the last inequality follows by Cauchy-Schwarz and a second application of the 2-level ω-RIP. Using the fact that S, S_1, S_ℓ and T, T_1, T_ℓ are each collections of mutually disjoint index sets for $\ell \geqslant 2$ and therefore respective restrictions of x and e on these sets are orthogonal, rewriting $S' = S \cup S_1 \cup S_\ell$ and $T' = T \cup T_1 \cup T_\ell$,

$$\begin{aligned} |\langle B(v_{S\cup T} + v_{S_1\cup T_1}), Bv_{S_\ell\cup T_\ell}\rangle| &= |\langle B^*_{S'\cup T'}B_{S'\cup T'}(v_{S\cup T} + v_{S_1\cup T_2}), v_{S_\ell\cup T_\ell}\rangle + \langle v_{S\cup T} + v_{S_1\cup T_1}, v_{S_\ell\cup T_\ell}\rangle| \\ &= |\langle (B^*_{S'\cup T'}B_{S'\cup T'} - I)(v_{S\cup T} + v_{S_1\cup T_2}), v_{S_\ell\cup T_\ell}\rangle| \\ &\leqslant \delta_{\omega,3s}\|v_{S\cup T} + v_{S_1\cup T_1}\|_2\|v_{S_\ell\cup T_\ell}\|_2. \end{aligned}$$

Splitting $v_{S_\ell \cup T_\ell} = [x_{S_\ell}; e_{T_\ell}]$ and using (4.25), we find

$$\begin{aligned} \|v_{S_\ell \cup T_\ell}\|_2 &\leqslant \|x_{S_\ell}\|_2 + \|e_{T_\ell}\|_2 \\ &\leqslant \frac{4}{\sqrt{s_1}}\|x_{S_{\ell-1}}\|_{\omega_1,1} + \frac{2}{\sqrt{s_2}}\|e_{T_{\ell-1}}\|_{\omega_2,1} \\ &\leqslant \frac{4}{\min\left\{\sqrt{s_1}, \lambda\sqrt{s_2}\right\}}\|x_{S_{\ell-1}}; e_{T_{\ell-1}}\|_{\omega_\lambda,1} \\ &= \frac{4}{\min\left\{\sqrt{s_1}, \lambda\sqrt{s_2}\right\}}\|v_{S_{\ell-1}\cup T_{\ell-1}}\|_{\omega_\lambda,1}, \end{aligned}$$

giving

$$|\langle B(v_{S\cup T} + v_{S_1\cup T_1}), Bv_{S_\ell \cup T_\ell}\rangle| \leqslant \frac{4\delta_{\omega,3s}}{\min\left\{\sqrt{s_1}, \lambda\sqrt{s_2}\right\}}\|v_{S\cup T} + v_{S_1\cup T_1}\|_2 \|v_{S_{\ell-1}\cup T_{\ell-1}}\|_{\omega_\lambda,1}.$$

Combining with (4.26) and dividing by $\|v_{S\cup T} + v_{S_1\cup T_1}\|_2$ gives

$$\begin{aligned} \|v_{S\cup T}\|_2 &\leqslant \|v_{S\cup T} + v_{S_1\cup T_1}\|_2 \\ &\leqslant \frac{\sqrt{1+\delta_{\omega,3s}}}{1-\delta_{\omega,3s}}\|Bv\|_2 + \frac{4\delta_{\omega,3s}}{(1-\delta_{\omega,3s})\min\left\{\sqrt{s_1}, \lambda\sqrt{s_2}\right\}}\sum_{\ell \geqslant 2}\|v_{S_{\ell-1}\cup T_{\ell-1}}\|_{\omega_\lambda,1} \\ &= \frac{\sqrt{1+\delta_{\omega,3s}}}{1-\delta_{\omega,3s}}\|Bv\|_2 + \frac{4\delta_{\omega,3s}\Theta}{(1-\delta_{\omega,3s})\sqrt{s_\lambda}}\left\|v_{(S\cup T)^c}\right\|_{\omega_\lambda,1}. \end{aligned}$$

Thus, B satisfies the 2-level weighted robust null space property of scale λ and order s with the discussed constants, so long as

$$\frac{4\delta_{\omega_3 s}\Theta}{1-\delta_{\omega,3s}} < 1,$$

which is satisfied precisely when $\delta_{\omega,3s} < 1/(1+4\Theta)$ as desired. □

THEOREM 4.8 ([1], Theorem 5.23, cf. Theorems 2.8 and 4.2). *Fix $\delta, \gamma \in (0,1)$, and let $(\phi_\nu)_{\nu\in\Lambda_0}$ be the tensorized Chebyshev or Legendre basis on the hyperbolic cross index set. Take the intrinsic weight sequence $\omega_{1,\nu} = \|\phi_\nu\|_\infty$, let ω_2 be arbitrary, and take*

$$K \gtrsim \mathcal{S}(s_1)\max\left\{L(s_1, M, \delta, \gamma), \delta^{-2}s_2\right\}$$

i.i.d. sampling points $\{Z^{(k)}\}_{k=1}^K$ drawn from the orthogonalization measure π dz of the basis, where $L(s, M, \delta, \gamma)$ is the polylogarithmic factor (4.7). With probability exceeding $1-\gamma$, $B = [\tilde{A}, I] \in \mathbb{C}^{K, M+K}$ with $\tilde{A} \in \mathbb{C}^{K,M}$ the normalized sampling matrix with entries $\tilde{A}_{k,\nu} = \frac{1}{\sqrt{K}}\phi_\nu(Z^{(k)})$ has 2-level ω-RIP constant of order $s = (\mathcal{S}(s_1), s_2)$ satisfying $\delta_{\omega,s} \leqslant \delta$.

PROOF. As in [2], we first relate $\tilde{A}$ having the ω_1-RIP of order $\mathcal{S}(s_1)$ to B having the 2-level ω-RIP of order $\mathcal{S}(s_1)$. Suppose that $v = [x; e]$, $\|x\|_{\omega_1,0} \leqslant \mathcal{S}(s_1)$, and $\|e\|_{\omega_2,0} \leqslant s_2$. We wish to calculate an upper bound on all $\delta_{\omega,s}$ such that

$$(1-\delta_{\omega,s})\|v\|_2^2 \leqslant \|Bv\|_2^2 \leqslant (1+\delta_{\omega,s})\|v\|_2^2,$$

which by the structure of B and v is equivalent to the condition that

$$(1-\delta_{\omega,s})(\|x\|_2^2 + \|e\|_2^2) \leqslant \left\|\tilde{A}x + e\right\|_2^2 \leqslant (1+\delta_{\omega,s})(\|x\|_2^2 + \|e\|_2^2).$$

We consider

$$\left\|\tilde{A}x + e\right\|_2^2 = \left\|\tilde{A}x\right\|_2^2 + \|e\|_2^2 + 2\mathrm{Re}\langle\tilde{A}x, e\rangle. \tag{4.27}$$

The first term can be handled by the fact that $\tilde{A}$ will have the ω-RIP of order $\mathcal{S}(s_2)$, and the second term is needs no modification. Thus, it suffices to provide upper and lower bounds on the cross term. To start, we have

$$-2\left|\langle \tilde{A}x, e\rangle\right| \leqslant 2\mathrm{Re}\langle \tilde{A}x, e\rangle \leqslant 2\left|\langle \tilde{A}x, e\rangle\right|.$$

Based on our choice of intrinsic weights, we see

$$\begin{aligned} 2\left|\langle \tilde{A}x, e\rangle\right| &\leqslant 2\left\|\tilde{A}x\right\|_\infty \|e\|_1 \\ &\leqslant 2\frac{1}{\sqrt{K}}\left(\sum_{\nu\in\Lambda_0} \|\phi_\nu\|_\infty |x_\nu|\right)\|e\|_1 \\ &\leqslant 2\frac{1}{\sqrt{K}}\|x\|_{\omega_1,1}\|e\|_{\omega_2,1} \\ &\leqslant \sqrt{\frac{\mathcal{S}(s_1)s_2}{K}}2\|x\|_2\|e\|_2. \end{aligned}$$

To combine this with the terms involving sums of $\|x\|_2^2$ and $\|e\|_2^2$, we apply Cauchy's inequality (after multiplying and dividing by some $\sqrt{\varepsilon}$ to be determined), calculating

$$\begin{aligned} 2\left|\langle \tilde{A}x, e\rangle\right| &\leqslant \sqrt{\frac{\mathcal{S}(s_1)s_2}{K}}\left(\frac{\|x\|_2^2}{\varepsilon} + \varepsilon\|e\|_2^2\right) \\ &=: D\left(\frac{\|x\|_2^2}{\varepsilon} + \varepsilon\|e\|_2^2\right). \end{aligned}$$

Assuming that $\delta_{\omega_1,\mathcal{S}(s_1)} \leqslant \delta$, applying the previous bound and the ω_1-RIP to $\left\|\tilde{A}x\right\|_2^2$ in (4.27), we find

$$(1-\delta-D/\varepsilon)\|x\|_2^2 + (1-D\varepsilon)\|e\|_2^2 \leqslant \left\|\tilde{A}x+e\right\|_2^2 \leqslant (1+\delta+D/\varepsilon)\|x\|_2^2 + (1+D\varepsilon)\|e\|_2^2.$$

We now choose ε so that $\delta + D/\varepsilon = D\varepsilon =: \tilde{\delta}$, which gives

$$\varepsilon = \frac{\delta+\sqrt{\delta^2-4D^2}}{2D}, \quad \tilde{\delta} = \frac{\delta+\sqrt{\delta-4D^2}}{2} \leqslant \delta$$

so long as $\delta \geqslant 2D$. Thus, $B = [\tilde{A}, I]$ satisfies the 2-level weighted RIP of order s with $\delta_{\omega,s} \leqslant \delta$.

We now ensure that our measurements are chosen correctly to meet our assumptions made throughout the proof. In particular, if we choose $K \gtrsim \mathcal{S}(s_1)L$, with probability exceeding $1-\gamma$, we have our previous assumption that $\delta_{\omega_1,\mathcal{S}(s_1)} \leqslant \delta$ by Theorem 4.2 and the fact that the lower RIP of order s_1 is equivalent to the weighted RIP of order $\mathcal{S}(s_1)$. We just need then that $\delta \geqslant 2D$ which is satisfied when

$$\delta \geqslant 2\sqrt{\frac{\mathcal{S}(s_1)s_2}{K}} \text{ which is equivalent to } K \gtrsim \mathcal{S}(s_1)\delta^{-2}s_2.$$

Thus, the specified number of measurements in the theorem suffice to show that B has 2-level weighted RIP constant bounded by δ. □

We finally provide the recovery estimates for generalized WLAD-LASSO minimization.

THEOREM 4.9 ([1], Theorem 5.25). *Let* $0 < \gamma < 1$, $2 \leqslant s \leqslant 2^{N+1}$, *and* $\Lambda_0 = \Lambda_{HC}(s)$ *with* $\{\phi_\nu\}_{\nu\in\Lambda_0}$ *tensor Legendre or Chebyshev polynomial bases. Additionally, suppose that*

$$\delta \leqslant \frac{1}{1+4\Theta}, \quad \Theta = \frac{\sqrt{\mathcal{S}(s)+\lambda^2 H}}{\min\left\{\sqrt{\mathcal{S}(s)}, \lambda\sqrt{H}\right\}}.$$

If we draw

$$K \gtrsim s^{\kappa} \max\left\{L(s,n,\delta,\gamma), \delta^{-2}H\right\}$$

i.i.d. measurements $\{Z^{(k)}\}_{k=1}^{K}$ *from the orthogonalization measure* $\pi\,dz$ *for* κ *as in* (4.8) *and* L *as in* (4.7)*, then with probability* $1-\gamma$*, the following holds. Letting* $\hat{u}^{\sharp}$ *be the solution of the generalized WLAD-LASSO minimization problem*

$$\hat{u}^{\sharp} = \arg\min_{z\in\mathbb{C}^n} \|z\|_{\omega_1,1} + \lambda\|\tilde{A}z - \tilde{y}\|_{\omega_2,1}, \tag{4.28}$$

defining $u^{\sharp} = \sum_{\nu\in\Lambda_0}\hat{u}^{\sharp}_{\nu}\phi_{\nu}$*, if* $\|\omega_2\|_{\infty}^2 \leqslant \frac{1}{2}H$*,*

$$\|u-u^{\sharp}\|_{\infty} + \lambda\|e - (\tilde{y} - \tilde{A}\hat{u}^{\sharp})\|_{\omega_2,1} \lesssim \sigma_{L,s}(u)_{\omega_1,1} + \lambda\sigma_H(e)_{\omega_2,1},$$

$$\|u-u^{\sharp}\|_{2} + \|e - (\tilde{y} - \tilde{A}\hat{u}^{\sharp})\|_{2} \lesssim (1+\sqrt{\Theta})\left(\frac{\sigma_{L,s}(u)_{\omega_1,1}}{s^{\kappa/2}} + \frac{\sigma_H(e)_{\omega_2,1}}{\sqrt{H}}\right) + \|u_{\Lambda_R}\|_2$$

where all constants depend on δ*.*

Proof. We first rewrite (4.28) as the constrained problem

$$(\hat{u}^{\sharp}, e^{\sharp}) = \arg\min_{(z,d)\in\mathbb{C}^M\times\mathbb{C}^K} \|z\|_{\omega_1,1} + \lambda\|d\|_{\omega_2,1} = \arg\min_{x=[z;d]\in\mathbb{C}^{M+K}} \|x\|_{\omega_\lambda,1} \text{ such that } \tilde{y} - \tilde{A}z = d.$$

By the fact that $\mathcal{S}(s) \leqslant s^{\kappa}$, the given number of measurements and Theorem 4.8 implies that with probability exceeding $1-\gamma$, $B = [\tilde{A}, I]$ satisfies the 2-level weighted RIP with RIP constant bounded so that by Theorem 4.7, B has the 2-level weighted robust null space property of scale λ and order $(\mathcal{S}(s), H)$. Here we have also used our assumption that $\frac{1}{2}H \geqslant \|\omega_2\|_{\infty}^2$ and Lemma 4.1 to give that $\frac{3}{4}s_1 \geqslant \|\omega_1\|_{\infty}^2$. We finally apply the distance bounds from Lemma 4.5 with $x = [\hat{u}_{\Lambda_0}; e]$ and $z = [\hat{u}^{\sharp}; e^{\sharp}]$. Note that by our restated formulation of the WLAD-LASSO minimization program and the definition of e, since $Bx = \tilde{A}\hat{u}_{\Lambda_0} + e = \tilde{y} = \tilde{A}\hat{u}^{\sharp} + e^{\sharp} = Bz$, we must have $B(x-z) = 0$.

For the L^{∞} bound, we use the 2-level ℓ_1 distance bound (4.21) which implies

$$\begin{aligned}\|u-u^{\sharp}\|_{\infty} + \lambda\|e - (\tilde{y} - \tilde{A}\hat{u}^{\sharp})\|_{\omega_2,1} &\leqslant \|\hat{u}_{\Lambda_0} - u^{\sharp}\|_{\omega_1,1} + \lambda\|e - e^{\sharp}\|_{\omega_2,1} + \|\hat{u}_{\Lambda_R}\|_{\omega_1,1} \\ &= \|x-z\|_{\omega_\lambda,1} + \|\hat{u}_{\Lambda_R}\|_{\omega_1,1} \\ &\lesssim \sigma_{(\mathcal{S}(s),H)}(x)_{\omega_\lambda,1} + \|\hat{u}_{\Lambda_R}\|_{\omega_1,1},\end{aligned}$$

since $[\hat{u}^{\sharp}; e^{\sharp}]$ minimizes $\|z\|_{\omega_\lambda,1}$. By definition,

$$\sigma_{(\mathcal{S}(s),H)}(x)_{\omega_\lambda,1} = \sigma_{\mathcal{S}(s)}(\hat{u}_{\Lambda_0})_{\omega_1,1} + \lambda\sigma_H(e)_{\omega_2,1}.$$

As before, since the first term is defined taking the infimum over approximations of $\hat{u}_{\Lambda_0}$ supported on index sets S with $\omega_1(S) \leqslant \mathcal{S}(s)$, this set includes all cardinality s lower sets by the definition of intrinsic sparsity. Thus, the error in the best s-term approximation to $\hat{u}_{\Lambda_0}$ in lower sets can only be larger as it the infimum over a smaller feasible set. Additionally, as before, since Λ_0 contains all lower sets, the error in the best s-term approximation to $\hat{u}_{\Lambda_0}$ in lower sets must contain $\|\hat{u}_{\Lambda_R}\|_{\omega_1,1}$. And finally, $\sigma_{L,s}(\hat{u}_{\Lambda_0})_{\omega_1,1} = \sigma_{L,s}(\hat{u})_{\omega_1,1}$, since the best approximation to $\hat{u}$ on a lower set must be supported in Λ_0. Combining these facts gives

$$\|u-u^{\sharp}\|_{\infty} + \lambda\|e - (\tilde{y} - \tilde{A}\hat{u}^{\sharp})\|_{\omega_2,1} \lesssim \sigma_{L,s}(u)_{\omega_1,1} + \lambda\sigma_H(e)_{\omega_2,1}$$

as desired.

For the L^2 error, we use Parseval's identity, the ℓ_2 bound (4.22) and the lower bound on the intrinsic sparsity in Lemma 4.3 giving

$$\begin{aligned}
\left\|u-u^\sharp\right\|_2+\left\|e-(\tilde{y}-\tilde{A}\hat{u}^\sharp)\right\|_2 &\leqslant \left\|\hat{u}_{\Lambda_0}-u^\sharp\right\|_2+\left\|e-e^\sharp\right\|_2+\|u_{\Lambda_R}\|_2\\
&\leqslant \sqrt{2}\|x-z\|_2+\|u_{\Lambda_R}\|_2\\
&\lesssim (1+\sqrt{\Theta})\frac{\sigma_{(\mathcal{S}(s),H)}(x)_{\omega_\lambda,1}}{\sqrt{\mathcal{S}(s)+\lambda^2 H}}+\|u_{\Lambda_R}\|_2\\
&\lesssim (1+\sqrt{\Theta})\left(\frac{\sigma_{L,s}(u)_{\omega_1,1}}{\sqrt{\mathcal{S}(s)+\lambda^2H}}+\lambda\frac{\sigma_H(e)_{\omega_2,1}}{\sqrt{\mathcal{S}(s)+\lambda^2H}}\right)+\|u_{\Lambda_R}\|_2\\
&\lesssim (1+\sqrt{\Theta})\left(\frac{\sigma_{L,s}(u)_{\omega_1,1}}{\sqrt{\mathcal{S}(s)}}+\frac{\sigma_H(e)_{\omega_2,1}}{\sqrt{H}}\right)+\|u_{\Lambda_R}\|_2\\
&\lesssim (1+\sqrt{\Theta})\left(\frac{\sigma_{L,s}(u)_{\omega_1,1}}{s^{\kappa/2}}+\frac{\sigma_H(e)_{\omega_2,1}}{\sqrt{H}}\right)+\|u_{\Lambda_R}\|_2
\end{aligned}$$

as desired. □

Remark 4.2. *Let us consider the conclusions of Theorem 4.9 for the case of our original WLAD-LASSO program* (4.14). *In this case, we must take our weights on the error terms to be one. Thus, for any* $H \geqslant 2$, $\|\omega_2\|_\infty^2 \leqslant \frac{1}{2}H$ *and Theorem 4.9 applies to give the error estimates*

$$\begin{aligned}
\left\|u-u^\sharp\right\|_\infty &\lesssim \sigma_{L,s}(u)_{\omega_1,1}+\lambda\sigma_H(e)_1,\\
\left\|u-u^\sharp\right\|_2 &\lesssim (1+\sqrt{\Theta})\left(\frac{\sigma_{L,s}(u)_{\omega_1,1}}{s^{\kappa/2}}+\frac{\sigma_H(e)_1}{\sqrt{H}}\right)+\|u_{\Lambda_R}\|_2
\end{aligned}$$

for $\Theta=\sqrt{\mathcal{S}(s)+\lambda^2H}/\min\left\{\sqrt{\mathcal{S}(s)},\lambda\sqrt{H}\right\}$. *Choosing* $\lambda=\sqrt{\mathcal{S}(s)}/\sqrt{H}$ *gives* $\Theta=\sqrt{2}$, *removing this factor from the error bound.*

In contrast to the other three minimization methods, our dependence on $\|e\|_2$ *in the error bounds has been replaced with* $\sigma_H(e)_1$ *in the WLAD-LASSO case. When the total measurement error fits our assumed decomposition of* $e=e^{\text{bounded}}+e^{\text{sparse}}$ *with* H *chosen to match the sparsity of* e^{sparse}, *we may rewrite*

$$\sigma_H(e)_1 \leqslant \|e-e^{\text{sparse}}\|_1=\left\|e^{\text{bounded}}\right\|_1 \leqslant \sqrt{K}\left\|e^{\text{bounded}}\right\|_2.$$

In the case that we choose $K \asymp s^\kappa \max\{L,H\}$ *and* $\lambda \asymp s^{\kappa/2}/\sqrt{H}$, *we have* $\sqrt{K/H} \lesssim s^{\kappa/2}$ *and therefore*

$$\begin{aligned}
\left\|u-u^\sharp\right\|_\infty &\lesssim \sigma_{L,s}(u)_{\omega_1,1}+s^\kappa\left\|e^{\text{bounded}}\right\|_2\\
\left\|u-u^\sharp\right\|_2 &\lesssim \frac{\sigma_{L,s}(u)_{\omega_1,1}}{s^{\kappa/2}}+s^{\kappa/2}\left\|e^{\text{bounded}}\right\|_2+\|u_{\Lambda_R}\|_2.
\end{aligned}$$

This matches the error bounds of the other three methods up to a factor of $s^{\kappa/2}$ *on the* $\left\|e^{\text{bounded}}\right\|_2$ *term, however, the WLAD-LASSO minimization removes any dependence on arbitrarily large, sparse error as well. We note though that for these bounds to hold, the tuning parameter* λ *still requires some information of the error unlike WSR-LASSO minimization. Rather than needing information on the norm of the error as in the WQCBP and WLASSO however, only the sparsity of* e^{sparse} *is required.*

Bibliography

[1] B. Adcock, A. Bao, and S. Brugiapaglia, *Correcting for unknown errors in sparse high-dimensional function approximation*, Numer. Math., 142 (2019), pp. 667–711.

[2] B. Adcock, A. Bao, J. D. Jakeman, and A. Narayan, *Compressed sensing with sparse corruptions: fault-tolerant sparse collocation approximations*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 1424–1453.

[3] B. Adcock, S. Brugiapaglia, and C. G. Webster, *Compressed sensing approaches for polynomial approximation of high-dimensional functions*, in Compressed sensing and its applications, Appl. Numer. Harmon. Anal., Birkhäuser/Springer, Cham, 2017, pp. 93–124.

[4] I. Babuška, F. Nobile, and R. Tempone, *A stochastic collocation method for elliptic partial differential equations with random input data*, SIAM J. Numer. Anal., 45 (2007), pp. 1005–1034.

[5] J. Bäck, F. Nobile, L. Tamellini, and R. Tempone, *Stochastic spectral Galerkin and collocation methods for PDEs with random coefficients: a numerical comparison*, in Spectral and high order methods for partial differential equations, vol. 76 of Lect. Notes Comput. Sci. Eng., Springer, Heidelberg, 2011, pp. 43–62.

[6] V. Barthelmann, E. Novak, and K. Ritter, *High dimensional polynomial interpolation on sparse grids*, in Multivariate polynomial interpolation, vol. 12, Baltzer Science Publishers BV, Bussum, 2000, pp. 273–288.

[7] J. P. Boyd, *Chebyshev and Fourier spectral methods*, Dover Publications, Inc., Mineola, NY, second ed., 2001.

[8] S. Brugiapaglia and B. Adcock, *Robustness to unknown error in sparse regularization*, IEEE Trans. Inform. Theory, 64 (2018), pp. 6638–6661.

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

[10] A. Chkifa, A. Cohen, and C. Schwab, *Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs*, J. Math. Pures Appl. (9), 103 (2015), pp. 400–428.

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

[12] A. Cohen, R. DeVore, and C. Schwab, *Convergence rates of best* N*-term Galerkin approximations for a class of elliptic sPDEs*, Found. Comput. Math., 10 (2010), pp. 615–646.

[13] A. Cohen, R. Devore, and C. Schwab, *Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's*, Anal. Appl. (Singap.), 9 (2011), pp. 11–47.

[14] L. C. Evans, *Partial differential equations*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, second ed., 2010.

[15] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, Applied and Numerical Harmonic Analysis, Birkhäuser/Springer, New York, 2013.

[16] M. Gunzburger, C. G. Webster, and G. Zhang, *Sparse collocation methods for stochastic interpolation and quadrature*, in Handbook of uncertainty quantification. Vol. 1, 2, 3, Springer, Cham, 2017, pp. 717–762.

[17] F. Krahmer, S. Mendelson, and H. Rauhut, *Suprema of chaos processes and the restricted isometry property*, Comm. Pure Appl. Math., 67 (2014), pp. 1877–1904.

[18] M. Motamed, *Pde-based uncertainty quantification.* Lecture notes, Argonne National Laboratory, May 2019.

[19] F. Nobile, R. Tempone, and C. G. Webster, *An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data*, SIAM J. Numer. Anal., 46 (2008), pp. 2411–2442.

[20] ———, *A sparse grid stochastic collocation method for partial differential equations with random input data*, SIAM J. Numer. Anal., 46 (2008), pp. 2309–2345.

[21] A. Quarteroni, A. Manzoni, and F. Negri, *Reduced basis methods for partial differential equations*, vol. 92 of Unitext, Springer, Cham, 2016. An introduction, La Matematica per il 3+2.

[22] H. Rauhut, *Compressive sensing and structured random matrices*, in Theoretical foundations and numerical methods for sparse recovery, vol. 9 of Radon Ser. Comput. Appl. Math., Walter de Gruyter, Berlin, 2010, pp. 1–92.

[23] H. Rauhut and C. Schwab, *Compressive sensing Petrov-Galerkin approximation of high-dimensional parametric operator equations*, Math. Comp., 86 (2017), pp. 661–700.

[24] H. Rauhut and R. Ward, *Interpolation via weighted ℓ_1 minimization*, Appl. Comput. Harmon. Anal., 40 (2016), pp. 321–351.

[25] C. Schwab, *QMC Galerkin discretization of parametric operator equations*, in Monte Carlo and quasi-Monte Carlo methods 2012, vol. 65 of Springer Proc. Math. Stat., Springer, Heidelberg, 2013, pp. 613–629.

[26] S. A. Smolyak, *Quadrature and interpolation formulas for tensor products of certain classes of functions*, in Doklady Akademii Nauk, vol. 148, Russian Academy of Sciences, 1963, pp. 1042–1045.

[27] R. Tempone and S. Wolfers, *Smolyak's algorithm: A powerful black box for the acceleration of scientific computations*, in Sparse Grids and Applications - Miami 2016, J. Garcke, D. Pflüger, C. G. Webster, and G. Zhang, eds., Cham, 2018, Springer International Publishing, pp. 201–228.

[28] L. N. Trefethen, *Is Gauss quadrature better than Clenshaw-Curtis?*, SIAM Rev., 50 (2008), pp. 67–87.

[29] R. Vershynin, *High-dimensional probability*, vol. 47 of Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, With a foreword by Sara van de Geer.

[30] D. Xiu, *Numerical methods for stochastic computations: A spectral methods approach*, Princeton University Press, Princeton, NJ, 2010.