

Sparse Reconstruction Techniques for Solutions of High-Dimensional Parametric PDEs

A Dissertation Presented for the
Doctor of Philosophy
Degree

The University of Tennessee, Knoxville

Nicholas Calvin Dexter

August 2018

© by Nicholas Calvin Dexter, 2018
All Rights Reserved.

To my wife, Gara

Acknowledgments

First, I would like to thank my doctoral committee, Profs. Dongarra, Karakashian, Salgado, and Wise. Each of you played a key role in directing my graduate education and inspiring in me a passion for mathematical research.

I would also especially like to express my gratitude for my academic advisor, Professor Clayton Webster, for his continuous support in my Ph.D. studies at the University of Tennessee and research at Oak Ridge National Laboratory. Since we first met six years ago, he has provided a constant source of wisdom and guidance, exposing me to many new areas of mathematics.

I'd also like to acknowledge and thank my collaborators, Guannan Zhang for our work on stochastic Galerkin approximations, Abdellah Chkifa for collaboration on the compressed sensing-based polynomial approximation methods, and Hoang Tran for multiple collaborations on structured sparse recovery, convergence of algorithms, and the sparse reconstruction techniques which make up most of the body of this work. Thanks also to Miroslav Stoyanov for our many useful conversations, and Peter Jantsch and Joe Daws for the many helpful conversations on a wide array of subjects from approximation theory, manifold learning, and sparse image reconstruction techniques.

I would also like to thank Dr. Lee Riedinger and the Bredesen Center for this amazing opportunity to study mathematics while conducting research at a national laboratory, and all of the resources they have provided through the years. Thanks also to the wonderful staff in the Department of Mathematics at UT, most especially to Pam Armentrout, without whose guidance, advice, and support, none of us would graduate.

To all of the friends who have made my studies at UT memorable: Eddie Tu, Brian Allen, Nate Pollesch, John Cummings, Tyler Massaro, Kelly Rooker, Kylie Berry, Khoa Dinh, Gabriel Dusing, Ibrahim Aslan, Montgomery Taylor, George Feng, Mustafa Elmas, Vincent Heningburg, Kolin Konjura, Cassie Micucci, Vy Nguyen, Kacy & Stefan Schnake, Kevin Sonnanburg, Joshua Mike and Grace McClurkin. I've enjoyed our many discussions on mathematics, books, movies, and ramen. To all of the wonderful people I've met over my last seven years in Knoxville: Annette DeCapite, Kaela O'Dell, and Hattie & Aaron Owsley. I've enjoyed all of our hikes, El Cantarito nights, and parties.

Finally, I would like to thank my family for their love and support during this extended period of graduate study. Most especially, I thank my wonderful wife Gara Dexter for her unwavering love and encouragement through this challenging time. To my mom, who taught me how to be independent and never give up on my dreams. To my loving grandparents, who were a constant source of inspiration and strength. I would not be who I am today if it weren't for all of the love and support you gave.

Abstract

This work studies sparse reconstruction techniques for approximating solutions of high-dimensional parametric PDEs. Such problems are relevant to mathematical modeling in engineering and the sciences, and are inherently difficult to solve due to the exponential relationship between dimension and computational cost of approximation, a phenomenon referred to as the *curse of dimensionality*. Nonetheless, for many practical applications, their solutions possess favorable properties, e.g., regularity and compressibility, enabling efficient numerical methods of solution.

For the past few decades, polynomial approximations, e.g., *stochastic Galerkin* (SG) methods [124], have remained popular for solving parameterized PDEs due to their ability to exploit these properties, achieving fast (sub-exponential) convergence rates. However, in practice the complexity of SG methods depends on the type of parametric dependence in the underlying PDE. Previous analysis on SG complexity neglected this fact, focusing only on the affine case. To resolve this, we prove sharp bounds on the cost of solving SG systems, resulting from discretizing elliptic PDEs with arbitrary parametric dependence, to a prescribed tolerance. We also verify our theoretical results with detailed numerical experiments over a variety of parameterizations.

Since the mid-2000s, increasing interest in sparse signal reconstruction through *compressed sensing* (CS) [20, 21, 47] has led to a wide array of applications, including sparse polynomial approximation. To improve convergence of CS polynomial approximations, we introduced a novel weighting scheme based on lower sets for smooth function approximation, also relevant to parameterized PDEs. However, standard CS is only capable of approximating functionals of solutions to parameterized PDEs. Therefore, we introduced a novel energy norm-based reformulation of basis pursuit denoising, which enables sparse recovery over tensor products of Hilbert spaces. We also derive the forward-backward iterations for its solution, proving strong convergence in the infinite dimensional setting. To improve efficiency of our solver, we combine Bregman iterations and fixed-point continuation, derived for this setting, with forward-backward iterations. We denote our approach sparse *Hilbert-valued vector* (HV) recovery, which we prove achieves sub-exponential rates of convergence. We also conduct detailed numerical experiments, demonstrating the minimal sample complexity requirements of our approach compared to several alternative methods of solution.

Table of Contents

1	Introduction	1
1.1	Explicit cost bounds of stochastic Galerkin approximations for parameterized PDEs	4
1.2	Polynomial approximation via compressed sensing of high-dimensional functions on lower sets	4
1.3	Strong convergence of forward-backward splitting in reconstructing jointly-sparse signals	5
1.4	Sparse reconstruction techniques for solutions of high-dimensional parametric PDEs	6
2	Parameterized partial differential equations	8
2.1	Parameterized elliptic PDE	8
3	Overview of methods	13
3.1	Finite element method for physical discretization	13
3.2	Stochastic Galerkin method	14
3.2.1	Stochastic Galerkin approximation with an orthogonal basis	15
3.2.2	Representations of the data and resulting linear systems	18
3.3	Monte Carlo method	20
3.4	Stochastic collocation method	22
3.5	Compressed sensing method	23
3.5.1	Improved RIP estimate for bounded orthonormal systems	28
3.5.2	Sparse recovery on lower sets	30
3.5.3	Application of standard compressed sensing to approximate functionals of solutions to parameterized PDEs	32
3.6	A novel method for sparse Hilbert-valued vector recovery	34
3.6.1	Sparse recovery of Hilbert-valued vectors	34
3.6.2	Setting of tensor products of Hilbert spaces	36
3.6.3	Relation between sparse Hilbert-valued and joint-sparse recovery problems	37
3.6.4	Uniform recovery of Hilbert-valued signals	40
4	Algorithms for sparse recovery	44
4.1	Forward-backward splitting algorithm	45
4.1.1	Gradient descent step	47

4.1.2	Proximal point step	48
4.1.3	Convergence results	49
4.2	Bregman iterations	56
4.2.1	Bregman iterations for Hilbert-valued recovery	58
4.3	Fixed-point continuation for improving forward-backward splitting algorithm performance	60
4.3.1	Fixed point continuation	61
4.3.2	Combined Bregman-FPC algorithm	64
5	Computational complexity	70
5.1	Complexity of stochastic Galerkin method	70
5.1.1	Cost of solving the generalized SG system	71
5.1.2	Complexity of SG matrix-vector products	73
5.1.3	ε -complexity analysis of the SG method	81
5.1.4	Conditioning aspects of the generalized SG system	87
5.2	Complexity of stochastic collocation method	87
5.2.1	Cost of solving the SC systems	87
5.2.2	Comparing the explicit cost bounds of the SG and SC methods	88
5.3	Complexity of the sparse Hilbert-valued recovery method	92
5.3.1	Compressibility results for the parameterized PDE setting	93
5.3.2	Error estimates for Hilbert-valued recovery	94
6	Numerical experiments	98
6.1	Computational complexity of SG approximations	99
6.1.1	Piecewise affine coefficients	100
6.1.2	Polynomial coefficients	103
6.1.3	Transcendental coefficients	105
6.2	Approximating high-dimensional smooth functions with lower sparse recovery	107
6.3	Numerical results for the approximation of solutions to the parameterized elliptic PDE with sparse Hilbert-valued recovery	113
6.3.1	Comparison of sparse HV recovery and CS point-wise functional recovery for global approximation of solutions to parameterized PDEs	114
6.3.2	Comparison of sparse HV recovery with SC, SG, and MC methods for parameterized PDEs	118
7	Conclusions	122
	Bibliography	124
	Appendices	135
A	Proofs from the text	136
A	Proofs of results from Chapter 4	136
B	Proofs of results from Chapter 5	140
Vita		142

List of Tables

3.1	Comparison of $N_p = \#(\mathcal{J}_p)$ SDOF for the multi-index sets $\mathcal{J}_p = \mathcal{J}_p^{\text{TD}}$ (total degree) and $\mathcal{J}_p = \mathcal{J}_p^{\text{TP}}$ (tensor product), where $d = \dim(\mathcal{U})$ and p is the maximal degree.	17
4.1	Parameters used in the Bregman-FPC Algorithm 2 for the numerical experiments on sparse HV recovery in Section 6.3. The values of $I_{\text{Max-Bregman}}$ & $I_{\text{Max-FPC}}$ are selected based on run-time considerations.	69
6.1	Growth of the bound (6.6) on the constant $\mathcal{M}(p, \bar{r})$, with respect to dimension d and $\bar{r}$ when $p = 5$ is fixed.	105
6.2	Comparison of cost in matrix-vector products for solving problem (2.2) with coefficient (6.7) and forcing $f(x_1, x_2) = 2\cos(x_1)\sin(x_2)$ using the SC-CC and SG-TD approximations, with the strategy of picking the CG tolerance to be $\ \mathbb{E}[\varepsilon_{\text{SG}}]\ _{L^\infty(D)}/10$ for the SG method and $\ \mathbb{E}[\varepsilon_{\text{SC}}]\ _{L^\infty(D)}/10$ for the SC method using the approximations from (6.1) and the cost metrics (5.5) and (5.30) (normalized to an $\mathcal{O}(K_h)$ FE stiffness matrix-vector product).	108

List of Figures

2.1	Domains of uniform ellipticity for some one-dimensional coefficients $a(x, y)$ are indicated by the gray regions in each plot. The blue and red curves represent the maximal discs $\mathcal{D}_\gamma$ and ellipses $\mathcal{E}_\gamma$, respectively, that can be contained in those domains, and the green lines represent the interval $\mathcal{U} = [-1, 1]$	12
3.1	Domains of uniform ellipticity for orthonormal Legendre expansions of order r of the one-dimensional coefficient $a(x, y) = 0.1 + \exp(2.5y)$ are indicated by the gray regions in each plot. The last plot shows the domain of uniform ellipticity of the original function $a(x, y)$. The blue and red curves represent the maximal discs $\mathcal{D}_\gamma$ and ellipses $\mathcal{E}_\gamma$, respectively, that can be contained in those domains, and the green lines represent the interval $\mathcal{U} = [-1, 1]$	21
5.1	Visualization of the number of nonzeros of $\sum_{\alpha \in \mathcal{J}_r^{\text{TD}}} \mathbf{G}_\alpha \in \mathbb{R}^{N_p \times N_p}$ with $\mathbf{G}_\alpha$ from (3.20), when p and $\mathcal{J}_p^{\text{TD}}$ are fixed and r and $\mathcal{J}_r^{\text{TD}}$ are allowed to grow. Each pixel implies a nonzero block $\langle \Psi_\nu \Psi_\mu \Psi_\alpha \rangle * \mathbf{B}_\alpha \in \mathbb{R}^{K_h \times K_h}$ in the matrix $\mathbf{K}_r$. Here $d = 8$ and $p = 3$ so that $N_p = \#(\mathcal{J}_p^{\text{TD}}) = 165$. Moving left to right r is increasing by one unit from 0 to 5. At $r = 6$, the matrix is entirely block-dense.	73
5.2	For a range of p values, we plot for (top) $r = 1$, (middle) $r = 2$, and (bottom) $r = 3$, the actual sparsity $\mathcal{M}(p, r)$ given by (5.3) of the Galerkin system $\mathbf{K}_r$ from (3.21) (blue), the bound on the sparsity from (5.9) (green), and the bound on the sparsity from (5.11) (red) in the cases $d = 4, 8, 16$ moving left to right.	79
5.3	For $r = p$ with a range of p values, we plot for (top) $d = 4$ (left) and $d = 8$ (right) and (bottom) $d = 20$ (left) and $d = 50$ (right), the actual sparsity $\mathcal{M}(p, r)$ given by (5.3) of the Galerkin system $\mathbf{K}_r$ from (3.21) (blue), the bound on the sparsity from (5.9) (green), and the bound on the sparsity from (5.11) (red).	80
5.4	Condition numbers of both the unpreconditioned matrix $\mathbf{K}_r$ and the preconditioned matrix $\mathbf{P}^{-1}\mathbf{K}_r$ with decreasing finite element mesh parameter h and $r = p$ for the coefficient (6.7) from Section 6.1.3 with $d = 4$ and $L_c = 1/2$. . .	92

6.1	We plot (left) a triangulation of the domain D generated with distmesh , with circular and square inclusions. Red nodes highlight the boundary of an inclusion or the domain D , blue nodes highlight nodes on the interior of an inclusion. We also plot (right) the expected value of the solution of (2.2) with stochastic conductivity coefficient (6.2).	101
6.2	We plot (left) error versus SDOF and (right) error versus cost for solving problem (2.2) with coefficient (6.2) and forcing function (6.3), using the approximations to the error metrics $\ \mathbb{E}[\varepsilon_{\text{SG}}]\ _{L^\infty(D)}$ and $\ \mathbb{E}[\varepsilon_{\text{SC}}]\ _{L^\infty(D)}$ from (6.1) for error and the cost metrics (5.5) and (5.30) (normalized to an $\mathcal{O}(K_h)$ FE stiffness matrix-vector product).	102
6.3	We plot (top) error vs SDOF and (bottom) error versus cost for solving problem (2.2) with coefficient (6.5) with $\bar{r} = 1$ (left) , $\bar{r} = 3$ (middle) , and $\bar{r} = 7$ (right) , with forcing $f(x) = 1 \forall x \in \bar{D}$, using the approximations to the error metrics $\ \mathbb{E}[\varepsilon_{\text{SG}}]\ _{L^\infty(D)}$ and $\ \mathbb{E}[\varepsilon_{\text{SC}}]\ _{L^\infty(D)}$ from (6.1) for error and the cost metrics (5.5) and (5.30) (normalized to an $\mathcal{O}(K_h)$ FE stiffness matrix-vector product).	104
6.4	We plot (left) error versus SDOF and (right) error versus cost in solving problem (2.2) with coefficient (6.7) and forcing deterministic $f(x_1, x_2) = 2 \cos(x_1) \sin(x_2)$, using the approximations to the error metrics $\ \mathbb{E}[\varepsilon_{\text{SG}}]\ _{L^\infty(D)}$ and $\ \mathbb{E}[\varepsilon_{\text{SC}}]\ _{L^\infty(D)}$ from (6.1) for error and the cost metrics (5.5) and (5.30) (normalized to an $\mathcal{O}(K_h)$ FE stiffness matrix-vector product).	107
6.5	Comparison of the averaged $L^2_\rho(\mathcal{U})$ error in approximating g from (6.11) using weighted ℓ^1 -minimization with various choices of weights in (top) $d = 8$, $N = 1843$, and $\ g_{\mathcal{H}^c_s}\ _2 = 1.5922e - 05$, and (bottom) $d = 16$, $N = 4129$, and $\ g_{\mathcal{H}^c_s}\ _2 = 1.3388e - 06$	108
6.6	Comparison of the averaged $L^2_\rho(\mathcal{U})$ error in approximating g from (6.12) using weighted ℓ^1 minimization with various choices of weights in (top) $d = 8$, $N = 1843$, and $\ g_{\mathcal{H}^c_s}\ _2 = 4.0232e - 07$, and (bottom) $d = 16$, $N = 4129$, and $\ g_{\mathcal{H}^c_s}\ _2 = 2.0155e - 06$	109
6.7	Comparison of the averaged $L^2_\rho(\mathcal{U})$ error in approximating g from (6.13) using weighted ℓ^1 minimization with various choices of weights with $d = 8$, $N = 1843$ and $\ g_{\mathcal{H}^c_s}\ _2 = 6.1018e - 3$	110
6.8	Comparison of the averaged $L^2_\rho(\mathcal{U})$ error in approximating g from (6.14) using weighted ℓ^1 minimization with various choices of weights in (top) $d = 8$, $N = 1843$, and $\ g_{\mathcal{H}^c_s}\ _2 = 7.2714e - 07$, and (bottom) $d = 16$, $N = 4129$, and $\ g_{\mathcal{H}^c_s}\ _2 = 3.7412e - 07$	112
6.9	Comparison of the averaged $L^\infty(\mathcal{U})$ error of approximations studied in Figure 6.6 in (left) $d = 8$ and (right) $d = 16$	112

6.10	(left) Comparison of relative error statistics $\varepsilon_{h,\text{approx}}^{\text{rel-}\mathbb{E}}$ from (6.15) for the sparse HV recovery method (HV-TD) and compressed sensing point-wise functional recovery (CS-TD) methods, both computed with a total degree basis of order $p = 2$ in $d = 100$ dimensions having cardinality $N = 5151$. (center) Magnitudes of the coefficients in (top) energy norm and (bottom) pointwise at three physical locations x_i sorted lexicographically, (right) same as center after sorting by largest in magnitude. Here $\mathbf{c}^{h,*}$ is obtained with the stochastic Galerkin method in the same total degree polynomial subspace.	117
6.11	Comparison of relative approximation errors $\varepsilon_{h,\text{approx}}^{\text{rel-}\mathbb{E}}$ (left) and $\varepsilon_{h,\text{approx}}^{\text{rel-}\sigma}$ (right) from (6.15) for the $L = 2, 3$ stochastic collocation (SC-CC), $p = 1, 2$ stochastic Galerkin (SG-TD), Monte Carlo (MC), and total degree order $p = 2$ sparse HV recovery (HV-TD) methods for solving (2.2) with coefficient (6.18) and correlation length $L_c = 1/4$. (top) $d = 20$, $N = 231$, (middle) $d = 60$, $N = 1891$, (bottom) $d = 100$, $N = 5151$	120
6.12	Comparison of relative approximation errors $\varepsilon_{h,\text{approx}}^{\text{rel-}\mathbb{E}}$ (left) and $\varepsilon_{h,\text{approx}}^{\text{rel-}\sigma}$ (right) from (6.15) for the $L = 2, 3$ stochastic collocation (SC-CC), $p = 1, 2$ stochastic Galerkin (SG-TD), Monte Carlo (MC), and total degree order $p = 2$ with $N = 5151$ sparse HV recovery (HV-TD) methods for solving (2.2) with coefficient (6.18) and correlation length $L_c = 1/2$ in $d = 100$ dimensions.	121
6.13	Comparison of relative approximation errors $\varepsilon_{h,\text{approx}}^{\text{rel-}\mathbb{E}}$ (left) and $\varepsilon_{h,\text{approx}}^{\text{rel-}\sigma}$ (right) from (6.15) for the $L = 2, 3$ stochastic collocation (SC-CC), Monte Carlo (MC), and total degree order $p = 4$ with $N = 5985$ sparse HV recovery (HV-TD) methods for solving (2.2) with coefficient (6.7) and correlation length $L_c = 1/8$ in $d = 17$ dimensions.	121

Chapter 1

Introduction

The primary focus of this work is on techniques for sparse approximation of solutions to high-dimensional parameterized partial differential equations (PDEs). Such equations arise in many science and engineering problems when the underlying physical model can be described by a PDE having dependence on a large number of parameters [61, 65]. In applications, each parameter may be used to describe a particular physical property of the system, e.g., material diffusivity, boundary conditions, source or sink terms, geometry of the domain or subdomains, and other details relevant to the model at hand [4, 6, 42]. In contrast to traditional approaches to mathematical modeling, which often treat these parameter values as immutable quantities, here we seek to characterize solution manifolds corresponding to ranges of parameter values. Depending on the particular application, it may be necessary to search the manifold as a part of an optimization procedure or compute a desired *quantity of interest* (QOI) as part of a simulation-informed decision making process.

Parameterized PDEs have been applied to mathematical modeling problems in both deterministic and stochastic contexts [56, 57, 78]. For example, when the system is entirely deterministic, i.e., depending on parameters which are subject to exact controls, parameterized PDEs can be used to study the response of the system with respect to small changes in the input parameter values. Common scenarios include optimizing the output of the PDE with respect to the parameter vector to maximize performance or efficiency, or minimize risk of failure [13]. To properly model complex physical systems comprised of multiple integrated parts, the underlying PDE may depend on hundreds or thousands of parameters, each describing a particular component. In such high-dimensional settings, it is paramount to identify the most important components, and their necessary structural or thermal properties, to guarantee favorable outcomes of the design process and ensure long-term reliability.

On the other hand, the underlying system may be assumed to possess some uncertainty, either of an *aleatoric* or *epistemic* nature. In either case it is desirable to quantify any uncertainties concerning the response of the system with respect to these random or unknown influences. Stochastic modeling through parameterized PDEs may be used to assess the volatility of a complex system in engineering and science applications, e.g., a nuclear reactor or combustion engine, when the operating conditions inside that system are partially unknown. Obtaining accurate measurements in such scenarios may be impossible for any number of reasons, including: great cost of gathering samples, physical limitations on

measuring, or otherwise lack of reliability or accuracy of the measurement process. The most straightforward way to incorporate such uncertainties can be found in probability theory, treating the unknown system properties as a vector of random variables distributed according to a joint *probability density function* (PDF) [50, 56, 78]. Using the probabilistic framework, applications may require computation of statistical QOIs, e.g., expectation or variance of the solution over a particular physical region, to aid in assessing risk or probabilities of rare events.

The underlying mathematical modeling problem relevant to both deterministic and stochastic modeling contexts can be described with PDEs in the following way. Consider the problem of simultaneously resolving a family of solutions to a PDE, defined on a domain $D \subset \mathbb{R}^n$, e.g., $n \in \{1, 2, 3\}$, referred to herein as the physical domain. Let $\mathcal{D}$ be a differential operator with respect to x on D , and, for $d \in \mathbb{N}$, define the parameter (or stochastic) domain $\mathcal{U} = \prod_{j=1}^d \mathcal{U}_j \subset \mathbb{R}^d$. The underlying model can then be described by a parameterized boundary value problem: for every $\mathbf{y} \in \mathcal{U}$, find $u(\cdot, \mathbf{y}) : \bar{D} \rightarrow \mathbb{R}$, such that

$$\mathcal{D}(u, \mathbf{y}) = 0, \quad \text{in } D, \quad (1.1)$$

subject to suitable boundary conditions. In the stochastic modeling setting, the vector $\mathbf{y} = \mathbf{y}(\omega) : \Gamma \rightarrow \mathcal{U}$ represents a vector of random variables, assumed here to be i.i.d., defined on the event space Γ , with $\mathcal{U} \equiv \mathbf{y}(\Gamma)$ the image of $\mathbf{y}$, and $\mathbf{y}$ distributed according to the joint PDF $\varrho = \prod_{j=1}^d \varrho_j$.

In the considered setting, a sample $u(\mathbf{y})$ at a particular $\mathbf{y} \in \mathcal{U}$ provides little information about the response of the system in the aforementioned stochastic and deterministic modeling contexts. Therefore, we seek to directly approximate the *solution map*

$$\mathbf{y} \in \mathcal{U} \mapsto u(\cdot, \mathbf{y}) \in \mathcal{V}. \quad (1.2)$$

Here the solution space $\mathcal{V}$ is typically a Sobolev space (e.g., $H_0^1(D)$ in the case of the elliptic PDE), or, more generally, a Banach space. Using this approximation, or *surrogate*, of the solution map, one can then compute any desired QOI of the solution or optimize the output of the system (1.1) with respect to the parameter vector $\mathbf{y}$ in the contexts discussed above. Of course, even obtaining a single evaluation $u(\mathbf{y})$ at a particular $\mathbf{y} \in \mathcal{U}$ often requires solving a large discretized system of equations, depending nonlinearly on $\mathbf{y}$. Therefore, in designing approximation schemes to (1.2), one hopes to obtain a surrogate that can be quickly and efficiently evaluated at arbitrary $\mathbf{y} \in \mathcal{U}$.

Modern mathematical modeling and simulation techniques include a wide variety of methods to approximate (1.2) in the high-dimensional setting, from *polynomial approximation*, to purely *sampling-based approximation*, to combinations of both approaches through *linear* and *nonlinear interpolation*. Of these methods, stochastic polynomial methods are the most widely used, due to their ability to exploit the regularity of the solution, with respect to the parameterization, in obtaining fast convergence rates. Such methods include the stochastic Galerkin (SG) method [124], a method of polynomial approximation which employs a Galerkin projection in the parametric domain, yielding fully discrete polynomial approximations to the solution with $L^2_q(\mathcal{U})$ -optimal error. The stochastic collocation (SC) method [91], on the other hand, constructs an interpolant of the solution from a carefully

selected set of points, also yielding a fully discrete polynomial approximation. Both methods enable fast evaluation of the surrogate map to u , and can be shown to satisfy approximation error bounds which decay sub-exponentially in the asymptotic regime as the overall order of the polynomial approximation increases to infinity [91, 112].

Nonetheless, all of the methods considered in this work suffer from a well-known phenomenon in high-dimensional approximation, the so-called *curse of dimensionality* [12, 11]. While the curse is universal, its effects can be highly problem and method-specific, manifesting in three key ways:

1. approximation error bounds for these methods always carry an explicit, unfavorable dependence on the dimension d , with rates of convergence deteriorating as $d \rightarrow \infty$,
2. the number of degrees of freedom (DOF) required by constructive approximations often grow exponentially with the dimension,
3. in all but the simplest cases of problem (1.1), computational complexity of constructing the approximation depends nonlinearly on the number of DOFs.

All of these effects contribute to the curse, leading to an exponential relationship between the computational effort required to find a numerical approximation to the solution and the dimension d of the parameter vector. This problem is only compounded by the fact that many physical problems require a large number of DOF for the spatial and temporal discretizations to achieve the required fidelity for making accurate predictions. Furthermore, as noted in [28], the theory of nonlinear Kolmogorov widths can be used to show that the detrimental effects on convergence with increasing dimension, alluded to by the first item above, can not be improved [44, 111].

Various strategies can be used to mitigate all three effects, e.g., anisotropic weighting schemes and slow-growth rules for the polynomial subspaces, sparsifying bases for obtaining sparse systems of equations, reduced order modeling techniques for identifying important parameters and DOF, and combinations of *a priori* and *a posteriori* error estimates for constructing efficient approximations. However, these too have their limitations in application for practical parameterized PDE problems, often requiring detailed knowledge of the dependence of the solution u on the parameters $\mathbf{y}$. On the other hand, a key observation can be made from examining each of the above mitigating strategies: *computational effort can be saved if either the system required to find the solution is sparse (or can be made so), or the solution itself is well-represented by a sparse approximation*. Indeed, by exploiting sparsity, either *structured* or *unstructured*, highly-accurate approximations can be obtained with minimal complexity. With this observation in mind, this work considers the concept of sparsity from multiple viewpoints with regards to approximation of solutions to parameterized systems.

The outline of this work is the following: Chapter 2 presents the problem setting of parameterized PDEs in detail. Chapter 3 gives an overview of methods for approximating their solution, including the novel contribution of this work, an extension of standard compressed-sensing and joint-sparse recovery techniques to enable recovery of sparse Hilbert-valued vectors. Chapter 4 describes algorithms for sparse Hilbert-valued vector recovery, providing convergence results for the considered setting. Chapter 5 thoroughly explores

the computational complexity of obtaining such approximations. Chapter 6 presents numerical experiments validating our theoretical results. Finally, Chapter 7 summarizes the achievements of this work and presents possible future research directions. We encapsulate the overarching themes of this thesis, giving detailed information on key sections below.

1.1 Explicit cost bounds of stochastic Galerkin approximations for parameterized PDEs

In Section 3.2, we examine the systems of equations that result from the SG discretization of a parameterized elliptic PDE having arbitrary parametric dependence. A key challenge in the application of the SG method lies in the fact that the resulting linear systems are fully-coupled systems of physical PDE problems, requiring simultaneous solution [84, 98, 122]. For example, when finite elements (FE) are used for the physical discretization, the SG system can be written as a sum of Kronecker products of SG matrices with FE stiffness matrices, i.e., coupled FE problems [46].

Complexity of the SG method had been studied for parameterized elliptic equations when the input data in the Laplacian has *affine* dependence on the parameters [55]. In these cases, it can be shown that the Kronecker factors are sparse, and the resulting system can be solved with optimal complexity with respect to the number of *stochastic degrees of freedom* (SDOF) [7]. However, when the same equations are given input data having *nonlinear* dependence on the parameters, the resulting systems become block-dense, i.e., having many nonzero Kronecker factors [52]. Therefore, the consideration of matrix sparsity turns out to be paramount in studying the complexity of the SG method in general.

We account for this fact in our complexity analysis in Section 5.1 by deriving sharp bounds on the sparsity of the Kronecker factors for general parameterizations, and using these bounds to estimate cost in *floating point operations* (FLOPs) to solve the system. These theoretical results are supported by detailed numerical experiments, comparing the overall computational complexity of the method for a variety of *affine* and *non-affine* parameterizations in Section 6.1. Our numerical results agree with the theoretical optimality of the method in the affine case, and highlight the loss of optimality for higher order parameterizations.

1.2 Polynomial approximation via compressed sensing of high-dimensional functions on lower sets

The last few years have seen a large number of papers on the subject of adapting methods from compressed sensing for the recovery of real or complex-valued signals to high-dimensional polynomial approximation problems, see, e.g., [2, 1, 103]. In such works, the signal to be reconstructed is a vector of coefficients $\mathbf{c} \in \mathbb{C}^N$, representing a target function g in an orthonormal basis $(\Psi_\nu)_{\nu \in \mathcal{J}}$, where $\mathcal{J}$ is a multi-index set having cardinality N . In Section 3.5, we describe the standard measurement scheme for this problem, which requires a set of $m \ll N$ random measurements $(\mathbf{y}_i)_{i=1}^m \subset \mathcal{U}$ drawn independently from a PDF

ϱ . With these measurements, we define the normalized sampling matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ by $[\mathbf{A}]_{i,\nu} = \Psi_\nu(\mathbf{y}_i)/\sqrt{m}$, and normalized measurement vector $\mathbf{g} \in \mathbb{C}^m$ by $[\mathbf{g}]_i = g(\mathbf{y}_i)/\sqrt{m}$. The goal is then to solve the underdetermined system

$$\mathbf{g} = \mathbf{A}\mathbf{c} + \mathbf{e}, \quad (1.3)$$

for the signal vector $\mathbf{c} \in \mathbb{C}^N$, where $\mathbf{e} \in \mathbb{C}^m$ is an unknown noise (or truncation) vector. Because the system is underdetermined, infinitely many solution vectors $\mathbf{c}$ satisfy (1.3). However, under the additional assumptions that $\mathbf{A}$ satisfies a certain *restricted isometry property* (RIP) and the signal vector $\mathbf{c}$ is sparse (or compressible) with respect to the basis $(\Psi_\nu)_{\nu \in \mathcal{F}}$, problem (1.3) possesses a unique solution.

Sparse solutions can be found by solving the following ℓ_1 -regularized constrained optimization problem, commonly referred to as *basis pursuit denoising* (BPDN):

$$\text{minimize}_{\mathbf{z} \in \mathbb{C}^N} \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{g}\|_2 \leq \eta/\sqrt{m}. \quad (1.4)$$

Here, the ℓ_1 norm promotes sparsity, the term $\|\mathbf{A}\mathbf{z} - \mathbf{g}\|_2$ promotes data fidelity, and $\eta/\sqrt{m} > 0$ is an estimate of the normalized expansion tail or noise, i.e., $\|\mathbf{e}\|_2 \leq \eta/\sqrt{m}$. Weighted versions of (1.4), i.e., those obtained by replacing the ℓ_1 -norm with its weighted counterpart $\|\mathbf{z}\|_{\omega,1} := \sum_{\nu \in \mathcal{J}} \omega_\nu |z_\nu|$, have been shown to enhance recovery properties in nonlinear polynomial interpolation problems when the target function is known to belong to a certain weighted ℓ_p -space with $p \leq 1$. The work [103] extended recovery guarantees for problem (1.4) to the weighted setting by introducing a modified weighted-RIP condition.

Motivated by the fact that smooth functions often have rapidly decaying orthogonal expansions, whose most important terms belong to lower sets, in Section 3.5.2 we introduce a novel weighting strategy and lower set-based weakening of the standard RIP that promotes both smoothness and sparsity for function approximation. Combined with an extension of chaining arguments by Bourgain for Hadamard systems [15] to general bounded orthonormal systems, and approximation on the hyperbolic cross (i.e., the union of all lower sets of cardinality s), we show that the resulting reconstruction problem has provably lower sample complexity requirements than unweighted ℓ_1 -minimization. We complement our theoretical results with in-depth numerical experiments in Section 6.2, showing that our weighted lower set reconstruction strategy outperforms unweighted ℓ_1 -minimization and alternative weighting strategies for smooth function approximation.

1.3 Strong convergence of forward-backward splitting in reconstructing jointly-sparse signals

In Section 3.6.3 we describe the problem of joint-sparse recovery, which seeks to simultaneously reconstruct a collection of sparse signals $\{\mathbf{c}^{(r)} \in \mathbb{C}^N : r \in \Omega\}$, where Ω is a countable (possibly infinite) set, given multiple measurements $\mathbf{u}^{(r)} = \mathbf{A}\mathbf{c}^{(r)} + \mathbf{e}^{(r)} \in \mathbb{C}^m$ affected by unknown noise vectors $\mathbf{e}^{(r)} \in \mathbb{C}^m$ [39, 53, 58]. One can group the signals, measurements, and noise into matrices, and the resulting reconstruction problem becomes: find $\mathbf{c} \in \mathbb{C}^{N \times \Omega}$

satisfying

$$\mathbf{u} = \mathbf{A}\mathbf{c} + \mathbf{e}. \quad (1.5)$$

Here, we assume that the signals $\mathbf{c}^{(r)}$ possess a joint sparsity pattern, so that all but few of the rows of $\mathbf{c}$ are non-negligible, and our goal is to find a sparse matrix solution consistent with (1.5). This problem may be solved to desired accuracy via the related convex minimization problem

$$\text{minimize}_{\mathbf{z} \in \mathbb{C}^{N \times \Omega}} \|\mathbf{z}\|_{2,1} + \frac{\mu}{2} \|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{2,2}^2, \quad (1.6)$$

where the matrix norm $\|\cdot\|_{p,q}$ is defined by $\|\mathbf{z}\|_{p,q} = (\sum_{\nu \in \mathcal{J}} \|\mathbf{z}_\nu\|_p^q)^{1/q}$, and $\mu > 0$ is a penalty parameter. The mixed $\ell_{2,1}$ -norm regularization term can be seen as promoting nonsparsity along the columns via the ℓ_2 -norm and joint-sparsity along the rows via the ℓ_1 -norm.

In Section 4.1, we analyze a particular formulation of the forward-backward splitting method for solving (1.6). To establish our strong convergence result, we prove several new results on the convergence properties of the method for convex minimization over Hilbert spaces. In particular, we establish angular convergence along the elements of the support of the solution, and finite convergence on the complement of the support. The proof of angular convergence is innovative in its use of the firmly nonexpansive property for the soft-thresholding operator [10]. Our strong convergence result is then obtained by combining the finite and angular convergence results with the well-known weak convergence from [36, 38]. Unlike other strong convergence results, our proof does not require simplifying assumptions such as strict convexity of the objective function of (1.6) or finiteness of Ω (hence local compactness of the space $\mathbb{C}^{N \times \Omega}$).

1.4 Sparse reconstruction techniques for solutions of high-dimensional parametric PDEs

The compressed sensing polynomial approximation methods discussed in Section 1.2 have recently been applied to approximate functionals of solutions to parameterized PDEs with great success, see, e.g., [48, 70, 86, 95, 96]. In Section 3.5.3, we describe the problem of reconstructing a function $g : \mathbf{y} \in \mathcal{U} \mapsto G(u(\mathbf{y}))$, where the functional G represents a desired QOI of the solution u to problem (1.1). We also discuss the limitations of this approach for global approximation of u in D , noting that, in many of the aforementioned studies, G is chosen to be the point-wise evaluation functional $G_{x^*}(u(\mathbf{y})) = u(x^*, \mathbf{y})$ for a fixed $x^* \in D$. Global recovery via compressed sensing therefore requires estimates of the point-wise truncation errors in parameterizing problem (1.4) at the required points $x^* \in D$, which are unavailable for many practical problems of interest.

On the other hand, as noted above, the solutions of parameterized PDEs are often elements of a functional space $\mathcal{V}$. Given an orthonormal basis $(\Psi_\nu)_{\nu \in \mathcal{J}}$ of the N -dimensional polynomial space $P_{\mathcal{J}}(\mathcal{U}) \subset L^2_\rho(\mathcal{U})$, we can approximate the solution of (1.1) by an expansion $u_{\mathcal{J}} = \sum_{\nu \in \mathcal{J}} \mathbf{c}_\nu \Psi_\nu$, where $\mathbf{c}_\nu \in \mathcal{V}$ for all $\nu \in \mathcal{J}$. Motivated by the sparse vector and matrix recovery techniques discussed in Sections 1.2 & 1.3, in Section 3.6.1 we propose a novel

Hilbert-valued vector recovery problem that seeks to recover sparse vectors $\mathbf{c} = (\mathbf{c}_\nu)_{\nu \in \mathcal{J}}$ in the tensor-product of functional spaces $\mathcal{V}^N$. This problem is based on a reformulation of problem (1.4) in terms of the $\|\cdot\|_{\mathcal{V},q}$ -norm defined by $\|\mathbf{z}\|_{\mathcal{V},q}^q := \sum_{\nu \in \mathcal{J}} \|\mathbf{z}_\nu\|_{\mathcal{V}}^q$ for vectors $\mathbf{z} \in \mathcal{V}^N$. To parameterize this problem, we require only global estimates of the truncation error $\|u - u_{\mathcal{J}}\|_{\mathcal{V}}$. Such estimates, and requisite compressibility results for solutions, have been shown for a wide-range of parameterized problems fitting the description of problem (1.1), as discussed in Section 5.3.

In Section 3.6.3 we also describe the relation between the joint-sparse and sparse Hilbert-valued recovery problems. In particular, we show how joint-sparse recovery can be considered an instance of Hilbert-valued recovery in which the components of the vector are elements of the space $\ell_2(\mathbb{C})$, and describe situations in which both problems are actually equivalent. We also give details on combining our approach with finite element approximations, describing methods of regularizing the problem with respect to the $\|\cdot\|_{\mathcal{V},q}$ norm when a piecewise-linear FE basis is employed, noting that other methods of discretization, e.g., finite difference and finite volume methods, can also be used.

In Section 3.6.4, we show uniform recovery guarantees for the Hilbert-valued recovery approach and equivalence of the standard RIP condition with a formulation of the RIP using the $\|\cdot\|_{\mathcal{V},2}$ -norm. With these results, and estimates from our previous work [29] on the sample complexity of sparse approximation (described in Section 3.5.1), we give conditions for sparse recovery to obtain errors comparable to the best s -term approximation error in $\|\cdot\|_{\mathcal{V},1}$ -norm and error of truncation. These guarantees are then combined with quasi-optimal error estimates from [113] in Section 5.3.2 to show sub-exponential rates of convergence for our approach under the minimal sample complexity requirements for sparse recovery.

Sections 4.3 and 4.2 describe how numerical solutions can be found using reformulations of the forward-backward iterations (with a continuation strategy on the step size) and Bregman iterations for the $\|\cdot\|_{\mathcal{V},q}$ -norm version of problem (1.6). Section 6.3 provides detailed numerical results for application of these techniques to the parameterized elliptic PDE introduced in Section 2.1. Our results reveal that sparse Hilbert-valued recovery techniques can achieve error estimates comparable to the stochastic Galerkin method, with minimal sample complexity requirements on the underlying physical PDE problems. Finally, concluding remarks and possible future directions for improving the sparse Hilbert-valued recovery techniques, e.g., combining the weighted strategy of Section 1.2, are presented in Chapter 7.

Chapter 2

Parameterized partial differential equations

In this chapter, we introduce the problem setting of parameterized PDEs and the main notation used throughout this work. As discussed in Chapter 1, this work is concerned with the simultaneous solution of a family of partial differential equations on a domain $D \subset \mathbb{R}^n$, $n \in \{1, 2, 3\}$, which arises in various deterministic and stochastic modelling contexts. Let $\mathcal{U} = \prod_{j=1}^d \mathcal{U}_j \subset \mathbb{R}^d$ be the parametric domain, and $\mathcal{D}$ be a differential operator on D . We consider the following parameterized boundary value problem: for all $\mathbf{y} \in \mathcal{U}$, find $u(\cdot, \mathbf{y}) : \overline{D} \rightarrow \mathbb{R}$, such that

$$\mathcal{D}(u, \mathbf{y}) = 0, \quad \text{in } D, \quad (2.1)$$

subject to suitable boundary conditions. The solution map $\mathbf{y} \mapsto u(\cdot, \mathbf{y})$, defined from $\mathcal{U}$ into the solution space $\mathcal{V}$, typically a Sobolev space (e.g., $H_0^1(D)$), is now well-known to be smooth for a wide class of parameterized PDEs, see, e.g., [14, 25, 28, 33, 34, 71, 72, 75, 76, 77, 112], for studies on parameterized elliptic, parabolic, and hyperbolic systems, with both affine and non-affine parameterizations. In such cases, global polynomial approximation is an appealing approach to solve (2.1) with a reasonable cost. In this chapter, we provide detailed examples of parameterized PDEs fitting model (2.1). Section 2.1 discusses the *model problem* for this work, that of a parameterized elliptic PDE having either affine, polynomial, or transcendental parametric dependence, and the necessary assumptions on the parameterization to show regularity results.

2.1 Parameterized elliptic PDE

In this section, we consider the simultaneous solution of a parameterized linear elliptic PDE, having arbitrary dependence on the parameterization. The considered problem can be written: for all $\mathbf{y} \in \mathcal{U}$, find $u(\cdot, \mathbf{y}) : D \rightarrow \mathbb{R}$ such that

$$\begin{cases} -\nabla \cdot (a(x, \mathbf{y}) \nabla u(x, \mathbf{y})) &= f(x) & \forall x \in D, \mathbf{y} \in \mathcal{U} \\ u(x, \mathbf{y}) &= 0 & \forall x \in \partial D, \mathbf{y} \in \mathcal{U} \end{cases} \quad (2.2)$$

where $f \in L^2(D)$ is a fixed function of x , $D \subset \mathbb{R}^n$, $n \in \{1, 2, 3\}$, is a bounded Lipschitz domain. We will often suppress the dependence on $x \in D$, writing $a(\mathbf{y}) = a(\cdot, \mathbf{y})$ and $u(\mathbf{y}) = u(\cdot, \mathbf{y})$. In the setting of deterministic modeling, $\mathbf{y} = (y_1, \dots, y_d) \in \mathcal{U}$ is a vector of parameters modeling, e.g., material properties associated with the diffusion coefficient $a : D \times \mathcal{U} \rightarrow \mathbb{C}$, and $\mathcal{U}$ is the *parameter domain*.

In the stochastic modeling context, we assume $\mathbf{y}(\omega) = (y_1(\omega), \dots, y_d(\omega)) : \Gamma \rightarrow \mathcal{U}$ is a vector of d i.i.d. random variables, referring to $\mathcal{U}$ as the *stochastic domain*. Here Γ is the set of outcomes, and $(\Gamma, \mathcal{F}, \mathbb{P})$ is a complete probability space with $\mathcal{F} \subset 2^\Gamma$ the σ -algebra of events and $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ a probability measure associated with the joint distribution of $\mathbf{y}$,

$$F_{\mathbf{y}}(\mathbf{z}) := \mathbb{P}(\mathbf{y}(\omega) \leq \mathbf{z}) \equiv \mathbb{P}(y_1(\omega) \leq z_1, \dots, y_d(\omega) \leq z_d) : \mathcal{U} \rightarrow [0, 1],$$

i.e., using the convention that $\mathbf{y} \leq \mathbf{z}$ if and only if $y_j \leq z_j \forall j$. In the considered setting, we assume $F_{\mathbf{y}}(\mathbf{z})$ is absolutely continuous with respect to the Lebesgue measure on $\mathcal{U}$. Therefore there exists a joint *probability density function* (PDF) of $\mathbf{y}$ denoted $\varrho(\mathbf{y}) : \mathcal{U} \rightarrow \mathbb{R}_+$, with $\varrho(\mathbf{y}) = \prod_{j=1}^d \varrho_j(y_j)$. Furthermore, the probability space $(\Gamma, \mathcal{F}, \mathbb{P})$ is mapped to the space $(\mathcal{U}, \mathcal{B}(\mathcal{U}), \varrho(\mathbf{y})d\mathbf{y})$ through the stochastic parameterization, where $\mathcal{B}(\mathcal{U})$ denotes the Borel σ -algebra on $\mathcal{U}$ and $\varrho(\mathbf{y})d\mathbf{y}$ is the probability measure of $\mathbf{y}$. Hence, we may drop the dependence on ω in writing $\mathbf{y} \equiv \mathbf{y}(\omega)$ when discussing the parameterization in stochastic modeling contexts. In this probabilistic framework, the mapping $u(\mathbf{y})$ from $(\mathcal{U}, \mathcal{B}(\mathcal{U}))$ to $(\mathbb{R}, \mathcal{R})$, where $\mathcal{R}$ is the Borel σ -algebra on $\mathbb{R}$, is measurable, implying $u(\mathbf{y})$ is also a random variable distributed according to ϱ . Hence, statistical *quantities of interest* (QOI) such as expectations, variance, and probabilities of $u(\mathbf{y})$ can be computed as

$$\mathbb{E}_{\mathbf{y}}[u] \equiv \mathbb{E}[u(\mathbf{y})] := \int_{\Gamma} u(\mathbf{y}(\omega)) d\mathbb{P}(\omega) \equiv \int_{\mathcal{U}} u(\mathbf{y}) \varrho(\mathbf{y}) d\mathbf{y}, \quad (2.3)$$

$$\text{Var}_{\mathbf{y}}[u] \equiv \text{Var}[u(\mathbf{y})] := \mathbb{E}[(u(\mathbf{y}) - \mathbb{E}[u(\mathbf{y})])^2] = \mathbb{E}[u^2(\mathbf{y})] - \mathbb{E}[u(\mathbf{y})]^2, \quad (2.4)$$

$$\begin{aligned} \mathbb{P}_{\mathbf{y}}(u \in A) &\equiv \mathbb{P}(u(\mathbf{y}) \in A) := \int_{\Gamma} \mathbb{1}_A(u(\mathbf{y}(\omega))) d\mathbb{P}(\omega) \\ &\equiv \int_{\mathcal{U}} \mathbb{1}_A(u(\mathbf{y})) \varrho(\mathbf{y}) d\mathbf{y}, \quad A \in \mathcal{B}(\mathcal{U}), \end{aligned} \quad (2.5)$$

respectively, (and similarly for functionals of $u(\mathbf{y})$). We note that the results of both (2.3) & (2.5) remain functions of $x \in D$, allowing point-wise interpretation of these QOI.

To establish the desired regularity results for the polynomial approximation methods of Chapter 3, we require the following assumptions related to the continuity, coercivity, and holomorphic dependence of the coefficient $a(\mathbf{y})$. Namely:

(A1) $\exists 0 < a_{\min} \leq a_{\max} < \infty$, constants, such that $a_{\min} \leq a \leq a_{\max}$ uniformly in $\overline{D} \times \mathcal{U}$.

(A2) The complex continuation of a , represented as the map $a^* : \mathbb{C}^d \rightarrow L^\infty(D)$, is a $L^\infty(D)$ -valued holomorphic function on $\mathbb{C}^d$.

The holomorphic dependence on $\mathbf{y}$ of the coefficient a holds in many examples, including polynomial, exponential, and trigonometric functions of the variables $y_1, \dots, y_d$ shown below.

Example 2.1 (The affine case). We consider an affine function of the deterministic or random parameters, e.g.,

$$a(x, \mathbf{y}) = a_0(x) + \sum_{j=1}^d y_j b_j(x), \quad x \in \overline{D}, \mathbf{y} \in \mathcal{U}, \quad (2.6)$$

where $a_0, \{b_j\}_{j=1}^d \subset L^2(D)$ are such that $a(x, \mathbf{y})$ satisfies (A1). Specific instances of (2.6) include piecewise constant random fields and Karhunen-Loève expansions, see, e.g., [85].

Example 2.2 (The non-affine, polynomial case). We consider a non-affine, polynomial function of the random parameters, e.g.,

$$a(x, \mathbf{y}) = a_0(x) + \sum_{1 \leq |\alpha| \leq \bar{r}} c_\alpha(x) \mathbf{y}^\alpha, \quad x \in \overline{D}, \mathbf{y} \in \mathcal{U}, \quad (2.7)$$

where $\alpha = (\alpha_1, \dots, \alpha_d)$ is a multi-index with $|\alpha| := \alpha_1 + \dots + \alpha_d$ and $\mathbf{y}^\alpha := y_1^{\alpha_1} \dots y_d^{\alpha_d}$, $\bar{r} < \infty$ is the polynomial order of $a(x, \mathbf{y})$, and $a_0, \{c_\alpha\}_{1 \leq |\alpha| \leq \bar{r}} \subset L^2(D)$ are such that $a(x, \mathbf{y})$ satisfies (A1). Specific instances of (2.7) include fixed-order Taylor or orthonormal expansions of a general random field.

Example 2.3 (The non-affine, transcendental case). We consider a non-affine, transcendental function of the random parameters, e.g.,

$$a(x, \mathbf{y}) = a_0(x) + g(x, \mathbf{y}), \quad x \in \overline{D}, \mathbf{y} \in \mathcal{U}, \quad (2.8)$$

where $a_0, g(\cdot, \mathbf{y}) \subset L^2(D)$, and $g(x, \mathbf{y})$ is a general transcendental function of x and $\mathbf{y}$, such that $a(x, \mathbf{y})$ satisfies (A1). Specific instances of suitable $g(x, \mathbf{y})$ in (2.8) include composing (2.6) or (2.7) with the trigonometric sine and cosine functions, their hyperbolic counterparts, or exponential functions, e.g., the well-known log-transformed Karhunen-Loève expansion [5].

Let $L_\varrho^2(\mathcal{U})$ be the space of square integrable functions with respect to the measure $\varrho(\mathbf{y})d\mathbf{y}$ and $L_\varrho^\infty(\mathcal{U})$ be the space of essentially bounded functions, with the norm

$$\|u\|_{L_\varrho^\infty(\mathcal{U})} := \operatorname{ess\,sup}_{\mathbf{y} \in \mathcal{U}} |u(\mathbf{y})|,$$

where the essential supremum is taken with respect to the weight ϱ . These spaces are relevant to both deterministic and stochastic modeling contexts, as ϱ may be taken as the identity. By $H^{-1}(D)$ we denote the dual of $H_0^1(D)$, the space of square integrable functions in D having zero trace on the boundary and square integrable distributional derivatives. We will often use the abbreviation $\mathcal{V} \equiv \mathcal{V}(D) = H_0^1(D)$. Also, $\mathcal{H}_\varrho^2$ and $\mathcal{H}_\varrho^\infty$ will occasionally be used to denote the spaces

$$\begin{aligned} L_\varrho^2(\mathcal{U}; \mathcal{V}) &:= \left\{ u : \overline{D} \times \mathcal{U} \rightarrow \mathbb{R} : u \text{ strongly measurable and } \int_{\mathcal{U}} \|u\|_{\mathcal{V}}^2 \varrho(\mathbf{y}) d\mathbf{y} < \infty \right\}, \\ L_\varrho^\infty(\mathcal{U}; \mathcal{V}) &:= \left\{ u : \overline{D} \times \mathcal{U} \rightarrow \mathbb{R} : u \text{ strongly measurable \& ess\,sup}_{\mathbf{y} \in \mathcal{U}} \|u(\cdot, \mathbf{y})\|_{\mathcal{V}} < \infty \right\}, \end{aligned}$$

respectively, especially when referring to specific norms where notation becomes cumbersome. For the space $\mathcal{V}$ we have the energy norm $\|v\|_{\mathcal{V}} = \|\nabla v\|_{L^2(D)}$, hence $\mathcal{H}_\varrho^2$ is a Hilbert space with norm $\|v\|_{\mathcal{H}_\varrho^2}^2 = \int_{\mathcal{U}} \|v\|_{\mathcal{V}}^2 \varrho d\mathbf{y}$. The *stochastic weak form* of problem (2.2) is given by: find $u \in \mathcal{H}_\varrho^2$ such that $\forall v \in \mathcal{H}_\varrho^2$

$$\int_{\mathcal{U}} \mathcal{B}[u, v](\mathbf{y}) \varrho(\mathbf{y}) d\mathbf{y} = \int_{\mathcal{U}} F(v) \varrho(\mathbf{y}) d\mathbf{y}, \quad (2.9)$$

where

$$\mathcal{B}[u, v](\mathbf{y}) = \int_D a(x, \mathbf{y}) \nabla u(x, \mathbf{y}) \cdot \nabla v(x, \mathbf{y}) dx, \quad F(v) = \int_D f(x) v(x, \mathbf{y}) dx. \quad (2.10)$$

For convenience, we will often use the abbreviation $\mathcal{B}(\mathbf{y}) = \mathcal{B}[\cdot, \cdot](\mathbf{y})$. It follows from (A1) that $\mathcal{B}(\mathbf{y})$ is a symmetric, uniformly coercive, and continuous bilinear operator on $\mathcal{V}$, parameterized by $\mathbf{y} \in \mathcal{U}$, and $\mathcal{B}(\mathbf{y})$ induces the norm

$$\|u\|_{\mathcal{B}(\mathbf{y})}^2 := \int_D a(x, \mathbf{y}) |\nabla u|^2 dx. \quad (2.11)$$

Assumption (A1) and the Lax-Milgram lemma also ensure the existence and uniqueness of the solution u to (2.9) in $\mathcal{H}_\varrho^2$.

The convergence of the global polynomial methods described in Chapter 3 exploits the uniform ellipticity of the coefficient $a(\mathbf{y})$ and regularity of the solution $u(\mathbf{y})$ with respect to $\mathbf{y}$, guaranteed for problem (2.2) by (A1) & (A2). We describe the necessary concepts and results, which can be found in [113], as follows. By $\text{Re}(z)$ and $\text{Im}(z)$ we denote the real and imaginary parts of $z \in \mathbb{C}$, and for $0 < \delta < a_{\min}$ from (A1), we define

$$U(a, \delta) = \{z \in \mathbb{C}^d : \text{Re}(a^*(x, z)) \geq \delta, \forall x \in \overline{D}\}. \quad (2.12)$$

If $U(a, \delta) \neq \emptyset$ for some $0 < \delta < a_{\min}$, we say that $a(x, z)$ is uniformly elliptic on the set $U(a, \delta)$ and we refer to $U(a, \delta)$ as its domain of uniform ellipticity. For $\gamma = (\gamma_1, \dots, \gamma_d)$ with $\gamma_j > 1 \forall j$, we define the polyellipse

$$\mathcal{E}_\gamma = \bigotimes_{1 \leq j \leq d} \left\{ z_j \in \mathbb{C} : \text{Re}(z_j) \leq \frac{\gamma_j + \gamma_j^{-1}}{2} \cos \theta, \text{Im}(z_j) \leq \frac{\gamma_j - \gamma_j^{-1}}{2} \sin \theta, \theta \in [0, 2\pi) \right\}.$$

In [113, Lemma 1] it was shown that if $a(\mathbf{y})$ satisfies (A1) and (A2), then for any $0 < \delta < a_{\min}$ there exists a $\gamma = (\gamma_1, \dots, \gamma_d)$ with $\gamma_j > 1 \forall j$ such that $\mathcal{E}_\gamma \subset U(a, \delta)$. Similarly, we can define the polydisc $\mathcal{D}_\gamma = \bigotimes_{1 \leq j \leq d} \{z_j \in \mathbb{C} : |z_j| \leq \gamma_j\}$, though, for arbitrary $0 < \delta < a_{\min}$, it is not always possible to find a γ with $\gamma_j > 1 \forall j$ such that $\mathcal{D}_\gamma \subset U(a, \delta)$. Figure 2.1 provides an illustration of this fact for various one-dimensional coefficients $a(\mathbf{y})$, $\mathbf{y} \in \mathbb{C}$. Note that in the case of the 6th degree polynomial and exponential random variables, no disc of radius $\gamma > 1$ containing $\mathcal{U} = [-1, 1]$ can fit in the region.

The following theorem, combining Lemma 1 and Theorem 1 from [113], shows the regularity of the solution u with respect to the parameterization under (A1) & (A2).

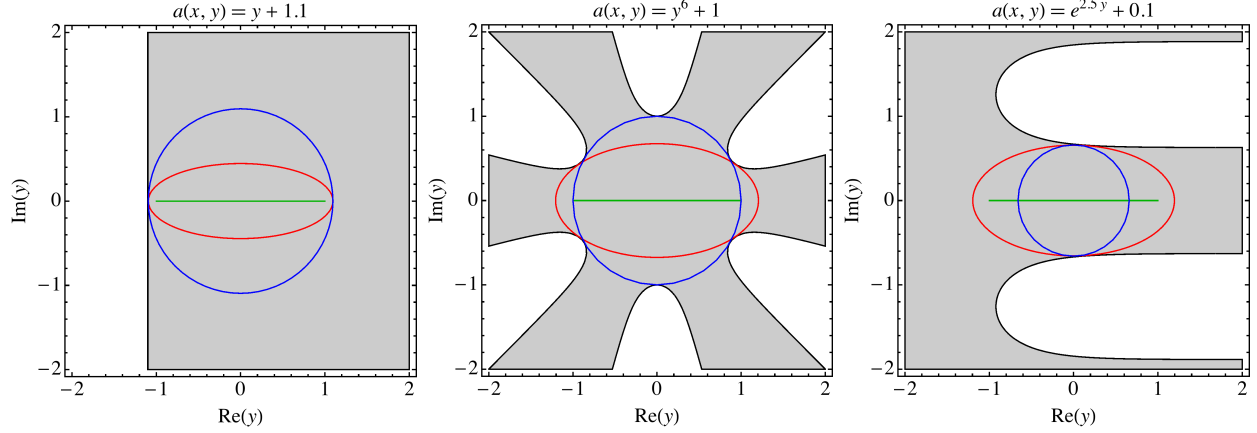


Figure 2.1: Domains of uniform ellipticity for some one-dimensional coefficients $a(x, y)$ are indicated by the gray regions in each plot. The blue and red curves represent the maximal discs $\mathcal{D}_\gamma$ and ellipses $\mathcal{E}_\gamma$, respectively, that can be contained in those domains, and the green lines represent the interval $\mathcal{U} = [-1, 1]$.

Theorem 2.4. *When the coefficient $a(x, \mathbf{y})$ satisfies (A1) and (A2), so that for some $0 < \delta < a_{\min}$ and $\gamma = (\gamma_1, \dots, \gamma_d)$ with $\gamma_j > 1 \ \forall j$ we have $\mathcal{E}_\gamma \subset U(a, \delta)$, then the function $\mathbf{z} \mapsto u(\mathbf{z})$ from (2.2) is holomorphic in an open neighborhood of $\mathcal{E}_\gamma$.*

This result states that a direct consequence of the uniform ellipticity of the function $a(x, \mathbf{y})$ on the polyellipse $\mathcal{E}_\gamma \subset U(a, \delta)$ is that the solution u of (2.2) has smooth analytic dependence on the parameterization $\mathbf{y}$. However, this result does not tie convergence rates of global polynomial approximations to the maximal parameter γ such that $\mathcal{E}_\gamma \subset U(a, \delta)$. The connection between γ , dimension d , and the sparsity level s in constructing sparse approximations to u is the main subject of [113], and will be discussed in greater detail in Section 5.3 while developing error estimates for the sparse Hilbert-valued recovery techniques of Section 3.6. Nonetheless, results such as Theorem 2.4 are the key motivation in applying the global polynomial approximation methods described in Chapter 3 to approximating solutions u to problems fitting (2.1), and have been shown for a variety of parameterized problems including parabolic and hyperbolic PDEs.

Chapter 3

Overview of methods

In this chapter, we discuss many of the popular methods for solving the parameterized PDEs introduced in Chapter 2. We begin in Section 3.1 with a description of the finite element method for discretizing problems such as (2.1) over the physical domain D . In Section 3.2, we describe application of the stochastic Galerkin method with orthonormal bases for the parametric discretization, including details of the resulting linear systems in the case of a few relevant parameterizations. In Section 3.3 we describe the Monte Carlo method, which seeks to directly approximate QOI by empirical averaging. Section 3.4 introduces the sparse grid-based stochastic collocation method, which constructs approximations from tensorizations and hierarchical differences of one-dimensional Lagrange interpolating operators. Section 3.5 describes compressive-sensing-based strategies for obtaining polynomial approximations of high-dimensional functions with applications to parameterized PDEs, and explains recovery guarantees in the context of structured sparsity and lower-set-based approximation. Finally Section 3.6 details the main theoretical contribution and focus of this thesis, a novel sparse Hilbert-valued vector reconstruction technique, relating this approach to the problem of joint-sparse recovery, and provides a framework for obtaining error estimates for approximating solutions to parameterized PDEs.

3.1 Finite element method for physical discretization

Equations such as (2.1) can be discretized over the physical domain using, e.g., finite elements, finite differences, or finite volume methods. Depending on the type of underlying PDE, certain methods may be more favorable to use than others. In this work, we elect to use the finite element method, noting that our techniques for sparse polynomial approximation outlined in Section 3.6.1 can be applied to any of the aforementioned methods. Therefore, we briefly present details relevant to the finite element discretization used for obtaining semi-discrete approximations of u over D when $\mathbf{y} \in \mathcal{U}$ is fixed. Let $(\mathcal{T}_h)_{h>0}$, be a family of shape regular triangulations of D , parameterized by maximum mesh size $h \rightarrow 0$, and let $\mathcal{V}_h \equiv \mathcal{V}_h(D) \subset \mathcal{V}(D)$ be the finite element space of piecewise continuous polynomials on $\mathcal{T}_h$. Given a basis $(\varphi_k)_{k=1}^{K_h}$ of $\mathcal{V}_h$, we consider the solution of the following problem: *find*

$u_h(\mathbf{y}) \in \mathcal{V}_h$ such that $\forall v \in \mathcal{V}_h$

$$\mathcal{B}[u_h(\mathbf{y}), v](\mathbf{y}) = F(v), \quad (3.1)$$

where $\mathcal{B}(\mathbf{y})$ and F are defined in (2.10). For almost every $\mathbf{y} \in \mathcal{U}$, problem (3.1) admits a unique solution of the form

$$u_h(x, \mathbf{y}) = \sum_{k=1}^{K_h} u_k(\mathbf{y}) \varphi_k(x), \quad (3.2)$$

referred to as the Galerkin projection of $u(\mathbf{y})$ onto $\mathcal{V}_h$. We discretize this system by defining the parameterized matrix equations, for each $l, k = 1, \dots, K_h$

$$[\mathbf{B}(\mathbf{y})]_{l,k} = \mathcal{B}[\varphi_k, \varphi_l](\mathbf{y}), \quad \mathbf{F}_k = F(\varphi_k). \quad (3.3)$$

Here $\mathbf{B}(\mathbf{y})$ is a parameterized finite element stiffness matrix, and the coefficients $\mathbf{c}_h(\mathbf{y}) = (u_1(\mathbf{y}), u_2(\mathbf{y}), \dots, u_{K_h}(\mathbf{y}))^T \in \mathbb{C}^{K_h}$ of $u_h(x, \mathbf{y})$ from (3.2) are determined by solving the linear system

$$\mathbf{B}(\mathbf{y})\mathbf{c}_h(\mathbf{y}) = \mathbf{F}, \quad (3.4)$$

at fixed realizations of $\mathbf{y} \in \mathcal{U}$. For the numerical experiments in Chapter 6, we apply the FE discretization with a piece-wise linear basis. More specifically, given a triangulation $\mathcal{T}_h$ having nodes $(x_k)_{k=1}^{K_h}$, the basis $(\varphi_k)_{k=1}^{K_h}$ is the unique basis of piece-wise linear polynomials satisfying the *Lagrange nodal basis property* for $j, k = 1, \dots, K_h$

$$\varphi_j(x_k) = \begin{cases} 1, & j = k, \\ 0, & j \neq k. \end{cases}$$

In the considered setting of problem (2.2) under Assumption (A1), the bilinear form $\mathcal{B}(\mathbf{y})$ is symmetric, coercive, and continuous, guaranteeing $\mathbf{B}(\mathbf{y})$ is symmetric and positive-definite. Hence (3.4) can be solved by iterative methods such as the conjugate gradient (CG) method. A more thorough discussion of the approximation properties of our choice of basis and errors of iterative solvers can be found in the ε -complexity analysis of Section 5.1.3.

3.2 Stochastic Galerkin method

In this section we explain the combined approach of stochastic Galerkin and finite element approximations for constructing fully discrete approximations to the solution u of problem (2.2). While both methods employ a Galerkin projection over their respective domains, the stochastic Galerkin method is generally applied with a globally-defined orthonormal bases, chosen with respect to the underlying orthogonalization measure ϱ in problem (2.1). This enables fast, sub-exponential rates of convergence to be shown for such approximations, see, e.g., [112]. In Section 3.2.2 we describe the linear systems that result from the SG

discretization when Examples 2.1, 2.2, and 2.3 are used in problem (2.2), saving a thorough examination of the computational complexity of these cases for Section 5.1.1.

3.2.1 Stochastic Galerkin approximation with an orthogonal basis

In this section we describe techniques for discretization over the parameter domain $\mathcal{U}$ by means of orthonormal Galerkin projection. These techniques are relevant to the methods of compressed sensing and sparse Hilbert-valued approximation introduced in Sections 3.5 & 3.6. Defining $\mathcal{F} := \mathbb{N}_0^d = \{\boldsymbol{\nu} = (\nu_1, \dots, \nu_d) : \nu_k \in \mathbb{N}_0\}$, in the considered setting, the solution $u(x, \mathbf{y})$ of the system (2.1) can be expanded in an $L_\varrho^2(\mathcal{U})$ -orthonormal basis $(\Psi_\nu)_{\nu \in \mathcal{F}}$ according to

$$u(x, \mathbf{y}) = \sum_{\nu \in \mathcal{F}} \mathbf{c}_\nu(x) \Psi_\nu(\mathbf{y}), \quad (3.5)$$

where $\Psi_\nu = \prod_{j=1}^d \Psi_{\nu_j}$ are tensor products of $L_{\varrho_j}^2(\mathcal{U}_j)$ -orthonormal polynomials, and the coefficients $\mathbf{c}_\nu$ belong to the space $\mathcal{V}$. The series (3.5) is referred to as the *generalized polynomial chaos* (GPC) expansion of u (see, e.g., [61, 123]), whose convergence rates are well-understood. We are particularly interested in the cases that ϱ is the uniform or Chebyshev density, and for the sake of clarity refer to the corresponding Legendre and Chebyshev bases of $L_\varrho^2(\mathcal{U})$ as $(L_\nu)_{\nu \in \mathcal{F}}$ and $(T_\nu)_{\nu \in \mathcal{F}}$.

For any finite subset $\mathcal{J} \subset \mathcal{F}$, we define the $\mathcal{V}$ -valued polynomial subspace

$$\mathcal{V}_\mathcal{J} := \left\{ \sum_{\nu \in \mathcal{J}} \hat{\mathbf{c}}_\nu(x) \Psi_\nu(\mathbf{y}) : \hat{\mathbf{c}}_\nu \in \mathcal{V} \right\}, \quad (3.6)$$

and, with an abuse of notation, often use $\mathcal{V}^N$ to denote $\mathcal{V}_\mathcal{J}$ when the particular set $\mathcal{J}$ with $N = \#(\mathcal{J})$ is understood. It follows that $\mathcal{V}_\mathcal{J} := P_\mathcal{J}(\mathcal{U}) \otimes \mathcal{V} \subset L_\varrho^2(\mathcal{U}; \mathcal{V})$ where $P_\mathcal{J}(\mathcal{U}) := \text{span}\{\Psi_\nu(\mathbf{y}) : \nu \in \mathcal{J}\}$. The Galerkin projection of u onto $\mathcal{V}_\mathcal{J}$ is given by

$$u_\mathcal{J}(x, \mathbf{y}) = \sum_{\nu \in \mathcal{J}} \mathbf{c}_\nu(x) \Psi_\nu(\mathbf{y}). \quad (3.7)$$

An efficient evaluation of $u_\mathcal{J}$ would require the index set $\mathcal{J}$ to enclose all effective multi-indices (i.e., corresponding to largest $\|\mathbf{c}_\nu\|_\mathcal{V}$) so that

$$\|u - u_\mathcal{J}\|_{L_\varrho^2(\mathcal{U}; \mathcal{V})}^2 = \|u_{\mathcal{J}^c}\|_{L_\varrho^2(\mathcal{U}; \mathcal{V})}^2 = \left\| \sum_{\nu \notin \mathcal{J}} \mathbf{c}_\nu \Psi_\nu \right\|_{L_\varrho^2(\mathcal{U}; \mathcal{V})}^2 = \sum_{\nu \notin \mathcal{J}} \|\mathbf{c}_\nu\|_\mathcal{V}^2 \quad (3.8)$$

is minimized. This condition can be easily fulfilled if we are able to select $\mathcal{J}$ to be a very large index set. However, for traditional approximation approaches, the size of $\mathcal{J}$ is often constrained by the computational budget, as the cost of computing grows at least linearly in N , while N often grows exponentially in d for many choices of $\mathcal{J}$, see, e.g., [7, 26, 31, 46, 65]. This consideration of cost motivates the use of adaptive strategies which seek to construct $\mathcal{J}$ in a (near) optimal manner, i.e., including in $\mathcal{J}$ only those multi-indices that likely correspond to large polynomial coefficients, see, e.g., [8, 113].

Remark 3.1 (Best s -term and quasi-optimal approximations). *With the estimate (3.8) and following discussion in mind, the optimal choice of $\mathcal{J}$ would be the set of cardinality s such that the corresponding approximation provides maximum accuracy out of all sets of size s . Such approximations are referred to as best s -term approximations, and recently much work has focussed on proving conditions for nonlinear (and non-constructive) best s -term Taylor, Chebyshev, and Legendre approximations to satisfy algebraic rates of convergence in the setting of infinite-dimensional parametric dependence, see, e.g., [14, 25, 28, 33, 34, 71, 72, 75, 76, 77]. On the other hand, constructive quasi-optimal approximations seek to approximate the solution u by choosing $\mathcal{J}$ corresponding to the s largest bounds on the coefficients $\|\mathbf{c}_\nu\|_\nu$, see, e.g., [8, 9, 113]. In Section 5.3.1, we present a more thorough discussion on both best s -term and quasi-optimal approximations, and the necessary assumptions on the bounds on the coefficients that enable sparse Hilbert-valued recovery.*

Many studies of SG approximations ignore issues of computational complexity with respect to choice of $\mathcal{J}$ in studying the SG method, opting instead to fix an index set $\mathcal{J}$ (or rule for growing the set) *a priori*. Without knowledge of the pattern of decay of the coefficients of (3.8), the “safe” choice is to apply isotropic approximation techniques, treating all directions as equally important. Common examples of isotropic multi-index sets include

$$\begin{aligned} \mathcal{J}_p^{\text{TP}} &= \left\{ \nu \in \mathcal{F} : \max_{1 \leq j \leq d} \nu_j \leq p \right\}, & \mathcal{J}_p^{\text{TD}} &= \left\{ \nu \in \mathcal{F} : \sum_{j=1}^d \nu_j \leq p \right\}, \\ \mathcal{J}_p^{\text{HC}} &= \left\{ \nu \in \mathcal{F} : \prod_{j=1}^d (\nu_j + 1) \leq p \right\}, & \mathcal{J}_p^{\text{SM}} &= \left\{ \nu \in \mathcal{F} : \sum_{j=1}^d f(\nu_j) \leq f(p) \right\}, \\ f(p) &= \begin{cases} 0, & p = 0 \\ 1, & p = 1 \\ \lceil \log_2(p) \rceil, & p \geq 2 \end{cases} \end{aligned} \quad (3.9)$$

corresponding to the Tensor Products (TP), Total Degree (TD), Hyperbolic Cross (HC), and Smolyak (SM) polynomial spaces $P_{\mathcal{J}_p^{\text{TP}}}(\mathcal{U})$, $P_{\mathcal{J}_p^{\text{TD}}}(\mathcal{U})$, $P_{\mathcal{J}_p^{\text{HC}}}(\mathcal{U})$, and $P_{\mathcal{J}_p^{\text{SM}}}(\mathcal{U})$, respectively. Any choice of these sets can be used in constructing the SG approximation to u , with some possessing more favorable approximation properties than others. For example, Table 3.1 exhibits the difference in stochastic degrees of freedom (SDOF), i.e., the cardinality N_p associated with a choice of index set $\mathcal{J}_p$, for the TD and TP choices of subspace over a range of dimensions for orders $p = 3, 5$. When the solution u exhibits an anisotropic dependence on the parameters $\mathbf{y}$, anisotropic weighted versions of the index sets defined in (3.9) can be introduced, further reducing the number of SDOF needed to approximate u at a desired accuracy. For related work on anisotropic approximations, see, e.g., [7, 92].

Remark 3.2 (Approximating stochastic Galerkin expansions). *There are multiple ways to approximate the Galerkin projection $u_{\mathcal{J}}$ from (3.7), including interpolatory approaches. Discrete least-squares can be applied by formulating a measurement system and solving the related overdetermined polynomial interpolation problem [26, 35]. The sparse Hilbert-valued vector recovery techniques outlined in Section 3.6 also attempt to find sparse approximations*

Table 3.1: Comparison of $N_p = \#(\mathcal{J}_p)$ SDOF for the multi-index sets $\mathcal{J}_p = \mathcal{J}_p^{\text{TD}}$ (total degree) and $\mathcal{J}_p = \mathcal{J}_p^{\text{TP}}$ (tensor product), where $d = \dim(\mathcal{U})$ and p is the maximal degree.

d	p	N_p using TD index set	N_p using TP index set
3	3	20	64
	5	56	216
5	3	56	1024
	5	252	7776
10	3	282	1048576
	5	3003	60466176
20	3	1771	$> 1 \times 10^{12}$
	5	53130	$> 3 \times 10^{15}$
100	3	176851	$> 1 \times 10^{60}$
	5	96560646	$> 6 \times 10^{77}$

to the solution vector $\mathbf{c} = (\mathbf{c}_\nu)_{\nu \in \mathcal{J}}$ from (3.7), and can be considered a type of sparse nonlinear interpolation problem [102]. In the interpolatory setting, the resulting systems of equations are totally decoupled and can be solved independently and in parallel. In general however, when discretizing problems such as (2.9) with the stochastic Galerkin and finite element methods, the resulting system is composed of fully coupled finite element systems which must be solved simultaneously.

We detail the discretization procedure as follows. With $(\varphi_k)_{k=1}^{K_h}$ as in Section 3.1 and $(\Psi_\nu)_{\nu \in \mathcal{J}_p}$ as above, we can write the fully discrete SG approximation as

$$u_{h,p}(x, \mathbf{y}) = \sum_{\nu \in \mathcal{J}_p} \sum_{k=1}^{K_h} u_{k,\nu} \varphi_k(x) \Psi_\nu(\mathbf{y}), \quad (3.10)$$

whose coefficients can be found by solving the following coupled problem: find $u_{h,p} \in \mathcal{V}_h(D) \otimes P_{\mathcal{J}_p}(\mathcal{U})$ such that for all $v \in \mathcal{V}_h(D) \otimes P_{\mathcal{J}_p}(\mathcal{U})$

$$\mathbb{E} [\mathcal{B}[u_{h,p}, v](\mathbf{y})] = \mathbb{E} [F(v)], \quad (3.11)$$

where $\mathcal{B}[\cdot, \cdot](\mathbf{y})$ and $F(\cdot)$ are defined in (2.10). To form the linear system of equations resulting from the SG approximation given by (3.10), we let

$$\mathbf{u}_{h,\nu} = (u_{1,\nu}, \dots, u_{K_h,\nu})^T \quad (3.12)$$

be the vector of nodal values of the finite element solution corresponding to the ν -th stochastic mode of $u_{h,p}$, and $\mathbf{u}_{h,p} = (\mathbf{u}_{h,\nu})_{\nu \in \mathcal{J}_p}^T$. Observe that when f is deterministic

$$\langle \Psi_\nu, \mathbf{F}_k \rangle := \int_{\mathcal{U}} \Psi_\nu(\mathbf{y}) \mathbf{F}_k \varrho(\mathbf{y}) d\mathbf{y} = \mathbf{F}_k \delta_{\mathbf{0},\nu},$$

for all $k = 1, \dots, K_h$, where $\delta_{\mathbf{0}, \boldsymbol{\nu}} = 1$ if $\boldsymbol{\nu} = \mathbf{0}$ and $\delta_{\mathbf{0}, \boldsymbol{\nu}} = 0$ otherwise. Performing a Galerkin projection onto $\text{span}\{(\Psi_{\boldsymbol{\nu}})_{\boldsymbol{\nu} \in \mathcal{J}_p}\}$ for the solution of (3.11) yields the following system: for each $\boldsymbol{\nu} \in \mathcal{J}_p$

$$\sum_{\boldsymbol{\mu} \in \mathcal{J}_p} \langle \Psi_{\boldsymbol{\nu}}(\mathbf{y}), \mathbf{B}(\mathbf{y}) \Psi_{\boldsymbol{\mu}}(\mathbf{y}) \rangle \mathbf{u}_{h, \boldsymbol{\mu}} = \langle \Psi_{\boldsymbol{\nu}}(\mathbf{y}), \mathbf{F} \rangle, \quad (3.13)$$

which can be written algebraically as a system of *fully coupled finite element problems*: for each $\boldsymbol{\nu} \in \mathcal{J}_p$

$$\sum_{\boldsymbol{\mu} \in \mathcal{J}_p} [\mathbf{K}]_{\boldsymbol{\nu}, \boldsymbol{\mu}} \mathbf{u}_{h, \boldsymbol{\mu}} = \mathbf{F} \delta_{\mathbf{0}, \boldsymbol{\nu}} \quad (3.14)$$

with $[\mathbf{K}]_{\boldsymbol{\nu}, \boldsymbol{\mu}} = \langle \Psi_{\boldsymbol{\nu}}(\mathbf{y}), \mathbf{B}(\mathbf{y}) \Psi_{\boldsymbol{\mu}}(\mathbf{y}) \rangle$ and $\mathbf{B}(\mathbf{y})$ as given in (3.3). We say that $\mathbf{K}$ is *block-dense* when $[\mathbf{K}]_{\boldsymbol{\nu}, \boldsymbol{\mu}} \neq \mathbf{0}$ for all $\boldsymbol{\nu}, \boldsymbol{\mu} \in \mathcal{J}_p$ and refer to the *block-sparsity* of $\mathbf{K}$ as the number of *nonzero blocks* $[\mathbf{K}]_{\boldsymbol{\nu}, \boldsymbol{\mu}}$ over all choices $\boldsymbol{\nu}, \boldsymbol{\mu} \in \mathcal{J}_p$.

Remark 3.3 (Storage and computation with $\mathbf{K}$). *Matrix-free methods are typically applied to solve (3.14) without ever explicitly forming $\mathbf{K}$ in a continuous block of memory, as described in [94]. When the resulting system is sparse, which occurs, e.g., when $a(x, \mathbf{y})$ is affine as in Example 2.1, this can lead to computationally efficient solution strategies having near optimal dependence on N_p . However, implementation of these matrix-free methods relies on the fact that the coefficient $a(x, \mathbf{y})$ has a separable representation in x and $\mathbf{y}$, e.g., $a(x, \mathbf{y}) = \sum_{j=1}^d b_j(x) c_j(\mathbf{y})$, so that (3.14) essentially decouples into a sum of Kronecker products (more details on this in the next section). For the transcendental function $a(x, \mathbf{y})$ from Example 2.3, this may not be the case, and one may be forced to perform an additional expansion of the coefficient (a costly procedure in high-dimensions). In the worst-case scenario, when $\mathbf{K}$ is entirely block-dense, matrix-vector multiplications require $\mathcal{O}(K_h N_p^2)$ floating point operations. For many practical problems, this quadratic dependence on N_p makes the solution of (3.14) unfeasible.*

3.2.2 Representations of the data and resulting linear systems

As mentioned in Remark 3.3, for a general coefficient $a(x, \mathbf{y})$, the matrix $\mathbf{K}$ in (3.14) requires the storage of at most N_p^2 block matrices of the size and sparsity of $\mathbf{B}(\mathbf{y})$, i.e., $\mathcal{O}(K_h N_p^2)$ elements. However, in several specific cases the actual block-sparsity of $\mathbf{K}$ is much less. We recall the coefficient from Example 2.1, where $\mathbf{K}$ can be written

$$[\mathbf{K}]_{\boldsymbol{\nu}, \boldsymbol{\mu}} = \langle \Psi_{\boldsymbol{\nu}}(\mathbf{y}), \Psi_{\boldsymbol{\mu}}(\mathbf{y}) \rangle \mathbf{B}_0 + \sum_{j=1}^d \langle y_j \Psi_{\boldsymbol{\nu}}(\mathbf{y}), \Psi_{\boldsymbol{\mu}}(\mathbf{y}) \rangle \mathbf{B}_j,$$

with $[\mathbf{B}_0]_{l, k} = \int_D a_0(x) \nabla \phi_k(x) \cdot \nabla \phi_l(x) dx$ and $[\mathbf{B}_j]_{l, k} = \int_D b_k(x) \nabla \phi_k(x) \cdot \nabla \phi_l(x) dx$. Defining the SG matrices $[\mathbf{G}_0]_{\boldsymbol{\nu}, \boldsymbol{\mu}} = \langle \Psi_{\boldsymbol{\nu}}(\mathbf{y}), \Psi_{\boldsymbol{\mu}}(\mathbf{y}) \rangle$ and $[\mathbf{G}_j]_{\boldsymbol{\nu}, \boldsymbol{\mu}} = \langle y_j \Psi_{\boldsymbol{\nu}}(\mathbf{y}), \Psi_{\boldsymbol{\mu}}(\mathbf{y}) \rangle$, $\mathbf{K}$ can be

represented by

$$\mathbf{K} = \mathbf{G}_0 \otimes \mathbf{B}_0 + \sum_{j=1}^d \mathbf{G}_j \otimes \mathbf{B}_j, \quad (3.15)$$

where $\mathbf{G} \otimes \mathbf{B}$ denotes the Kronecker product of $\mathbf{G}$ and $\mathbf{B}$. We note that a similar construction can be obtained for any coefficient $a(x, \mathbf{y})$ which can be written as a sum of separable functions of x and $\mathbf{y}$, such as the polynomial function of Example 2.2.

However, when $a(x, \mathbf{y})$ is not separable in x and $\mathbf{y}$, this construction is no longer valid, and the resulting matrix $\mathbf{K}$ may be block-dense if we simply carry out the Galerkin projections and compute $\mathbf{K}$ directly. In certain cases, e.g., when the diffusion coefficient is given by a log-transformed Karhunen-Loève expansion of a random field, the problem can be reformulated as a convection-diffusion problem. Because the reformulation results in an affine parameterization in the elliptic operator of (2.2), the resulting system is close to block-diagonal, and can be solved much more efficiently than the original problem [117]. However, attention to details such as the Péclet number associated with the reformulation is required.

In general, such reformulation strategies are unavailable, and other methods of simplifying the problem must be used. For a general transcendental coefficient $a(x, \mathbf{y})$, as given in Example 2.3, it can be advantageous to project the coefficient onto an additional subspace $P_{\mathcal{J}_r}(\mathcal{U})$, $r \in \mathbb{N}_0$, in order to obtain a separable representation. With the same basis $(\Psi_\alpha)_{\alpha \in \mathcal{F}}$ as in Section 3.2.1, we consider the problem of truncating the formal expansion of the coefficient $a(x, \mathbf{y}) = \sum_{\alpha \in \mathcal{F}} a_\alpha(x) \Psi_\alpha(\mathbf{y})$, with coefficients $a_\alpha(x) = \int_{\mathcal{U}} a(x, \mathbf{y}) \Psi_\alpha(\mathbf{y}) \varrho(\mathbf{y}) d(\mathbf{y})$. Let $\mathcal{J}_r$ be an index set of the type described in Section 3.2.1 and define the Galerkin projection

$$a^r(x, \mathbf{y}) = \sum_{\alpha \in \mathcal{J}_r} a_\alpha(x) \Psi_\alpha(\mathbf{y}) \quad (3.16)$$

onto the subspace $P_{\mathcal{J}_r}(\mathcal{U})$, where $r \in \mathbb{N}_0$ is to be chosen later. Now the approximation (3.16) is a sum of separable functions in x and $\mathbf{y}$, and can be considered an instance of Example 2.2 when $(\Psi_\alpha)_{\alpha \in \mathcal{F}}$ are polynomials. In the case that $a^r(x, \mathbf{y}) \neq a(x, \mathbf{y})$, e.g., in the case that the projection order r is chosen to minimize error independent of the SG discretization, we let $u_{h,p}^r$ denote the corresponding solution to the fully discrete SG problem with $a(x, \mathbf{y})$ replaced with $a^r(x, \mathbf{y})$. Substituting $a^r(x, \mathbf{y})$ into (2.10), we obtain

$$\int_D \left(\sum_{\alpha \in \mathcal{J}_r} a_\alpha(x) \Psi_\alpha(\mathbf{y}) \right) \nabla \phi_k(x) \cdot \nabla \phi_l(x) dx = \sum_{\alpha \in \mathcal{J}_r} [\mathbf{B}_\alpha]_{l,k} \Psi_\alpha(\mathbf{y}), \quad (3.17)$$

$$[\mathbf{B}_\alpha]_{l,k} = \int_D a_\alpha(x) \nabla \phi_k(x) \cdot \nabla \phi_l(x) dx. \quad (3.18)$$

Equation (3.17) represents the expansion of the stochastic finite element stiffness matrix $\mathbf{B}(\mathbf{y})$ in the basis $(\Psi_\alpha)_{\alpha \in \mathcal{J}_r}$ and equation (3.18) represents the α -th mode of the expansion.

Let

$$\mathbf{u}_{h,\nu}^r = (u_{1,\nu}^r, \dots, u_{K_h,\nu}^r)^T \quad (3.19)$$

denote the vector of nodal values of the finite element solution corresponding to the ν -th stochastic mode of $u_{h,p}^r$, and $\mathbf{u}_{h,p}^r = (\mathbf{u}_{h,\nu}^r)_{\nu \in \mathcal{J}_p}$. We substitute the truncated expansion of $\mathbf{B}(\mathbf{y})$ into the Galerkin equations (3.13), to obtain the coupled system: for each $\nu \in \mathcal{J}_p$

$$\sum_{\alpha \in \mathcal{J}_r} \sum_{\mu \in \mathcal{J}_p} [\mathbf{G}_\alpha]_{\nu,\mu} \mathbf{B}_\alpha \mathbf{u}_{h,\mu}^r = \langle \Psi_\nu, \mathbf{F} \rangle, \quad [\mathbf{G}_\alpha]_{\nu,\mu} = \langle \Psi_\nu \Psi_\mu \Psi_\alpha \rangle. \quad (3.20)$$

Alternatively, similar to (3.15), we may define $\mathbf{K}_r = \sum_{\alpha \in \mathcal{J}_r} \mathbf{G}_\alpha \otimes \mathbf{B}_\alpha$ to again obtain the coupled system of finite element problems: for all $\nu \in \mathcal{J}_p$

$$\sum_{\mu \in \mathcal{J}_p} [\mathbf{K}_r]_{\nu,\mu} \mathbf{u}_{h,\mu}^r = \mathbf{F} \delta_{0,\nu} \quad \forall \nu \in \mathcal{J}_p. \quad (3.21)$$

Note that in forming $\mathbf{K}_r$, we now need only store the matrices $(\mathbf{G}_\alpha)_{\alpha \in \mathcal{J}_r}$ and $(\mathbf{B}_\alpha)_{\alpha \in \mathcal{J}_r}$, so that the efficient, matrix-free, solution strategies discussed in Remark 3.3 can be applied.

Remark 3.4 (Projection and well-posedness). *In the case that the coefficient is a transcendental function of the random variables, as in Example 2.3, there does not exist an $r \in \mathbb{N}_0$ such that the projection (3.16) is exact. Due to the orthogonality of the basis, setting $r = 2p$ in the construction of $\mathbf{K}_r$ yields an entirely block-dense system, as first noted in [87], that is equivalent to (3.14), and computationally infeasible to solve. A more practical approach is to choose the expansion order $0 \leq r \leq 2p$, based on a priori estimates of the error in the solution introduced by the truncation $a^r(x, \mathbf{y})$, so that the error when using the truncated expansion does not exceed that of the SG approximation. In this approach however, it becomes important to consider whether the truncated projection violates the well-posedness of problem (2.2) by failing to satisfy assumption (A1). One way to ensure this does not happen is to choose $\tilde{r} \leq r \leq 2p$ such that*

$$\tilde{r} := \min\{r \in \mathbb{N}_0 : \|a - a^q\|_{L^\infty(\mathcal{U}; L^\infty(D))} \leq a_{\min}, \quad \forall q \in \mathbb{N}_0, \quad q \geq r\}. \quad (3.22)$$

An example of this problem can be seen in Figure 3.1 where for the function $a(x, y) = 0.1 + \exp(2.5y)$, uniform ellipticity of the truncated projection $a^r(x, y)$ does not hold on $\mathcal{U} = [-1, 1]$ for $r < 4$.

3.3 Monte Carlo method

In this section, we include a brief description of Monte Carlo (MC) approximations to quantities of interest involving the solution u to (2.1). The standard Monte Carlo method, when combined with finite element approximations described in Section 3.1, approximates QOI by empirical averages of i.i.d. realizations of the solution u_h , or functionals of u_h . We detail the basic MC averaging procedure for approximating $\mathbb{E}[u(\mathbf{y})]$ as follows:

Step 1. Generate m random samples $(\mathbf{y}_i)_{i=1}^m$ from the measure ϱ .

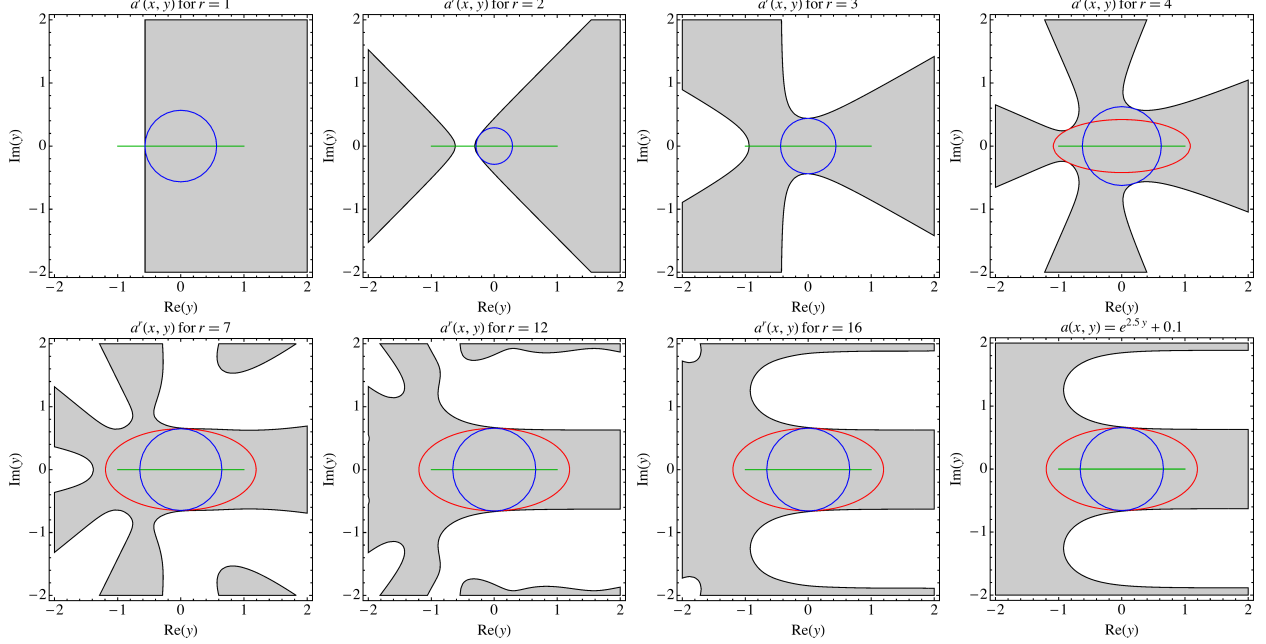


Figure 3.1: Domains of uniform ellipticity for orthonormal Legendre expansions of order r of the one-dimensional coefficient $a(x, y) = 0.1 + \exp(2.5y)$ are indicated by the gray regions in each plot. The last plot shows the domain of uniform ellipticity of the original function $a(x, y)$. The blue and red curves represent the maximal discs $\mathcal{D}_\gamma$ and ellipses $\mathcal{E}_\gamma$, respectively, that can be contained in those domains, and the green lines represent the interval $\mathcal{U} = [-1, 1]$.

Step 2. For each $i = 1, \dots, m$, compute the FE approximation (or alternative physical discretization) of the solution u at the point $\mathbf{y}_i$, i.e., $u_h(\mathbf{y}_i) \in \mathcal{V}_h$.

Step 3. Compute the empirical average $\mathbb{E}[u(\cdot, \mathbf{y})] \approx \frac{1}{m} \sum_{i=1}^m u_h(\cdot, \mathbf{y}_i)$.

This procedure can be also used to construct an approximation to the solution map $\mathbf{y} \mapsto u(\mathbf{y})$ by approximating the spectral coefficients $(\mathbf{c}_\nu)_{\nu \in \mathcal{J}} \in \mathcal{V}^N$ of (3.7) as

$$\mathbf{c}_\nu^\# = \frac{1}{m} \sum_{i=1}^m u_h(\cdot, \mathbf{y}_i) \Psi_\nu(\mathbf{y}_i) \varrho(\mathbf{y}_i) \approx \int_{\mathcal{U}} u(\cdot, \mathbf{y}) \Psi_\nu(\mathbf{y}) \varrho(\mathbf{y}) d\mathbf{y} =: \mathbf{c}_\nu, \quad (3.23)$$

Construction of the approximate orthonormal expansion to u via (3.23) is generally referred to as *non-intrusive spectral projection* via MC, see, e.g., [65, Section 4.4.3]. The strong law of large numbers [50, Theorem 2.4.1] ensures the convergence of $\mathbf{c}_\nu^\# \rightarrow \mathbf{c}_\nu$ almost surely as $m \rightarrow \infty$, and hence the MC approximation to the solution map converges as well. However, the convergence rate with respect to the number of samples is $\mathcal{O}(1/\sqrt{m})$ [6], which, while dimension independent, is too slow for most practical applications. For example, if the error of an approximation found using the MC method is $\mathcal{O}(10^{-1})$ using $m = 100$ samples, we require approximately $100 \times m - m = 9900$ additional samples to reduce the error to $\mathcal{O}(10^{-2})$.

One can also consider more structured sampling rules, i.e., not purely random, in Step 1. above. Such approaches include the so called *quasi-Monte Carlo* [90] and *latin hypercube*

sampling [74] methods, which have been shown to have convergence of $\mathcal{O}(\log(m)^{r(d)}/m)$ where the parameter $r(d) > 0$ grows as $d \rightarrow \infty$. In this work, we focus only on the Monte Carlo method as a baseline for performance in the high-dimensional numerical examples presented in Chapter 6, noting that the stochastic collocation methods presented in the next section often feature much faster rates of convergence [91].

3.4 Stochastic collocation method

Like the MC methods discussed in the previous section, stochastic collocation (SC) methods also apply a sampling based approach in approximating solutions to parameterized PDEs. However, rather than directly approximating a QOI through the empirical averaging, SC methods seek to reconstruct the coefficients of a global polynomial approximation via a method of tensor-product Lagrange interpolation [4]. The now well-known *sparse grid* approach to stochastic collocation dates back to the 1963 work by the Russian mathematician S.A. Smolyak [105]. Sparse grid approximations construct multi-dimensional multilevel bases of interpolating polynomials via specialized truncations, i.e., hierarchical differences, of tensor product expansions of one-dimensional bases.

Sparse grid SC approximations have been analyzed for multivariate tensor product approximation problems in [121]. Compared to full tensor-product grid collocation methods, which require $\mathcal{O}(m^d)$ samples to construct an approximation (where m is the number of points required by the underlying one-dimensional rule), sparse grid methods require only $\mathcal{O}(m \log(m)^{d-1})$ samples. For high-dimensional problems, this results in a dramatically reduced sample complexity on the underlying parameterized PDE for constructing fully discrete approximations. Furthermore, when the solution u of the parameterized PDE problem (2.1) has smooth dependence on the parameters, such methods can feature sub-exponential rates of convergence with respect to the overall polynomial order of the approximation [91]. Anisotropic approaches can also be applied to further reduce the sample complexity [92].

To construct the sparse grid SC approximation, we consider a class of multi-index sets defined in terms of increasing functions $\mathbf{m} : \mathbb{N}_+^d \rightarrow \mathbb{N}_+^d$ and $g : \mathbb{N}_+^d \rightarrow \mathbb{N}_+$. By $\mathbf{m}$ we specify the multivariate function $\mathbf{m}(\mathbf{l}) := (m_1(l_1), \dots, m_d(l_d))$ where each $m_j : \mathbb{N}_+ \rightarrow \mathbb{N}_+$ is an increasing function, possibly different for each $j = 1, \dots, d$. Here the m_j are referred to as growth functions, specifying how the number of points grows in the direction j . Associated with m_j we define the *left-inverse* $m_j^\dagger : \mathbb{N}_+ \rightarrow \mathbb{N}_+$ by $m_j^\dagger(q) = \min\{k \in \mathbb{N}_+ : m_j(k) \geq q\}$, and let $\mathbf{m}^\dagger(\mathbf{q}) = (m_1^\dagger(q_1), \dots, m_d^\dagger(q_d))$. In this case, we note that $m_j^\dagger(m_j(k)) = k$ and $m_j(m_j^\dagger(k)) \geq k$ for each $k \in \mathbb{N}_+$ and $j = 1, \dots, d$. Given $\mathbf{m}$ and g we can define the multi-index set

$$\mathcal{J}_L^{\mathbf{m},g} = \{\boldsymbol{\mu} \in \mathbb{N}_+^d : g(\mathbf{m}^\dagger(\boldsymbol{\mu} + \mathbf{1})) \leq L\}, \quad (3.24)$$

to be used in constructing polynomial approximations. In particular, setting $m_j(n) = n$ for all $n \in \mathbb{N}_+$ and $j = 1, \dots, d$, and defining

$$g_{\text{TP}}(\boldsymbol{\nu}) = \max_{1 \leq j \leq d} \nu_j, \quad g_{\text{TD}}(\boldsymbol{\nu}) = \sum_{j=1}^d (\nu_j - 1), \quad g_{\text{SM}}(\boldsymbol{\nu}) = \sum_{j=1}^d f(\nu_j), \quad (3.25)$$

where f is given in (3.9), and using the definition of $\mathcal{J}_L^{\boldsymbol{m},g}$ from (3.24), we obtain the TP, TD, and SM index sets $\mathcal{J}_L^{\text{TP}}$, $\mathcal{J}_L^{\text{TD}}$, and $\mathcal{J}_L^{\text{SM}}$, respectively, given in (3.9), with L generally referred to as the overall level of the grid.

We introduce a sequence of one-dimensional Lagrange interpolation operators $\mathcal{A}^{m_j(l_j)} : C^0(\mathcal{U}_j) \rightarrow \mathcal{P}_{m_j(l_j)-1}(\mathcal{U}_j)$. Then for $v \in C^0(\mathcal{U})$ the generalized multi-dimensional approximation operator $\mathcal{I}_L^{\boldsymbol{m},g} : C^0(\mathcal{U}) \rightarrow \mathcal{P}_{\mathcal{J}_L^{\boldsymbol{m},g}}(\mathcal{U})$ is given by

$$\mathcal{I}_L^{\boldsymbol{m},g}[v](\boldsymbol{y}) = \sum_{g(\mathbf{l}) \leq L} \sum_{\mathbf{i} \in \{0,1\}^d} (-1)^{|\mathbf{i}|} \left(\bigotimes_{j=1}^d \mathcal{A}_j^{m_j(l_j-i_j)} \right) [v](\boldsymbol{y}). \quad (3.26)$$

Construction of the approximation $\mathcal{I}_L^{\boldsymbol{m},g}[v](\boldsymbol{y})$ requires the independent evaluation of samples $v(\boldsymbol{y})$ on a deterministic set of distinct collocation points $\mathcal{G}_L^{\boldsymbol{m},g}$ having cardinality $M_L = \#\mathcal{G}_L^{\boldsymbol{m},g}$. Applying $\mathcal{I}_L^{\boldsymbol{m},g}[\cdot]$ from (3.26) to the semi-discrete solution $u_h(x, \boldsymbol{y})$ of problem (3.1), we obtain the fully discrete SC approximation

$$u_{h,L}(x, \boldsymbol{y}) = \mathcal{I}_L^{\boldsymbol{m},g}[u_h](x, \boldsymbol{y}). \quad (3.27)$$

One-dimensional abscissas In this work, we use three key examples for constructing the fully discrete approximation. The first is that of a fully-nested rule constructed on the Clenshaw-Curtis choice of abscissas [30] with function $g_{\text{TD}}(\boldsymbol{\nu})$ from (3.25) and an isotropic growth rule $\boldsymbol{m} = (m, \dots, m)$ with m given by

$$m(1) = 1, \quad m(l_j) = 2^{l_j-1} + 1 \quad \text{for } l_j > 1. \quad (3.28)$$

This is the classical Smolyak sparse-tensorization construction [105], and here the choice of $\boldsymbol{m}$ corresponds to a doubling growth rule that leads to a nested sequence of multi-dimensional grids, e.g., $\mathcal{G}_L^{\boldsymbol{m},g} \subset \mathcal{G}_{L+1}^{\boldsymbol{m},g}$. On the other hand, we can construct a sparse-Smolyak approximation on the Gauss-Legendre abscissas corresponding to the zeros of the Legendre polynomials $(\Psi_\nu)_{\nu \in \mathcal{F}}$, as defined in Section 3.2. However, when the points are grown isotropically according to the linear growth rule with $\boldsymbol{m} = (m, \dots, m)$ and m defined as

$$m(l_j) = l_j \quad \text{for } l_j \in \mathbb{N}, \quad (3.29)$$

and $g_{\text{TD}}(\boldsymbol{\nu})$ from (3.25), we obtain a grid that is not nested. The third construction we consider is based on the Leja points, i.e., the sequence of points defined inductively by $y_{n+1} := \arg\max_{y \in \mathcal{U}_j} \prod_{l=1}^n |y - y_l|$ (see, e.g., [41, 79, 80]). If we take the Leja sequence of

points with g_{TD} from (3.25) and the isotropic linear growth function $\mathbf{m}$ from (3.29), we obtain another sequence of nested grids.

3.5 Compressed sensing method

Compressed sensing (CS) is an appealing approach for approximating polynomial expansions from underdetermined systems of equations, with far fewer measurements than the number of terms in the expansion [20, 47, 102]. Under a sparsity (or compressibility) assumption on the coefficient vector, this approach enjoys a significant improvement in sample complexity in contrast to traditional methods such as discrete least squares, projection, and interpolation. As the solutions of many parameterized partial differential equations (PDEs) are known to be compressible (in the sense that they are well approximated by a sparse expansion in an orthonormal system, see, e.g., [32] and the references therein), it is no surprise that interest in applying compressed sensing techniques to approximating high-dimensional functions and parameterized systems has been growing rapidly in recent years [48, 70, 86, 89, 95, 96, 101, 125, 126].

In these works, the target function is a QOI of the solution of a parameterized PDE of the form of (2.1). As noted in Chapter 1, one can view the solution u to such problems as a map $\mathbf{y} \in \mathcal{U} \mapsto u(\mathbf{y}) \in \mathcal{V}$ where $\mathcal{V}$ is the solution space, typically a Sobolev space, e.g., $\mathcal{V} = H_0^1(D)$. The algorithms proposed in the previously cited works are designed to approximate a QOI, consisting of a function $g : \mathbf{y} \in \mathcal{U} \mapsto G(u(\mathbf{y}))$ which, e.g., is either the evaluation of u at a fixed point of the space/time domain or a linear functional in u such as a norm over D . With the measure $\varrho : \mathcal{U} \rightarrow \mathbb{R}_+$ and the $L^2_\varrho(\mathcal{U})$ -orthonormal basis $(\Psi_\nu)_{\nu \in \mathcal{F}}$ as in Section 2.1, and $\mathcal{F}$ as in Section 3.2.1, the considered functions are smooth, complex-valued, and can be expanded as

$$g(\mathbf{y}) = \sum_{\nu \in \mathcal{F}} c_\nu \Psi_\nu(\mathbf{y}), \quad (3.30)$$

where the coefficients c_ν belong to $\mathbb{C}$. The polynomial approximation of the function g in the CS setting is fairly straightforward. First, as in the context of Galerkin projection in (3.7), one truncates the expansion (3.30) in a multivariate polynomial space

$$P_{\mathcal{J}}(\mathcal{U}) := \text{span}\{\Psi_\nu(\mathbf{y}) : \nu \in \mathcal{J}\} \quad (3.31)$$

with $\mathcal{J}$ a finite set of indices whose cardinality $N := \#(\mathcal{J})$ is large enough to yield $g \simeq \sum_{\nu \in \mathcal{J}} c_\nu \Psi_\nu$. Then, as in the case of the MC method from 3.3, for some $m \in \mathbb{N}_+$ (generally satisfying $m \ll N$), we generate m samples $(\mathbf{y}_i)_{i=1}^m$ in the parametric (or stochastic) domain $\mathcal{U}$ independently from the orthogonalization measure ϱ associated with $(\Psi_\nu)_{\nu \in \mathcal{F}}$. Our goal is to find an approximation $g_{\mathcal{J}}^\#$ to g of the form

$$g_{\mathcal{J}}^\#(\mathbf{y}) = \sum_{\nu \in \mathcal{J}} c_\nu^\# \Psi_\nu(\mathbf{y}), \quad (3.32)$$

where $\mathbf{c}^\# := (c_\nu^\#)_{\nu \in \mathcal{J}}$ is the sparsest signal with an inherent interpolatory aspect. This problem can be written in terms of solving an underdetermined measurement system $\mathbf{A}\mathbf{z} = \mathbf{g}$

as follows. Let $\mathbf{A} \in \mathbb{R}^{m \times N}$ be a normalized measurement matrix, i.e., containing the samples of the orthonormal basis $(\Psi_\nu)_{\nu \in \mathcal{J}}$ at the points $(\mathbf{y}_i)_{i=1}^m$ and similarly let $\mathbf{g}$ be the normalized measurement vector containing the observations of the target function at the random points, i.e.,

$$\mathbf{A} := \left(\frac{\Psi_\nu(\mathbf{y}_i)}{\sqrt{m}} \right)_{\substack{1 \leq i \leq m \\ \nu \in \mathcal{J}}}, \quad \text{and} \quad \mathbf{g} := \left(\frac{g(\mathbf{y}_i)}{\sqrt{m}} \right)_{i=1}^m. \quad (3.33)$$

respectively. In practice, noisy formulations of this problem are also considered by investigating the expansion tail $\sum_{\nu \notin \mathcal{J}} c_\nu \Psi_\nu$.

To date, the sparse recovery of the polynomial expansion (3.30) via CS has shown to be very promising. However, this approach requires a low uniform bound of the underlying basis, given by

$$\Theta = \sup_{\nu \in \mathcal{J}} \|\Psi_\nu\|_{L^\infty(\mathcal{U})}, \quad (3.34)$$

as the sample complexity m required to recover the best s -term approximation (up to a multiplicative constant) scales with the following bound (see, e.g., [59])

$$m \gtrsim \Theta^2 s \times \log \text{ factors}. \quad (3.35)$$

This poses a challenge for many multivariate polynomial approximation strategies as Θ is prohibitively large in high dimensions. In particular, for d -dimensional problems, $\Theta = 2^{d/2}$ for Chebyshev systems and preconditioned Legendre systems [102]. Moreover, when using the standard Legendre expansion, the theoretical number of samples can exceed the cardinality of the polynomial subspace, unless the subspace a priori excludes all terms of high total order (see, e.g., [70, 125]). Therefore, the advantages of sparse polynomial recovery methods, coming from reduced query complexity, are eventually overcome by the *curse of dimensionality* discussed in Chapter 1: such techniques require at least as many samples as the traditional sparse interpolation techniques of Section 3.4 in high dimensions [65, 91, 92].

Nevertheless, in many engineering and science applications, the target functions, despite being high-dimensional, are smooth and often characterized by rapidly decaying polynomial expansions, whose most important coefficients are of *low order* [28, 32, 34, 75]. In such situations, the quest for finding the approximation containing the s largest terms can be restricted to polynomial spaces associated with lower (or anchored, or downward closed) multi-index sets.

Definition 3.5 (Lower set). *An index set $\mathcal{J} \subset \mathcal{F}$ is called a lower set (also called downward closed set) if and only if*

$$\boldsymbol{\nu} \in \mathcal{J} \quad \text{and} \quad \boldsymbol{\mu} \leq \boldsymbol{\nu} \implies \boldsymbol{\mu} \in \mathcal{J}, \quad (3.36)$$

where $\boldsymbol{\mu} \leq \boldsymbol{\nu}$ if and only if $\mu_j \leq \nu_j$ for all $1 \leq j \leq d$.

The practicality of downward closed sets is mainly computational, and has been demonstrated in different approaches such as quasi-optimal strategies [8], Taylor expansions [25], interpolatory approaches [27], and discrete least-squares [26]. For instance, in the

context of parametric PDEs such as (2.1), it was shown in [28] that, for a large class of operators $\mathcal{D}$ with a certain type of anisotropic dependence on $\mathbf{y}$, the solution map $\mathbf{y} \mapsto u(\mathbf{y})$ can be well-approximated by best s -term GPC expansions associated with index sets of cardinality s . Indeed, these best s -term approximations can be shown to satisfy algebraic rates of convergence $\mathcal{O}(s^{-\alpha})$, $\alpha > 0$ in the uniform and/or mean average sense. The same rates are preserved with index sets that are lower. In addition, for $\mathcal{U} = [-1, 1]^d$, such lower sets of cardinality s also enable the equivalence property $\|\cdot\|_{L^2_q(\mathcal{U})} \leq \|\cdot\|_{L^\infty} \leq s^\gamma \|\cdot\|_{L^2_q(\mathcal{U})}$ in arbitrary dimensions d with, e.g., $\gamma = 2$ for the uniform measure and $\gamma = \frac{\log 3}{\log 2}$ for Chebyshev measure.

Therefore, the focus of this section is on developing and analyzing CS approximations confined to downward closed sets, with the goal of overcoming the negative effects of the *curse of dimensionality* implied by the lower bound on the sampling complexity from (3.35). As such, our work also provides a fair comparison with existing numerical approaches for polynomial approximation in high dimensions [8, 25, 26, 27, 113]. To achieve our goal, we study a sparse recovery approach for imposing the downward closed structure, namely a weighted ℓ_1 -minimization with the specific choice of weights $\omega_\nu = \|\Psi_\nu\|_{L^\infty(\mathcal{U})}$. In addition, we describe a rigorous theoretical framework that provides analytical evidence for improved performance of our proposed methods in reconstructing smooth functions, noting detailed proofs of the results discussed can be found in our previous work [29].

In the context of CS, it is a well-established fact that sparse recovery is strongly tied to the concept of the restricted isometry property (RIP) of the normalized sampling matrix $\mathbf{A}$. However, motivated by the fact that the best s -term approximation is typically associated with a lower set, herein we describe a weaker version of the RIP, referred to as the *lower RIP*. Unlike the standard RIP which requires all sub-matrices formed by s columns of $\mathbf{A}$ to be well conditioned, the lower RIP involves only s -tuples of columns whose indices form a lower set. Given the lower RIP assumption, stable and robust reconstruction guarantees for the *best lower s -term approximation* of g , i.e., best among all approximations of g supported on lower sets of cardinality s , can be shown. It is reasonable to expect that this approximation, while weaker, is close to best s -term approximation for the smooth functions g considered throughout this work.

More importantly, the improved sample complexity for high-dimensional function recovery, using our methods, can be deduced directly from the sufficient condition for the lower RIP. For clarification, a complete technical description of (3.35) is given by the condition

$$m \geq C\Theta^2 s \log^3(s) \log(N), \quad (3.37)$$

which was developed in [23, 100, 104]. This result is often cited in the case of the standard RIP, and is used to guarantee uniform recovery with probability exceeding $1 - N^{-\log^3(s)}$. In this work, we describe three critical components that enable a systematic reduction in the number of samples compared to (3.37):

1. As mentioned above, the lower RIP is associated with downward closed sets, and allows us to employ efficient bounds of basis functions defined on those sets (derived in [26, 31] in the setting of discrete least-squares). Under the lower RIP assumption,

we can replace the term $\Theta^2 s$ from (3.37) with

$$K(s) = \sup_{\substack{\Lambda \subset \mathcal{J}, \Lambda^{\text{lower}} \\ \#(\Lambda) = s}} \left\| \sum_{\nu \in \Lambda} |\Psi_\nu|^2 \right\|_{L^\infty(\mathcal{U})}, \quad (3.38)$$

which, as we will soon see, is *significantly* smaller.

2. We can reasonably choose $\mathcal{J}$ as the Hyperbolic Cross index set from (3.9), repeated here for convenience

$$\mathcal{J}_s^{\text{HC}} := \left\{ \nu \in \mathcal{F} : \prod_{j=1}^d (\nu_j + 1) \leq s \right\}. \quad (3.39)$$

This choice represents the smallest set that surely contains the s best lower indices, since it can be seen as the union of all lower sets of cardinality s . An added benefit of this choice is that the cardinality of $\mathcal{J}_s^{\text{HC}}$ grows mildly in s and d when compared to other common choices, e.g., the tensor product and total degree sets from (3.9). Indeed, from [24, Theorem 3.7] and [82], we have $N_s := \#(\mathcal{J}_s^{\text{HC}}) \leq \min\{2s^3 4^d, e^2 s^{2+\log_2(d)}\}$. Since m and N are intimately connected in the context of CS, this choice also facilitates slow growth of m with respect to the dimension d and therefore minimizes complexity of matrix-vector multiplication.

3. By extending the chaining arguments, recently developed in [15, 73] for unitary matrices, to general bounded orthonormal systems, it is possible to decrease the logarithm factor in (3.37) by one unit. Following the approach in [15], the covering argument used there can be modified for this task, see Theorem 2.2 (restated below) and its proof in Appendix A.2 from [29]. The technical details presented there are necessary in quantifying the universal constants and the constraint of the number of samples m on the success probability. It is worth noting that this analysis shows a success probability slightly weaker than that associated with (3.37).

By combining all the above ingredients one can show that the sufficient condition required to reconstruct the best lower s -term approximation, with probability exceeding $1 - N^{-\log(s)}$, is given by

$$m \geq CK(s) \log^2(s) \min \left\{ \log(s) + d, \log(s) \log(d) \right\}. \quad (3.40)$$

As shown in [29, Lemma 3.5], for $\mathcal{J} = \mathcal{J}_s^{\text{HC}}$ in high dimensions, i.e., $2^d > s$,

$$\Theta^2 s \geq \begin{cases} s^2/2, & \text{if } (\Psi_\nu) \text{ is Chebyshev basis,} \\ s^{\frac{\log 3}{\log 2} + 1}/3, & \text{if } (\Psi_\nu) \text{ is Legendre basis,} \end{cases} \quad (3.41)$$

while [29, Lemma 3.7] yields

$$K(s) \leq \begin{cases} s^{\frac{\log 3}{\log 2}}, & \text{if } (\Psi_\nu) \text{ is Chebyshev basis,} \\ s^2, & \text{if } (\Psi_\nu) \text{ is Legendre basis.} \end{cases} \quad (3.42)$$

Therefore, the advantage in sample complexity given by (3.40) when compared to the well-known condition (3.37) implies that sufficient requirements for best lower s -term recovery possesses: an lower order dependence on s ; an lower order in the logarithmic factor $\log(s)$; and an efficient and explicit dependence on the dimension d through $\log(N)$, given by $\min\{\log(s) + d, \log(s) \log(d)\}$.

We proceed to outline the major results and strategies of [29] as follows. The next section details specific results on the improvement to the RIP estimate for bounded orthonormal systems. Section 3.5.2 describes the lower set recovery strategy, further improvement to the sample complexity under lower RIP condition, and related concepts. Section 3.5.3 concludes with a more thorough description of the functional recovery strategy, detailed in (3.32) and the subsequent discussion above, with particular attention to difficulties in reconstructing the solution u to (2.1) globally in the physical domain D .

3.5.1 Improved RIP estimate for bounded orthonormal systems

For over a decade, the RIP, defined below, has been an important ingredient for establishing sparse recovery guarantees in compressive-sensing contexts.

Definition 3.6 (RIP). *Given a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$, we say that $\mathbf{A}$ satisfies the restricted isometry property of order s if there exists $0 < \delta_s < 1$ such that*

$$(1 - \delta_s) \|\mathbf{z}\|_2^2 \leq \|\mathbf{A}\mathbf{z}\|_2^2 \leq (1 + \delta_s) \|\mathbf{z}\|_2^2, \quad (3.43)$$

for all $\mathbf{z} \in \mathbb{C}^N$ satisfying $\#(\text{supp}(\mathbf{z})) \leq s$. The smallest positive number δ_s for (3.43) to hold is referred to as the restricted isometry constant of $\mathbf{A}$.

The standard RIP concerns the action of the measurement matrix $\mathbf{A}$ from (3.33) on subspaces of $\mathbb{C}^N$ of dimension s . More specifically, (3.43) can be interpreted as an approximate isometry condition (giving rise to the name RIP), meaning that for $\mathbf{A}$ to satisfy the RIP, it must act on all s -dimensional subspaces of $\mathbb{C}^N$ as an isometry up to a (small) constant δ_s . The following RIP estimate for bounded orthonormal systems and its proof, inspired by the approach in [15], can be found in [29, Theorem 2.2 & Appendix A.2].

Theorem 3.7. *Let δ, γ be fixed parameters with $0 < \delta < 1/13$, $0 < \gamma < 1$ and $(\Psi_\nu)_{\nu \in \mathcal{J}}$ be an orthonormal system of finite size $N = \#(\mathcal{J})$. Assume that*

$$m \geq 2^6 e \frac{\Theta^2 s}{\delta^2} \log\left(\frac{\Theta^2 s}{\delta^2}\right) \max\left\{ \frac{2^5}{\delta^4} \log\left(40 \frac{\Theta^2 s}{\delta^2} \log\left(\frac{\Theta^2 s}{\delta^2}\right)\right) \log(4N), \right. \\ \left. \frac{1}{\delta} \log\left(\frac{1}{\gamma \delta} \log\left(\frac{\Theta^2 s}{\delta^2}\right)\right) \right\}, \quad (3.44)$$

and $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m$ are drawn independently from the orthogonalization measure ϱ associated to $\{\Psi_\nu\}$. Then with probability exceeding $1 - \gamma$, the normalized sampling matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ satisfies

$$(1 - 13\delta)\|\mathbf{z}\|_2^2 < \|\mathbf{A}\mathbf{z}\|_2^2 < (1 + 13\delta)\|\mathbf{z}\|_2^2, \quad (3.45)$$

for all $\mathbf{z} \in \mathbb{C}^N$, $\#(\text{supp}(\mathbf{z})) \leq s$.

Remark 3.8. We note that the result of Theorem 3.7 is not limited to the Legendre and Chebyshev systems on the unit hypercube, which are the main consideration in this work. Such estimates hold for any bounded orthonormal system on any underlying domain, discrete or continuous.

The complete detailed proof of Theorem 3.7 is given in [29, Appendix A.2]. However, to assist the reader in better understanding the logic of that proof we now provide a sketch explaining the essential features of the argument achieving the improved RIP estimate.

Sketch of proof. To begin, let us denote

$$\begin{aligned} \psi(\mathbf{y}, \mathbf{z}) &:= \sum_{\nu \in \mathcal{J}} z_\nu \Psi_\nu(\mathbf{y}), \quad \forall \mathbf{y} \in \mathcal{U}, \mathbf{z} \in \mathbb{C}^N, \\ \text{and } \mathcal{E}_s &:= \{\mathbf{z} \in \mathbb{C}^N : \|\mathbf{z}\|_2 = 1, \#(\text{supp}(\mathbf{z})) \leq s\}. \end{aligned}$$

Our goal is to derive conditions on m such that, for a set of m random samples $\{\mathbf{y}_i\}_{i=1}^m \subset \mathcal{U}$ drawn according to ϱ , with high probability there holds $\forall \mathbf{z} \in \mathcal{E}_s$:

$$\frac{1}{m} \sum_{i=1}^m |\psi(\mathbf{y}_i, \mathbf{z})|^2 \approx \int_{\mathcal{U}} |\psi(\mathbf{y}, \mathbf{z})|^2 d\varrho. \quad (3.46)$$

Note that the LHS of the above expression equals $\|\mathbf{A}\mathbf{z}\|_2^2$, and the RHS is $\|\mathbf{z}\|_2^2 = 1$. Hence (3.46) is nothing but (3.45) rewritten with the new notation. We construct a finite set $\mathcal{S}$ of piecewise constant functions on $\mathcal{U}$, which provides a good approximation of $\psi(\cdot, \mathbf{z})$, denoted by $\tilde{\psi}(\cdot, \mathbf{z})$, for all $\mathbf{z} \in \mathcal{E}_s$, and reduce the task of showing (3.46) holds for infinitely many $\mathbf{z} \in \mathcal{E}_s$ to that of proving

$$\frac{1}{m} \sum_{i=1}^m |\tilde{\psi}(\mathbf{y}_i, \mathbf{z})|^2 \approx \int_{\mathcal{U}} |\tilde{\psi}(\mathbf{y}, \mathbf{z})|^2 d\varrho \quad (3.47)$$

for all $\tilde{\psi}(\cdot, \mathbf{z}) \in \mathcal{S}$. In particular, the family of “discrete” approximations $\tilde{\psi}(\cdot, \mathbf{z})$ is constructed in such a way that (see [29, Appendix A.2], Step 1):

1. for any $\mathbf{z} \in \mathcal{E}_s$, $\tilde{\psi}(\cdot, \mathbf{z}) \approx \psi(\cdot, \mathbf{z})$;
2. $\tilde{\psi}(\cdot, \mathbf{z})$ can be represented as a piecewise constant function on $\mathcal{U}$: $\tilde{\psi}(\cdot, \mathbf{z}) = \sum_{l \in \mathcal{L}} \tilde{\psi}_l^{\mathbf{z}}(\cdot)$, where each $\tilde{\psi}_l^{\mathbf{z}}$ is a constant function, pair-wise disjointly supported on a subset of $\mathcal{U}$, representing a scale of $\tilde{\psi}(\cdot, \mathbf{z})$ and $\mathcal{L}$ is a finite set of scales; and
3. for each $l \in \mathcal{L}$, the class $\{\tilde{\psi}_l^{\mathbf{z}} : \mathbf{z} \in \mathcal{E}_s\}$ is finite, whose cardinality is as small as possible.

With the use of 1. and 2. one can establish the bound (see [29, Appendix A.2], Step 2):

$$\begin{aligned} \left| \frac{1}{m} \sum_{i=1}^m |\psi(\mathbf{y}_i, \mathbf{z})|^2 - \int_{\mathcal{U}} |\psi(\mathbf{y}, \mathbf{z})|^2 d\varrho \right| &\lesssim \left| \frac{1}{m} \sum_{i=1}^m |\tilde{\psi}(\mathbf{y}_i, \mathbf{z})|^2 - \int_{\mathcal{U}} |\tilde{\psi}(\mathbf{y}, \mathbf{z})|^2 d\varrho \right| \\ &\leq \sum_{l \in \mathcal{L}} \left| \frac{1}{m} \sum_{i=1}^m |\tilde{\psi}_l^{\mathbf{z}}(\mathbf{y}_i)|^2 - \int_{\mathcal{U}} |\tilde{\psi}_l^{\mathbf{z}}(\mathbf{y})|^2 d\varrho \right|. \end{aligned} \quad (3.48)$$

Using the basic tail estimate given by [29, Lemma A.2] yields, for any l and $\mathbf{z}$, with high probability

$$\frac{1}{m} \sum_{i=1}^m \tilde{\psi}_l^{\mathbf{z}}(\mathbf{y}_i) \approx \int_{\mathcal{U}} \tilde{\psi}_l^{\mathbf{z}}(\mathbf{y}) d\varrho. \quad (3.49)$$

We can then obtain (3.46) by employing (3.48) and applying the union bound for (3.49) over all $l, \mathbf{z}$ (see [29, Appendix A.2], Step 3). For this argument to yield small m , it is critical to construct $\tilde{\psi}(\cdot, \mathbf{z})$ so that the total number of functions $\tilde{\psi}_l^{\mathbf{z}}$ (over $l \in \mathcal{L}, \mathbf{z} \in \mathcal{E}_s$) is finite and minimal, justifying the requirement given by 3. above.

As an example, for each $l \in \mathcal{L}$, we can define $\tilde{\psi}_l^{\mathbf{z}}$ according to a covering of $\mathcal{E}_s$ under the pseudo-metric

$$d(\mathbf{z}, \mathbf{z}') = \sup_{\mathbf{y} \in \mathcal{U}} |\psi(\mathbf{y}, \mathbf{z} - \mathbf{z}')|,$$

so that $\#\{\tilde{\psi}_l^{\mathbf{z}} : \mathbf{z} \in \mathcal{E}_s\}$ is roughly a covering number of $\mathcal{E}_s$. However, one can check that in our high-dimensional setting, this covering number grows exponentially in the dimension of $\mathcal{U}$. Fortunately, an inspection of $|\psi(\cdot, \mathbf{z} - \mathbf{z}')|$ reveals that these functions often have tall spikes in a small subregion of $\mathcal{U}$, while are relatively small for the rest of the domain. This motivates us to consider a new “distance” between $\mathbf{z}$ and $\mathbf{z}'$, which is significantly smaller than $d(\mathbf{z}, \mathbf{z}')$, given by an upper bound of $|\psi(\mathbf{y}, \mathbf{z} - \mathbf{z}')|$ for *most* $\mathbf{y} \in \mathcal{U}$. More rigorously, we define

$$d_{\varsigma}(\mathbf{z}, \mathbf{z}') := \inf_{\substack{\tilde{\mathcal{U}} \subset \mathcal{U} \\ \varrho(\tilde{\mathcal{U}}) = 1 - \varsigma}} \sup_{\mathbf{y} \in \tilde{\mathcal{U}}} |\psi(\mathbf{y}, \mathbf{z} - \mathbf{z}')|.$$

Although d_{ς} is not a proper pseudo-metric, an adaptation of the covering number result can still be derived in this case (see [29, Lemma A.3]). This argument is similar in spirit to [73, Lemma 3.5]. The approximation $\tilde{\psi}$, constructed with d_{ς} , may not agree with ψ in a small subdomain of $\mathcal{U}$, but one can tune ς so as to not affect the estimate (3.48). This completes the sketch of the proof of Theorem 3.7. ■

Remark 3.9. *In brief, the RIP (and subsequently, best s -term reconstruction) occurs with probability exceeding $1 - \gamma$ under the condition*

$$m \geq C\Theta^2 s \max\{\log^2(\Theta^2 s) \log(N), \log(\Theta^2 s) \log(\log(\Theta^2 s)/\gamma)\}. \quad (3.50)$$

The first constraint in (3.50) therefore reduces the order of $\log(s)$ in (3.37) by one unit. The second constraint, on the other hand, has an additional log factor compared to the well-known one, i.e., $m \geq C\Theta^2 s \log(1/\gamma)$, see [100], after balancing leading to a weaker success probability, as discussed in the beginning of Section 3.5.

3.5.2 Sparse recovery on lower sets

In this section we focus on a smooth function g , given by (3.30), and exploit the fast decay of its polynomial expansion to further improve (3.50). Central to this task is the concept of lower or downward closed sets, given Definition 3.5. With this in mind, instead of best s -term approximations, we are interested in the *best lower s -term approximation* of g , which is the best among all approximations of g supported on lower sets of cardinality s . More rigorously, let Λ^* be a lower subset of $\mathcal{F}$ which realizes the infimum

$$\Lambda^* := \arg \min_{\substack{\Lambda \text{ lower} \\ \#(\Lambda) \leq s}} \|g - g_\Lambda\|, \quad (3.51)$$

with the norm to be specified later. Here $\|g - g_{\Lambda^*}\|$ is the *best lower s -term error*, and our goal is to find approximations of g with error scaling linearly in $\|g - g_{\Lambda^*}\|$. We expect the best lower s -term error, while generally larger, is close to the best s -term error in our setting. In particular, these quantities are identical provided that g is s -sparse and $\text{supp}(g)$ is lower, so that g is well-represented by finite Legendre and Chebyshev expansions.

To achieve our goal, it is reasonable to consider a relaxed version of the RIP that concerns s -tuples of columns associated with lower sets. Given a multi-index set $\Lambda \subset \mathcal{F}$, we introduce the quantity

$$K(\Lambda) := \left\| \sum_{\nu \in \Lambda} |\Psi_\nu|^2 \right\|_{L^\infty}, \quad (3.52)$$

and, with an abuse of notation, denote

$$K(s) := \sup_{\substack{\Lambda \subset \mathcal{F}, \Lambda \text{ lower} \\ \#(\Lambda) = s}} K(\Lambda), \quad (3.53)$$

which was previously defined in (3.38). We define next the *lower restricted isometry property* (lower RIP). This property is exclusive to the present setting and defined here for submatrices whose columns are associated with indices $\nu \in \mathcal{F}$.

Definition 3.10 (Lower RIP). *For $\mathbf{A} \in \mathbb{C}^{m \times N}$ as in (3.33), we say that $\mathbf{A}$ satisfies the lower restricted isometry property of order s if there exists $0 < \delta_{\ell,s} < 1$ such that*

$$(1 - \delta_{\ell,s}) \|\mathbf{z}\|_2^2 \leq \|\mathbf{A}\mathbf{z}\|_2^2 \leq (1 + \delta_{\ell,s}) \|\mathbf{z}\|_2^2 \quad (3.54)$$

for all $\mathbf{z} \in \mathbb{C}^N$ satisfying $K(\text{supp}(\mathbf{z})) \leq K(s)$. The smallest positive number $\delta_{\ell,s}$ for (3.54) to hold is referred to as the lower restricted isometry constant of $\mathbf{A}$.

Remark 3.11. The lower RIP is a specific case of the weighted RIP, introduced in [103] for general weights, here with the weights $\omega_{\nu} = \|\Psi_{\nu}\|_{L^{\infty}}$. By introducing the notation $\|\mathbf{z}\|_{0,\omega} = \sum_{\nu \in \text{supp}(\mathbf{z})} \omega_{\nu}^2$ for $\mathbf{z} \in \mathbb{C}^N$, the weighted RIP constant $\delta_{\omega,s}$ was defined as the smallest number $\delta_{\omega,s}$ for which

$$(1 - \delta_{\omega,s})\|\mathbf{z}\|_2^2 \leq \|\mathbf{A}\mathbf{z}\|_2^2 \leq (1 + \delta_{\omega,s})\|\mathbf{z}\|_2^2, \quad \|\mathbf{z}\|_{0,\omega} \leq s. \quad (3.55)$$

By (3.52), observe that $K(\text{supp}(\mathbf{z})) \leq \|\mathbf{z}\|_{0,\omega}$, hence given $\mathbf{z}$ such that $\|\mathbf{z}\|_{0,\omega} \leq K(s)$, then $K(\text{supp}(\mathbf{z})) \leq K(s)$ so that (3.54) is satisfied showing that $\delta_{\omega,K(s)} \leq \delta_{\ell,s}$. For the Chebyshev and Legendre systems, the polynomials Ψ_{ν} all attain their supremums at the point $\mathbf{1} = (1, \dots, 1) \in \mathcal{U}$, hence for any $\mathbf{z} \in \mathbb{C}^N$, then $K(\text{supp}(\mathbf{z})) = \|\mathbf{z}\|_{0,\omega}$, showing that

$$\delta_{\ell,s} = \delta_{\omega,K(s)}. \quad (3.56)$$

Note the change of order in this relation: loosely speaking, the lower RIP of order s corresponds to the weighted RIP of order $K(s)$.

An important subclass of $\mathbf{z}$ satisfying (3.54) is $\mathbf{z} \in \mathbb{C}^N$ with $\#(\text{supp}(\mathbf{z})) \leq s$ and $\text{supp}(\mathbf{z})$ lower. One may want to consider a more natural isometry property which requires (3.54) for only vectors $\mathbf{z}$ in the above class. We can see from the following analysis that this property is weaker but requires the same sampling cost as (3.54). The sample complexity for lower RIP is established in the following theorem.

Theorem 3.12. Let δ, γ be fixed parameters with $0 < \delta < 1/13$, $0 < \gamma < 1$ and $\{\Psi_{\nu}\}_{\nu \in \mathcal{J}}$ be an orthonormal system of finite size $N = \#(\mathcal{J})$. Assume that

$$m \geq 2^6 e \frac{K(s)}{\delta^2} \log\left(\frac{K(s)}{\delta^2}\right) \max\left\{\frac{2^5}{\delta^4} \log\left(40 \frac{K(s)}{\delta^2} \log\left(\frac{K(s)}{\delta^2}\right)\right) \log(4N), \frac{1}{\delta} \log\left(\frac{1}{\gamma\delta} \log\left(\frac{K(s)}{\delta^2}\right)\right)\right\}, \quad (3.57)$$

and $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m$ are drawn independently from the orthogonalization measure ϱ associated to $\{\Psi_{\nu}\}$. Then with probability exceeding $1 - \gamma$, the normalized sampling matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ satisfies

$$(1 - 13\delta)\|\mathbf{z}\|_2^2 < \|\mathbf{A}\mathbf{z}\|_2^2 < (1 + 13\delta)\|\mathbf{z}\|_2^2, \quad (3.58)$$

for all $\mathbf{z} \in \mathbb{C}^N$, $K(\text{supp}(\mathbf{z})) \leq K(s)$.

The proof of Theorem 3.12 is discussed in [29, Appendix A.3]. This proof essentially follows the same path as the proof of Theorem 3.7 with few minor changes.

Remark 3.13. In brief, the random matrix $\mathbf{A}$ satisfies the lower RIP of order s and, subsequently, guarantees lower reconstruction with probability exceeding $1 - \gamma$ if the sample size m satisfies

$$m \geq CK(s) \max\{\log^2(K(s)) \log(N), \log(K(s)) \log(\log(K(s))/\gamma)\}. \quad (3.59)$$

3.5.3 Application of standard compressed sensing to approximate functionals of solutions to parameterized PDEs

In this section, we detail the application of standard compressive-sensing polynomial approximation techniques to the recovery of functionals of the solution to parameterized PDE problems. As discussed in Section 3.5, previous works have applied compressed sensing to approximate functionals of solutions u to problems such as (2.1) under the assumption that such functionals have a sparse representation in a given basis, e.g., the Legendre or Chebyshev systems. In those works, the functional $G : \mathcal{V} \rightarrow \mathbb{R}$ is typically taken to be either the pointwise evaluation functional $G_{x^*}(u(\mathbf{y})) = u(x^*, \mathbf{y})$ for a fixed $x^* \in D$, or a QOI involving u such as $G(u(\mathbf{y})) = \|u(\mathbf{y})\|_{\mathcal{X}}$ for $\mathcal{X} = L^2(D)$ or $\mathcal{V}(D)$. One can realize a sparse vector $\mathbf{c}^\# = (c_\nu^\#)_{\nu \in \mathcal{J}} \in \mathbb{C}^N$ with corresponding functional $g_{\mathcal{J}}^\#$ as in (3.32), approximating $\mathbf{c} = (c_\nu)_{\nu \in \mathcal{J}} \in \mathbb{C}^N$ and $g_{\mathcal{J}}$ (the projection of g onto $P_{\mathcal{J}}(\mathcal{U})$ from (3.31)), respectively, by solving the standard *basis pursuit denoising* (BPDN) problem:

$$\text{minimize}_{\mathbf{z} \in \mathbb{C}^N} \|\mathbf{z}\|_1 \quad \text{subject to} \quad \|\mathbf{A}\mathbf{z} - \mathbf{g}\|_2 \leq \tilde{\eta}/\sqrt{m}. \quad (3.60)$$

Here $\mathbf{A}$ and $\mathbf{g}$ are the normalized measurement matrix and vector defined in (3.33), $\|\cdot\|_1$ and $\|\cdot\|_2$ are the standard ℓ_1 and Euclidean norms on $\mathbb{C}^N$ and $\mathbb{C}^m$, respectively, and $\tilde{\eta}$ is an estimate of the tail, i.e., $\|g_{\mathcal{J}^c}(\mathbf{y})\|_{L^2_{\mathbf{y}}(\mathcal{U})} = \|\sum_{\nu \notin \mathcal{J}} c_\nu \Psi_\nu(\mathbf{y})\|_{L^2_{\mathbf{y}}(\mathcal{U})} \leq \tilde{\eta}$.

One can apply the functional recovery approach via compressed sensing to reconstruct the solution globally over D , as was described in [126, Section 4]. We give an example of this construction in the context of finite element approximations with a triangulation $\mathcal{T}_h$ of D and associated piecewise linear finite element basis $(\varphi_k)_{k=1}^{K_h}$, as described in Section 3.1, as follows. We consider solving the BPDN problem (3.60) for an approximation $g_k^\#$ to

$$g_k(\mathbf{y}) := G_{x_k}(u(\mathbf{y})) = u(x_k, \mathbf{y}) \equiv \sum_{\nu \in \mathcal{F}} c_\nu(x_k) \Psi_\nu(\mathbf{y}), \quad (3.61)$$

i.e., the parametric GPC expansion of the point-wise evaluation functional associated to a point $x_k \in D$ of the triangulation $\mathcal{T}_h$. Under the assumption that (3.61) has sparse representation at all points $x \in D$ in a finite basis $(\Psi_\nu)_{\nu \in \mathcal{J}}$, this process can be repeated at all points on the mesh, yielding a collection of sparse vectors $(\tilde{\mathbf{c}}^{(k),\#})_{k=1}^{K_h}$ and corresponding truncated pointwise GPC expansions

$$g_{k,\mathcal{J}}^\#(\mathbf{y}) = \sum_{\nu \in \mathcal{J}} \tilde{c}_{\nu,k}^{h,\#} \Psi_\nu(\mathbf{y}). \quad (3.62)$$

From these point-wise approximations to the GPC expansion we can construct a global approximation to the solution u via piecewise linear interpolation. More specifically, applying the basis $(\varphi_k)_{k=1}^{K_h}$ of the finite element space $\mathcal{V}_h$ associated with $\mathcal{T}_h$, we can interpolate the approximate point-wise GPC expansions, yielding a fully discrete global polynomial

approximation which we denote

$$u_{h,\mathcal{J}}^{\text{pwCS},\#}(x, \mathbf{y}) := \sum_{\nu \in \mathcal{J}} \sum_{k=1}^{K_h} \tilde{c}_{\nu,k}^{h,\#} \varphi_k(x) \Psi_{\nu}(\mathbf{y}). \quad (3.63)$$

Using this fully discrete approximation, we can compute any desired QOI or perform an optimization with respect to $\mathbf{y}$, as discussed in the parametric and stochastic modeling contexts of Chapter 1.

Remark 3.14 (Pointwise estimates of truncation error & disjoint supports of the point-wise GPC expansion vectors). *We remark that successful application of the above global approximation technique via CS requires an estimate of the point-wise truncation error of (3.61) corresponding to the choice of index set $\mathcal{J}$ in parameterizing (3.60). For many parameterized PDE problems fitting (2.1), such estimates are inaccessible without highly-detailed knowledge of the physical and parametric dependence of the solution u . Furthermore this process will ultimately result in a matrix $\tilde{c}^{h,\#} \in \mathbb{C}^{N \times K_h}$, with each column associated with a particular point $x_k \in D$. In many scenarios, the coefficient vectors of the pointwise GPC expansions have disjoint supports due to the fact that the best s terms in reconstruction have dependence on x which varies over the domain D .*

3.6 A novel method for sparse Hilbert-valued vector recovery

The main contribution of this work, discussed in this section, is an extension of standard compressed sensing and joint-sparse recovery theory to the recovery of sparse vectors $\mathbf{c}$ whose coefficients $(\mathbf{c}_{\nu})_{\nu \in \mathcal{J}}$ belong to a general Hilbert space $\mathcal{V}$, i.e., in the context of (2.1) are functions of $x \in D$. Through this extension, we are able to derive problems, and algorithms for their solution, which enable sparse *Hilbert-valued vector* (HV) recovery. In certain cases, the problem of joint-sparse recovery, can be considered an instance of our HV recovery problem. For the sake of generality, we conduct our analysis in the infinite-dimensional setting, i.e., avoid imposing a specific discretization of the space $\mathcal{V}$, noting that the application of such results to the finite-dimensional setting is straightforward.

We organize our results as follows. Section 3.6.1 presents the problem setting and measurement scheme required for sparse HV reconstruction. The key extension from standard compressive sensing is a reformulation of problem (3.60) in terms of a mixed $\ell_{\mathcal{V},p}$ -norm which enables energy norm regularized sparse reconstruction of solutions to parameterized PDEs. Section 3.6.2 discusses facts relevant to the setting of tensor products of Hilbert spaces, which are useful in the analysis of the HV recovery version of Bregman iterative regularization introduced in Section 4.2. During the course of the development of our sparse HV recovery technique, we discovered connections to the problem of joint-sparse recovery. These connections are described in Section 3.6.3, where issues related to the context of finite element approximations are also presented. Section 3.6.4 concludes the chapter by discussing uniform recovery results and error estimates for solving the energy

norm regularized sparse reconstruction problems, relevant to the context of the parameterized elliptic PDE (2.2).

3.6.1 Sparse recovery of Hilbert-valued vectors

In this section, we discuss the necessary norms and methods of regularization which enable sparse HV recovery. We describe our approach as follows. For any $N \in \mathbb{N}_0$, $p \geq 1$, and $\mathbf{z}, \mathbf{z}' \in \mathcal{V}^N$, define the norm and inner product

$$\|\mathbf{z}\|_{\mathcal{V},p} := \|\mathbf{z}\|_{\ell_p(\mathcal{V}^N)} = \left(\sum_{\nu \in \mathcal{J}} \|\mathbf{z}_\nu\|_{\mathcal{V}}^p \right)^{1/p}, \quad \langle \mathbf{z}, \mathbf{z}' \rangle_{\mathcal{V},2} := \sum_{\nu \in \mathcal{J}} \langle \mathbf{z}_\nu, \mathbf{z}'_\nu \rangle_{\mathcal{V}}. \quad (3.64)$$

In the case of $\mathcal{V} = H_0^1(D)$, of course $\langle \mathbf{z}_\nu, \mathbf{z}'_\nu \rangle_{\mathcal{V}} := \int_D \nabla \mathbf{z}_\nu \cdot \nabla \mathbf{z}'_\nu \, dx$ and $\|\mathbf{z}_\nu\|_{\mathcal{V}} = \sqrt{\langle \mathbf{z}_\nu, \mathbf{z}_\nu \rangle_{\mathcal{V}}}$. Let $\text{supp}(\mathbf{z}) := \{\nu \in \mathcal{J} : \mathbf{z}_\nu \neq \mathbf{0}\}$ denote the support of $\mathbf{z} \in \mathcal{V}^N$ and $\|\mathbf{z}\|_{\mathcal{V},0} := \#(\text{supp}(\mathbf{z}))$. Given $p \geq 0$, we define the $\|\cdot\|_{\mathcal{V},p}$ -error of the best s -term approximation to $\mathbf{z} \in \mathcal{V}^N$ as

$$\sigma_s(\mathbf{z})_{\mathcal{V},p} = \inf_{\hat{\mathbf{z}} \in \mathcal{V}^N, \|\hat{\mathbf{z}}\|_{\mathcal{V},0} \leq s} \|\mathbf{z} - \hat{\mathbf{z}}\|_{\mathcal{V},p}. \quad (3.65)$$

We say that $\mathbf{z} \in \mathcal{V}^N$ is s -sparse when $\|\mathbf{z}\|_{\mathcal{V},0} \leq s$, and compressible when $\sigma_s(\mathbf{z})_{\mathcal{V},p} \rightarrow 0$ quickly in s . The measurement scheme for HV recovery borrows directly from compressive-sensing polynomial approximation schemes, which we repeat in the Hilbert-valued setting as follows. As discussed in Sections 3.2 & 3.5, we require an index set $\mathcal{J}$ with $N = \#(\mathcal{J})$ chosen large enough so that the tail $u_{\mathcal{J}^c}$ is negligible, relative to the total error of the approximation. For some $m \ll N$, we generate m samples $\mathbf{y}_1, \dots, \mathbf{y}_m$ in the parametric domain $\mathcal{U}$ independently from the measure ϱ . The goal is then to find an approximation $u_{\mathcal{J}}^\#$ of $u_{\mathcal{J}}$ from (3.7) of the form

$$u_{\mathcal{J}}^\#(x, \mathbf{y}) = \sum_{\nu \in \mathcal{J}} \mathbf{c}_\nu^\#(x) \Psi_\nu(\mathbf{y}), \quad (3.66)$$

satisfying

$$\|u_{\mathcal{J}} - u_{\mathcal{J}}^\#\|_{L^2_{\varrho}(\mathcal{U}, \mathcal{V})} = \left\| \sum_{\nu \in \mathcal{J}} (\mathbf{c}_\nu - \mathbf{c}_\nu^\#) \Psi_\nu \right\|_{L^2_{\varrho}(\mathcal{U}, \mathcal{V})} \equiv \|\mathbf{c} - \mathbf{c}^\#\|_{\mathcal{V},2} \leq \varepsilon^\# \quad (3.67)$$

for some prescribed tolerance $\varepsilon^\# > 0$ below the desired total error of approximation. Just as in (3.33), we let $\mathbf{A} \in \mathbb{C}^{m \times N}$ be the normalized sampling matrix containing the samples of the basis $(\Psi_\nu)_{\nu \in \mathcal{J}}$, $\mathbf{u}$ the normalized observations of the parameterized solution at the points $\mathbf{y}_i$, and $\mathbf{e}$ the normalized expansion tail

$$\mathbf{A} := \left(\frac{\Psi_\nu(\mathbf{y}_i)}{\sqrt{m}} \right)_{\substack{1 \leq i \leq m \\ \nu \in \mathcal{J}}}, \quad \mathbf{u} := \left(\frac{u(\mathbf{y}_i)}{\sqrt{m}} \right)_{1 \leq i \leq m}, \quad \mathbf{e} := \left(\frac{u_{\mathcal{J}^c}(\mathbf{y}_i)}{\sqrt{m}} \right)_{1 \leq i \leq m}, \quad (3.68)$$

noting that $\mathbf{u}$ and $\mathbf{e}$ are elements of the space $\mathcal{V}^m$. Then the vector $\mathbf{c} = (\mathbf{c}_\nu)_{\nu \in \mathcal{J}} \in \mathcal{V}^N$ of exact coefficients satisfies $\mathbf{A}\mathbf{c} + \mathbf{e} = \mathbf{u}$. Assuming an upper estimate of the tail $\|\mathbf{e}\|_{\mathcal{V},2} \leq \eta/\sqrt{m}$ is known, one has $\|\mathbf{A}\mathbf{c} - \mathbf{u}\|_{\mathcal{V},2} \leq \eta/\sqrt{m}$, which motivates us to find the approximation $\mathbf{c}^\# = (\mathbf{c}_\nu^\#)_{\nu \in \mathcal{J}}$ among all $\mathbf{z} \in \mathcal{V}^N$ such that $\|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2} \leq \eta/\sqrt{m}$. Having the ℓ_1 minimization approach for real and complex sparse recovery in mind, it is natural to consider the following HV version of the BPDN problem:

$$\text{minimize}_{\mathbf{z} \in \mathcal{V}^N} \|\mathbf{z}\|_{\mathcal{V},1} \quad \text{subject to} \quad \|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2} \leq \eta/\sqrt{m}. \quad (3.69)$$

Here, the Hilbert-valued vector is minimized with respect to an ℓ_1 -norm defined as the sum of the magnitude of coordinates in the energy norm. We will sometimes refer to problem (3.69) as the HV-BPDN problem. In the next section, we will see how this problem can be seen as a generalization of the now well-known joint-sparse recovery problem, and examine situations in which both problems are actually equivalent. Solutions to this problem can be found by solving the related *unconstrained* convex minimization problem:

$$\text{minimize}_{\mathbf{z} \in \mathcal{V}^N} \|\mathbf{z}\|_{\mathcal{V},1} + \frac{\mu}{2} \|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2}^2, \quad (3.70)$$

We leave a further discussion of methods of solving the HV-BPDN problem via problem (3.70), including strategies for parameterizing μ to achieve the tail bound constraint on the residual from (3.69), to Chapter 4.

3.6.2 Setting of tensor products of Hilbert spaces

In this section we present a few useful properties of the spaces $(\mathcal{V}^N, \|\cdot\|_{\mathcal{V},p})$ for $1 \leq p \leq \infty$. As in the previous section, we assume that $\mathcal{J}$ is a finite multi-index set of cardinality $N = \#(\mathcal{J})$. We introduce the shorthand $\mathcal{H}_p$ to denote the space $(\mathcal{V}^N, \|\cdot\|_{\mathcal{V},p}) \equiv \{\mathbf{z} \in \mathcal{V}^N : \|\mathbf{z}\|_{\mathcal{V},p} < \infty\}$, and, in the case $p = 2$, drop the dependence on p in writing $\mathcal{H}$ when referring to the Hilbert space $(\mathcal{V}^N, \langle \cdot, \cdot \rangle_{\mathcal{V},2})$. We note that the space $\mathcal{H}$ is often referred to as the direct sum of N Hilbert spaces $\mathcal{V}$. Otherwise, for $1 \leq p \leq \infty$, $\mathcal{H}_p$ is a Banach space. Our first result concerns a Hölder-type inequality for the $\langle \cdot, \cdot \rangle_{\mathcal{V},2}$ inner-product.

Lemma 3.15. *Assume that $\mathbf{u} \in \mathcal{H}_p$ and $\mathbf{v} \in \mathcal{H}_q$ with $1 \leq p \leq \infty$ and $1/p + 1/q = 1$. Then*

$$|\langle \mathbf{u}, \mathbf{v} \rangle_{\mathcal{V},2}| \leq \|\mathbf{u}\|_{\mathcal{V},p} \|\mathbf{v}\|_{\mathcal{V},q}. \quad (3.71)$$

Proof. First, let $1 < p < \infty$. Then

$$|\langle \mathbf{u}, \mathbf{v} \rangle_{\mathcal{V},2}| \leq \sum_{\nu \in \mathcal{J}} \|\mathbf{u}_\nu\|_{\mathcal{V}} \|\mathbf{v}_\nu\|_{\mathcal{V}} \leq \left(\sum_{\nu \in \mathcal{J}} \|\mathbf{u}_\nu\|_{\mathcal{V}}^p \right)^{1/p} \left(\sum_{\mu \in \mathcal{J}} \|\mathbf{v}_\mu\|_{\mathcal{V}}^q \right)^{1/q} = \|\mathbf{u}\|_{\mathcal{V},p} \|\mathbf{v}\|_{\mathcal{V},q},$$

where the first inequality follows from the Cauchy-Schwarz inequality in each component, and the second inequality follows from Hölder's inequality for the counting measure since

each $\|\mathbf{u}_\nu\|_\nu$ and $\|\mathbf{v}_\nu\|_\nu$ are finite. Now if $p = 1$, then

$$|\langle \mathbf{u}, \mathbf{v} \rangle_{\nu,2}| \leq \sum_{\nu \in \mathcal{J}} \|\mathbf{u}_\nu\|_\nu \|\mathbf{v}_\nu\|_\nu \leq \max_{\mu \in \mathcal{J}} \|\mathbf{v}_\mu\|_\nu \sum_{\nu \in \mathcal{J}} \|\mathbf{u}_\nu\|_\nu = \|\mathbf{u}\|_{\nu,1} \|\mathbf{v}\|_{\nu,\infty}.$$

The case $p = \infty$ follows after interchanging p and q . ■

With Lemma 3.15, we can show the following result about the dual of $\mathcal{H}_p$, which follows from standard results about $(\mathbb{R}^N, \|\cdot\|_p)$, see [49, Theorem IV.3.9], and properties of the inner product $\langle \cdot, \cdot \rangle_\nu$.

Theorem 3.16. *If $1 \leq p \leq \infty$ and $1/p + 1/q = 1$, then the mapping $\mathbf{v} \longleftrightarrow (\mathbf{v}_\nu)_{\nu \in \mathcal{J}}$ determined by the equation*

$$\mathbf{v}(\mathbf{u}) := \langle \mathbf{v}, \mathbf{u} \rangle_{\nu,2} = \sum_{\nu \in \mathcal{J}} \langle \mathbf{v}_\nu, \mathbf{u}_\nu \rangle_\nu, \quad \mathbf{u} = (\mathbf{u}_\nu)_{\nu \in \mathcal{J}} \in \mathcal{H}_p,$$

is an isometric isomorphism of $\mathcal{H}_p^*$ onto $\mathcal{H}_q$.

Proof. It is clear that the mapping is an isomorphism. To see that it is also isometric, we apply Lemma 3.15 to obtain

$$\|\mathbf{v}\| := \sup_{\substack{\mathbf{u} \in \mathcal{H}_p \\ \|\mathbf{u}\|_{\nu,p} \leq 1}} |\langle \mathbf{v}, \mathbf{u} \rangle_{\nu,2}| \leq \sup_{\substack{\mathbf{u} \in \mathcal{H}_p \\ \|\mathbf{u}\|_{\nu,p} \leq 1}} \|\mathbf{v}\|_{\nu,q} \|\mathbf{u}\|_{\nu,p} = \|\mathbf{v}\|_{\nu,q},$$

where $\|\cdot\|$ is the operator norm. Now let $\mathbf{u}$ be defined by

$$\mathbf{u}_\nu = \begin{cases} \frac{\|\mathbf{v}_\nu\|_\nu^q}{\|\mathbf{v}_\nu\|_\nu^2} \cdot \mathbf{v}_\nu & \mathbf{v}_\nu \neq \mathbf{0}, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Then $\mathbf{u} \in \mathcal{H}_p$, and

$$\langle \mathbf{v}, \mathbf{u} \rangle_{\nu,2} = \sum_{\nu \in \mathcal{J}} \langle \mathbf{v}_\nu, \mathbf{u}_\nu \rangle_\nu = \sum_{\nu \in \mathcal{J}} \frac{\|\mathbf{v}_\nu\|_\nu^q}{\|\mathbf{v}_\nu\|_\nu^2} \langle \mathbf{v}_\nu, \mathbf{v}_\nu \rangle_\nu = \sum_{\nu \in \mathcal{J}} \|\mathbf{v}_\nu\|_\nu^q$$

and

$$\|\mathbf{u}\|_{\nu,p}^p = \sum_{\nu \in \mathcal{J}} \|\mathbf{u}_\nu\|_\nu^p = \sum_{\nu \in \mathcal{J}} \left(\frac{\|\mathbf{v}_\nu\|_\nu^q}{\|\mathbf{v}_\nu\|_\nu^2} \right)^p \|\mathbf{v}_\nu\|_\nu^p = \sum_{\nu \in \mathcal{J}} \|\mathbf{v}_\nu\|_\nu^{(q-1)p} = \sum_{\nu \in \mathcal{J}} \|\mathbf{v}_\nu\|_\nu^q$$

or $\|\mathbf{u}\|_{\nu,p} = (\sum_{\nu \in \mathcal{J}} \|\mathbf{v}_\nu\|_\nu^q)^{1/p}$, which, together with the inequality

$$|\langle \mathbf{v}, \mathbf{u} \rangle_{\nu,2}| \leq \|\mathbf{v}\| \|\mathbf{u}\|_{\nu,p},$$

and $1/q = 1 - 1/p$ gives

$$\frac{|\langle \mathbf{v}, \mathbf{u} \rangle_{\mathcal{V}, 2}|}{\|\mathbf{u}\|_{\mathcal{V}, p}} = \frac{\sum_{\nu \in \mathcal{J}} \|\mathbf{v}_\nu\|_{\mathcal{V}}^q}{(\sum_{\nu \in \mathcal{J}} \|\mathbf{v}_\nu\|_{\mathcal{V}}^q)^{1/p}} = \left(\sum_{\nu \in \mathcal{J}} \|\mathbf{v}_\nu\|_{\mathcal{V}}^q \right)^{1/q} \leq \|\mathbf{v}\|.$$

Thus $\|\mathbf{v}\| = \|\mathbf{v}\|_{\mathcal{V}, q}$, which shows that the mapping is an isometry. ■

The above result implies the reflexivity of $\mathcal{H}_p$ for any $1 \leq p \leq \infty$. While this result is not particularly surprising since $\mathcal{J}$ is a finite set and the underlying space $\mathcal{V}$ is a Hilbert space (hence reflexive), it will be useful in establishing the existence of weak-* convergent subsequences of the iterations resulting from Bregman iterative regularization discussed in Section 4.2. We note that in the case that $\mathcal{J}$ is not a finite set, classical analysis of infinite sequence spaces can be used to show that the above result does not hold [49].

3.6.3 Relation between sparse Hilbert-valued and joint-sparse recovery problems

Joint-sparse recovery problem

Following the significant advances in establishing sufficient conditions for recovery of sparse signals from underdetermined systems of equations [20, 47], many works in the last decade have sought to expand such problems to recovering not just one signal, but a collection of sparse signals, see, e.g., [37, 39, 45, 53, 58, 64, 88, 106, 114, 115]. In these works, the collection is assumed to possess a shared structure, e.g. a common (unknown) support set among all signal vectors, or a shared block-sparsity pattern. In this setting, the goal is to simultaneously reconstruct the entire set from noisy, possibly corrupted measurements, exploiting this structure in designing methods of regularizing the problems and efficient algorithms for their solution.

As described in our previous work [45], the model of structured sparsity most closely connected to the setting of sparse polynomial approximations of parameterized PDE problems is that of joint-sparsity. We describe the joint-sparse recovery problem as follows. Consider the problem of the simultaneously reconstructing a set of sparse signal vectors $\{\mathbf{c}^{(r)} \in \mathbb{C}^N : r \in \Omega\}$, sharing common support, given multiple measurements

$$\mathbf{u}^{(r)} = \mathbf{A}\mathbf{c}^{(r)} + \mathbf{e}^{(r)}, \quad \forall r \in \Omega. \quad (3.72)$$

Here Ω is a countable (possibly infinite) set, $\mathbf{u}^{(r)} \in \mathbb{C}^m$ are normalized measurement vectors, $\mathbf{A} \in \mathbb{C}^{m \times N}$ is a predefined normalized sampling matrix, and $\mathbf{e}^{(r)} \in \mathbb{C}^m$ are unknown noise vectors. We may group the signal, measurement, and noise vectors into the matrices

$$\begin{aligned} \mathbf{c} &= [\mathbf{c}^{(1)} \ \mathbf{c}^{(2)} \ \dots \ \mathbf{c}^{(r)} \ \dots] \in \mathbb{C}^{N \times \Omega}, \\ \mathbf{u} &= [\mathbf{u}^{(1)} \ \mathbf{u}^{(2)} \ \dots \ \mathbf{u}^{(r)} \ \dots] \in \mathbb{C}^{m \times \Omega}, \quad \mathbf{e} = [\mathbf{e}^{(1)} \ \mathbf{e}^{(2)} \ \dots \ \mathbf{e}^{(r)} \ \dots] \in \mathbb{C}^{m \times \Omega} \end{aligned}$$

so that our problem becomes to reconstruct the signal matrix $\mathbf{c}$ from

$$\mathbf{u} = \mathbf{A}\mathbf{c} + \mathbf{e}. \quad (3.73)$$

Here the common support assumption implies all but a few rows of $\mathbf{c}$ are negligible.

Previous works have studied problem (3.73) in the context of underdetermined systems of equations, i.e., $\mathbf{A} \in \mathbb{C}^{m \times N}$ where $m \ll N$. Both theoretical and numerical results have shown that measurement strategies for single-sparse recovery problems can be applied with great success to the joint-sparse reconstruction problem [64], with some works suggesting that joint-sparse recovery enjoys a reduced sample complexity over single-sparse recovery [39, 53]. One avenue for solving the joint-sparse recovery problem is through a simple reformulation of the BPDN problem (3.60), see, e.g., [88, 114, 115]:

$$\text{minimize}_{\mathbf{z} \in \mathbb{C}^{N \times \Omega}} \|\mathbf{z}\|_{2,1} \quad \text{subject to} \quad \|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{2,2} \leq \eta/\sqrt{m}, \quad (3.74)$$

with η chosen proportional to $\|\mathbf{e}\|_{2,2}$. Here, similar to (3.64), for any $N \in \mathbb{N}_0$, $1 \leq p, q < \infty$, and $\mathbf{z}, \mathbf{z}' \in \mathbb{C}^{N \times \Omega}$, the matrix norm and inner product are given by

$$\|\mathbf{z}\|_{q,p} := \left(\sum_{j \in [N]} \|\mathbf{x}_j\|_q^p \right)^{1/p}, \quad \langle \mathbf{z}, \mathbf{z}' \rangle_{2,2} := \sum_{j \in [N]} \langle \mathbf{z}_j, \mathbf{z}'_j \rangle_2, \quad (3.75)$$

where $\mathbf{x}_i$ denotes the i -th row of $\mathbf{x}$, and $[N] = \{1, \dots, N\}$. Of course, if $\|\mathbf{e}\|_{2,2} = 0$ we can replace the inequality constraint in (3.74) with the equality constraint $\mathbf{A}\mathbf{z} = \mathbf{u}$ and this becomes the joint-sparse *basis pursuit* (BP) problem. Just as in single sparse approximation, the $\ell_{2,1}$ part of the objective of (3.74) is a convex relaxation of $\ell_{2,0}$ -minimization, promoting sparsity of the rows, while the row-wise ℓ_2 -norm calculation promotes nonsparsity of the columns and the $\ell_{2,2}$ part promotes data fidelity in the model. Extensions of error estimates for the single-sparse BPDN problem, when the matrix $\mathbf{A}$ satisfies an RIP condition, see Definition 3.43, to problem (3.74) are trivial, see, e.g., [45, Proposition 2.2].

Remark 3.17 (Relaxation of the constrained minimization problem (3.74)). *We note that the constrained joint-sparse BPDN problem (3.74) can be solved by solving the related unconstrained convex minimization problem discussed in Section 1.3:*

$$\text{minimize}_{\mathbf{x} \in \mathbb{C}^{N \times \Omega}} \|\mathbf{x}\|_{2,1} + \frac{\mu}{2} \|\mathbf{A}\mathbf{x} - \mathbf{u}\|_{2,2}^2, \quad (3.76)$$

This problem provides solutions to problem (3.74) when the penalty parameter μ is chosen proportionally to η . In general, an estimate of η is required in order to obtain accurate solutions to (3.74) by solving (3.76). Many existing algorithms for single-sparse recovery can be modified in a straightforward manner to obtain solutions to (3.76), e.g., forward-backward splitting [45, 67, 68], Bregman iterations [93], and the spectral projected gradient method SPGL1 [118, 119]. Our previous work [45] examined a formulation of the forward-backward splitting algorithm for the solution of (3.76). In Section 4.1 we will discuss a novel strong convergence result for this algorithm in the context of solving (3.76), noting that the analysis that holds in this context can also be applied in the Hilbert-valued recovery setting.

Equivalence of the joint-sparse and Hilbert-valued recovery problems

From the discussion in Sections 3.2.1 & 3.6.1, after projecting the expansion (3.5) onto the $\mathcal{V}$ -valued polynomial space indexed by $\mathcal{J}$ from (3.6), one can identify the approximate solution to a parameterized PDE problem with a vector of Hilbert-valued coefficients $(\mathbf{c}_\nu)_{\nu \in \mathcal{J}} \in \mathcal{V}^N$. Assuming $\mathcal{V}$ is separable, there exists a countable orthonormal basis $(\phi_r)_{r \in \Omega}$ of $\mathcal{V}$. In this case we can write, for each $\nu \in \mathcal{J}$,

$$\mathbf{c}_\nu = \sum_{r \in \Omega} \hat{c}_{\nu,r} \phi_r, \quad (3.77)$$

where $\hat{c}_{\nu,r} \in \mathbb{C}$, $\forall r \in \Omega$. Hence, we have a natural bijection between elements $\mathbf{c}_\nu \in \mathcal{V}$ and vectors $\hat{\mathbf{c}}_\nu := (\hat{c}_{\nu,1}, \hat{c}_{\nu,2}, \dots, \hat{c}_{\nu,r}, \dots) \in \mathbb{C}^{1 \times \Omega}$, for each $\nu \in \mathcal{J}$. Grouping the vectors $(\hat{\mathbf{c}}_\nu)_{\nu \in \mathcal{J}}$ into the rows of a matrix, we see that the solution (3.7) can be identified with $\hat{\mathbf{c}} \in \mathbb{C}^{N \times \Omega}$. Furthermore, due to Parseval's identity, we have

$$\|\mathbf{c}\|_{\mathcal{V},q}^q := \sum_{\nu \in \mathcal{J}} \|\mathbf{c}_\nu\|_{\mathcal{V}}^q \equiv \sum_{\nu \in \mathcal{J}} \left(\sum_{r \in \Omega} |\hat{c}_{\nu,r}|^2 \right)^{q/2} \equiv \sum_{\nu \in \mathcal{J}} \|\hat{\mathbf{c}}_\nu\|_2^q =: \|\hat{\mathbf{c}}\|_{2,q}^q, \quad (3.78)$$

suggesting that problems (3.69) and (3.74) are equivalent in this case.

Hilbert-valued recovery in the context of finite element approximations

In practice, one may choose to employ another basis in approximating the solution u over the physical domain, e.g., due to computational constraints on the problem. Recalling the discussion on finite element approximations from Section 3.1, we can differentiate problems (3.69) and (3.74) in the absence of an orthonormal basis of $\mathcal{V}$ as follows. Given a basis $(\varphi_k)_{k=1}^{K_h}$ for the finite dimensional space $\mathcal{V}_h \subset \mathcal{V}(D)$ with $D \subset \mathbb{R}^n$ and $K_h = \dim(\mathcal{V}_h)$, for each $\nu \in \mathcal{J}$, we can approximate $\mathbf{c}_\nu$ by its FE expansion

$$\mathbf{c}_\nu^h := \sum_{k=1}^{K_h} \hat{c}_{\nu,k}^h \varphi_k \in \mathcal{V}_h, \quad (3.79)$$

where again $\hat{c}_{\nu,k}^h \in \mathbb{C}$ for $k = 1, \dots, K_h$. As in the countable orthonormal basis case, we can again identify the coefficients of the expansion (3.2) as rows of a matrix $\hat{\mathbf{c}}^h \in \mathbb{C}^{N \times K_h}$. Due to the regularization provided by the ℓ_1 -part of the $\ell_{\mathcal{V},1}$ -norm in problem (3.69), we expect that the approximate solution matrix $\hat{\mathbf{c}}^{h,\#}$ obtained by solving the HV-BPDN problem will be row-sparse, or approximately row-sparse, as in the case of joint-sparse recovery with mixed norm $\ell_{2,1}$. However, unlike in that case, we do not have $\|\hat{\mathbf{c}}^h\|_{2,q} \equiv \|\mathbf{c}^h\|_{\mathcal{V},q}$, but instead

$$c_{h,n} \|\hat{\mathbf{c}}^h\|_{2,q} \leq \|\mathbf{c}^h\|_{\mathcal{V},q} \leq C_{h,n} \|\hat{\mathbf{c}}^h\|_{2,q}, \quad (3.80)$$

for some constants $0 < c_{h,n} \leq C_{h,n}$ which may depend on h and n . Using problem (2.2) as an example, if $D \subset \mathbb{R}^n$ is a Lipschitz domain in $n = 2$ physical dimensions, and $(\varphi_k)_{k=1}^{K_h}$ is a piecewise-linear finite element basis, associated with a quasi-uniform and shape-regular mesh $\mathcal{T}_h$ of D , then $c_{h,n}$ is $\mathcal{O}(h^2)$ and $C_{h,n}$ is $\mathcal{O}(1)$, for more details, see, e.g., [17]. Nonetheless,

given the representation (3.79), it is easy to compute $\|\mathbf{c}_\nu^h\|_\nu$ in finding solutions to (3.69). More details on the changes to algorithms required for solving (3.69) with FE discretizations will be described in Sections 4.2-4.3 and 6.3.1.

3.6.4 Uniform recovery of Hilbert-valued signals

Straightforward extensions of concepts and results from compressed sensing and joint-sparse recovery can be made to ensure uniform recovery of Hilbert-valued signals via $\ell_{\nu,1}$ -relaxation. Well-known concepts such as the null space property (NSP) and restricted isometry property (RIP) have Hilbert-valued counterparts. We conclude this chapter by presenting Hilbert-valued versions of the NSP and RIP to guarantee uniform recovery for the Hilbert-valued BPDN problem, i.e., in the presence of noise or sparsity defects. To our knowledge, these extensions have not yet been provided in the considered setting, but are necessary for establishing recovery guarantees.

We first consider the noiseless case, i.e., when $\eta = 0$ in (3.69). Let $[N]$ denote the set $\{1, \dots, N\}$ and $\mathbf{z}_S$ be the Hilbert-valued vector consisting of only the elements from $\mathbf{z}$ indexed by the set $S \subseteq [N]$, i.e., $\mathbf{z}_j = \mathbf{0}$ for $j \in S^c$. Following simple extensions of results from [83, 120], it can be shown that every s -sparse $\mathbf{c} \in \mathcal{V}^N$ can be uniquely recovered from $\mathbf{u} = \mathbf{A}\mathbf{c}$ by solving (3.69) if and only if

$$\|\mathbf{z}_S\|_{\nu,1} < \|\mathbf{z}_{S^c}\|_{\nu,1},$$

for all $\mathbf{z} \neq \mathbf{0}$ with $\mathbf{z} \in \ker \mathbf{A} \setminus \{\mathbf{0}\}$, and all $S \subset [N]$ with $\#(S) = s$. In the noisy case $\eta > 0$, best s -term approximations of vectors $\mathbf{c} \in \mathcal{V}^N$ can be guaranteed (up to a constant and noise level) by the $\ell_{\nu,2}$ -robust NSP.

Definition 3.18 ($\ell_{\nu,2}$ -robust null space property). *The matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ is said to satisfy the $\ell_{\nu,2}$ -robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$ if*

$$\|\mathbf{z}_S\|_{\nu,2} \leq \frac{\rho}{\sqrt{s}} \|\mathbf{z}_{S^c}\|_{\nu,1} + \tau \|\mathbf{A}\mathbf{z}\|_{\nu,2} \quad \forall \mathbf{z} \in \mathcal{V}^N, \forall S \subset [N] \text{ with } \#(S) \leq s. \quad (3.81)$$

Just as in the setting of joint-sparse recovery, replacing the vector norms by $\ell_{\nu,q}$ -norms, we can obtain the following error estimate for solving the Hilbert-valued BPDN problem.

Proposition 3.19. *Let $1 \leq q \leq 2$. Suppose that the matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ satisfies the $\ell_{\nu,2}$ -robust NSP of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $\mathbf{c} \in \mathcal{V}^N$, the solution $\mathbf{c}^\#$ to (3.69) with $\mathbf{u} = \mathbf{A}\mathbf{c} + \mathbf{e}$ and $\|\mathbf{e}\|_{\nu,2} \leq \eta/\sqrt{m}$ approximates $\mathbf{c}$ with errors*

$$\|\mathbf{c} - \mathbf{c}^\#\|_{\nu,q} \leq \frac{C_1}{s^{1-1/q}} \sigma_s(\mathbf{c})_{\nu,1} + C_2 s^{1/q-1/2} \eta / \sqrt{m}, \quad (3.82)$$

for some constants $C_1, C_2 > 0$ depending only on ρ and τ . In particular, for $q = 1$ and $q = 2$, (3.82) gives

$$\|\mathbf{c} - \mathbf{c}^\#\|_{\mathcal{V},1} \leq C_1 \sigma_s(\mathbf{c})_{\mathcal{V},1} + C_2 \eta \sqrt{s/m}, \quad (3.83)$$

$$\|\mathbf{c} - \mathbf{c}^\#\|_{\mathcal{V},2} \leq \frac{C_1}{\sqrt{s}} \sigma_s(\mathbf{c})_{\mathcal{V},1} + C_2 \eta / \sqrt{m}. \quad (3.84)$$

Similar extensions to the standard RIP property are required to quantify the sample complexity of solving (3.69) to a prescribed accuracy. Our previous work [45] established the equivalence of the standard RIP with the $\ell_{2,2}$ -RIP relevant for joint-sparse recovery. The same can be shown for an $\ell_{\mathcal{V},2}$ -version of the RIP, given in the proof of the following proposition.

Proposition 3.20. *Suppose that the matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ satisfies the RIP, that is*

$$(1 - \delta_{2s}) \|\mathbf{z}\|_2^2 \leq \|\mathbf{A}\mathbf{z}\|_2^2 \leq (1 + \delta_{2s}) \|\mathbf{z}\|_2^2, \quad \forall \mathbf{z} \in \mathbb{C}^N, \quad \mathbf{z} \text{ } 2s\text{-sparse}, \quad (3.85)$$

with $\delta_{2s} < \frac{4}{\sqrt{41}}$. Then, $\mathbf{A}$ satisfies the $\ell_{\mathcal{V},2}$ -robust NSP of order s with constants $0 < \rho < 1$ and $\tau > 0$ depending only on δ_{2s} .

Proof. We show that (3.85) is equivalent to the following Hilbert-valued $\ell_{\mathcal{V},2}$ -version of the RIP, i.e.,

$$(1 - \delta_{2s}) \|\mathbf{z}\|_{\mathcal{V},2}^2 \leq \|\mathbf{A}\mathbf{z}\|_{\mathcal{V},2}^2 \leq (1 + \delta_{2s}) \|\mathbf{z}\|_{\mathcal{V},2}^2, \quad \forall \mathbf{z} \in \mathcal{V}^N, \quad \mathbf{z} \text{ } 2s\text{-sparse}. \quad (3.86)$$

The implication of the $\ell_{\mathcal{V},2}$ -robust NSP from Definition 3.18 easily follows by arguments from vector recovery proofs, see, e.g., [59, Theorem 6.13].

Assume that (3.85) holds, and let $\mathbf{z} \in \mathcal{V}^N$ be a $2s$ -sparse Hilbert-valued vector. Then there exists a $2s$ -row sparse matrix $\hat{\mathbf{z}} \in \mathbb{C}^{N \times \Omega}$ such that

$$\mathbf{z}_\nu = \sum_{r \in \Omega} \hat{\mathbf{z}}_{\nu,r} \phi_r, \quad \forall \nu \in \mathcal{J},$$

where $(\phi_r)_{r \in \Omega}$ is an orthonormal basis for $\mathcal{V}$ and $\|\mathbf{z}\|_{\mathcal{V},2} \equiv \|\hat{\mathbf{z}}\|_{2,2}$, as in (3.77) and (3.78), respectively. Moreover

$$\begin{aligned} \|\mathbf{A}\mathbf{z}\|_{\mathcal{V},2}^2 &:= \sum_{j=1}^m \|(\mathbf{A}\mathbf{z})_j\|_{\mathcal{V}}^2 = \sum_{j=1}^m \sum_{\nu \in \mathcal{J}} \sum_{\mu \in \mathcal{J}} \mathbf{A}_{j,\nu} \mathbf{A}_{j,\mu} \left\langle \sum_{r \in \Omega} \hat{\mathbf{z}}_{\nu,r} \phi_r, \sum_{t \in \Omega} \hat{\mathbf{z}}_{\mu,t} \phi_t \right\rangle_{\mathcal{V}} \\ &= \sum_{j=1}^m \sum_{\nu \in \mathcal{J}} \sum_{\mu \in \mathcal{J}} \mathbf{A}_{j,\nu} \mathbf{A}_{j,\mu} \langle \hat{\mathbf{z}}_\nu, \hat{\mathbf{z}}_\mu \rangle_2 \\ &= \sum_{j=1}^m \langle (\mathbf{A}\hat{\mathbf{z}})_j, (\mathbf{A}\hat{\mathbf{z}})_j \rangle_2 =: \|\mathbf{A}\hat{\mathbf{z}}\|_{2,2}^2, \end{aligned}$$

which, with [45, Proposition 2.3], implies (3.86).

On the other hand, assuming (3.86), let $\hat{\mathbf{z}}$ be a $2s$ -sparse vector in $\mathbb{C}^N$. Then the Hilbert valued vector $\mathbf{z} \in \mathcal{V}^N$ defined by $\mathbf{z}_j := \hat{\mathbf{z}}_j \phi_r \quad \forall j \in [N]$, for $r \in \Omega$ and ϕ_r arbitrary from the

basis $(\phi_r)_{r \in \Omega}$, satisfies

$$\|\mathbf{z}\|_{\mathcal{V},2}^2 := \sum_{j=1}^N \langle \mathbf{z}_j, \mathbf{z}_j \rangle_{\mathcal{V}} = \sum_{j=1}^N \langle \hat{z}_j \phi_r, \hat{z}_j \phi_r \rangle_{\mathcal{V}} = \sum_{j=1}^N |\hat{z}_j|^2 =: \|\hat{\mathbf{z}}\|_2^2$$

and, similarly

$$\begin{aligned} \|\mathbf{A}\mathbf{z}\|_{\mathcal{V},2}^2 &:= \sum_{j=1}^m \|(\mathbf{A}\mathbf{z})_j\|_{\mathcal{V}}^2 = \sum_{j=1}^m \left\langle \sum_{\nu \in \mathcal{J}} \mathbf{A}_{j,\nu} \hat{z}_{\nu} \phi_r, \sum_{\mu \in \mathcal{J}} \mathbf{A}_{j,\mu} \hat{z}_{\mu} \phi_r \right\rangle_{\mathcal{V}} \\ &= \sum_{j=1}^m \left(\sum_{\nu \in \mathcal{J}} \mathbf{A}_{j,\nu} \hat{z}_{\nu} \right) \left(\sum_{\mu \in \mathcal{J}} \mathbf{A}_{j,\mu} \hat{z}_{\mu} \right) \langle \phi_r, \phi_r \rangle_{\mathcal{V}} \\ &= \sum_{j=1}^m |(\mathbf{A}\hat{\mathbf{z}})_j|^2 =: \|\mathbf{A}\hat{\mathbf{z}}\|_2^2. \end{aligned}$$

Since $\mathbf{z}$ is $2s$ -sparse, the result follows. ■

Proposition 3.20 implies that query-complexity results shown for solving the single-sparse BPDN problem (3.60) hold for the Hilbert-valued BPDN problem (3.69) as well. Summarizing the results of Theorem 3.7 and the discussion in Remark 3.9 from Section 3.5, given $m \asymp \Theta^2 s \log^2(s) \log(N)$, Hilbert-valued reconstruction through problem (3.69) achieves an error of $\|\mathbf{c} - \mathbf{c}^{\#}\|_{\mathcal{V},2} \lesssim \sigma_s(\mathbf{c})_{\mathcal{V},1} / \sqrt{s} + \eta / \sqrt{m}$ with probability $1 - N^{-\log(s)}$. When exact estimates of the tail are available, the dependence on the truncation error can be removed, see, e.g., [29, Theorem 4.8]. Moreover, the lower bound (3.50) can be further reduced to as in (3.59) under the lower sparsity assumption and weighted recovery strategy from Section 3.5.2. We leave a more detailed study of efficiency of weighted sparse recovery in the Hilbert-valued case to future work.

Chapter 4

Algorithms for sparse recovery

In this chapter, we describe algorithms for the solution of several ℓ_1 -regularized constrained and unconstrained convex optimization problems presented in Chapter 3. Section 4.1 describes a formulation of the forward-backward splitting algorithm, presented in the setting of joint sparse recovery via problem (3.76). A novel strong convergence result is shown in the infinite-dimensional setting using only properties of the composed gradient-descent and soft-thresholding operators, which comprise the algorithm. We note that the analysis to follow requires only facts involving the $\|\cdot\|_{2,p}$ -norm and associated $\langle\cdot,\cdot\rangle_{2,2}$ inner product from (3.75) and problem (3.76). The required properties also hold for the $\|\cdot\|_{\mathcal{V},p}$ -norm and $\langle\cdot,\cdot\rangle_{\mathcal{V},2}$ inner product from (3.64) and problem (3.70), implying our novel strong convergence result for infinite-dimensional joint-sparse recovery via forward-backward splitting extends naturally to the infinite-dimensional Hilbert-valued setting.

Beginning with Section 4.2, the remainder of the chapter will concentrate on modifying algorithms for single-sparse and joint-sparse recovery to recover sparse Hilbert-valued vectors in the context of Section 3.6. Section 4.2 details the combination of the forward-backward iterations with Bregman iterations. While a strong convergence result for Bregman iterations remains to be shown in the infinite-dimensional HV and joint-sparse recovery settings of Sections 3.6.1 & 3.6.3, respectively, other results that have been shown for *total variation* (TV)-based image restoration can be extended to these settings. However, we note that these extensions are trivial, and follow from the analysis presented in the works [93, 129]. Therefore, the results extending such analysis to the Hilbert-valued setting are included only for completeness, and their proofs are reserved for Appendix A. Section 4.3 describes the *fixed-point continuation* (FPC) strategy on the step-size parameter of the forward-backward algorithm from [67], providing extensions to HV recovery. More details on the selection of the parameter μ from (3.70), key to satisfying the bound on the constrained residual from problem (3.69), are also presented.

4.1 Forward-backward splitting algorithm

In this section, we present a forward-backward splitting algorithm for solving problem (3.76). This algorithm can be slightly modified to solve problem (3.70), as will be detailed in Section 4.3. In Section 4.1.3, we prove convergence results for the forward-backward splitting algorithm, relevant to both problems (3.76) & (3.70). For simplicity, we assume that $\mu = 1$ in (3.76), noting that all of the analysis to follow holds in the case of arbitrary μ . We also assume that $\mathbf{u} \in \mathcal{H} := \{\mathbf{x} \in \mathbb{C}^{N \times \Omega} : \|\mathbf{x}\|_{2,2} < \infty\}$, so that (3.76) has solutions in $\mathcal{H}$.

Let $\phi_1(\mathbf{x}) = \|\mathbf{x}\|_{2,1}$ and $\phi_2(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{u}\|_{2,2}^2$. Then $\phi = \phi_1 + \phi_2$ represents a splitting of the objective of (3.76) into the non-differentiable and Fréchet differentiable parts, ϕ_1 and ϕ_2 , respectively. Define $T_1 = \partial\phi_1$, $T_2 = \partial\phi_2 = \{\nabla\phi_2\}$, and $T = \partial\phi = T_1 + T_2$, the solutions $\mathbf{x}^* \in \mathcal{H}$ of (3.76) are characterized by

$$\mathbf{0} \in \partial\phi_1(\mathbf{x}^*) + \{\nabla\phi_2(\mathbf{x}^*)\}, \quad \text{or} \quad \mathbf{0} \in (T_1 + T_2)(\mathbf{x}^*), \quad (4.1)$$

where $\partial\phi_1$ represents the subdifferential of ϕ_1 : for all $\mathbf{x} \in \mathcal{H}$,

$$\partial\phi_1(\mathbf{x}) = \{\mathbf{v} \in \mathcal{H} : \langle \mathbf{v}' - \mathbf{x}, \mathbf{v} \rangle_{2,2} + \phi_1(\mathbf{x}) \leq \phi_1(\mathbf{v}'), \quad \forall \mathbf{v}' \in \mathcal{H}\}. \quad (4.2)$$

Let $X^* := \{\mathbf{x} \in \mathcal{H} : \mathbf{0} \in (T_1 + T_2)(\mathbf{x})\}$, we aim to find $\mathbf{x}^* \in X^*$ via a formulation of the forward-backward splitting algorithm [18, 62, 22], which makes use of the splitting of T into T_1 and T_2 , derived in the setting of joint-sparse recovery. The forward-backward algorithm is a two-step fixed-point algorithm that involves an explicit (forward) step composed with an implicit (backward) step. It is efficient, in that it only involves alternating steps requiring relatively cheap computations, using the functions T_1 and T_2 separately. In this way it avoids direct computation of $(T_1 + T_2)^{-1}(\mathbf{0})$, which may not be feasible.

We derive the algorithm as follows. Let $\tau > 0$, then from (4.1) we have

$$\begin{aligned} \mathbf{0} \in T(\mathbf{x}) &\iff \mathbf{0} \in (\mathbf{x} + \tau T_1(\mathbf{x})) - (\mathbf{x} - \tau T_2(\mathbf{x})) \\ &\iff (I - \tau T_2)\mathbf{x} \in (I + \tau T_1)\mathbf{x} \\ &\iff \mathbf{x} = (I + \tau T_1)^{-1}(I - \tau T_2)\mathbf{x}. \end{aligned} \quad (4.3)$$

(4.3) is well-defined, since, as we shall see, $(I - \tau T_2)$ is single-valued and $(I + \tau T_1)$ is invertible. The last identity in (4.3) leads to the *forward-backward splitting* algorithm: given initial guess $\mathbf{x}^0 \in \mathcal{H}$, compute

$$\mathbf{x}^{k+1} = (I + \tau T_1)^{-1}(I - \tau T_2)\mathbf{x}^k, \quad (4.4)$$

where $\mathbf{x}^k$ denotes the approximation at the k -th iterate.

Let us define

$$J_\tau := (I + \tau T_1)^{-1}, \quad G_\tau := (I - \tau T_2), \quad \text{and} \quad S_\tau := J_\tau \circ G_\tau,$$

then (4.4) can be written as $\mathbf{x}^{k+1} = S_\tau(\mathbf{x}^k)$, and it is clear from (4.1) and (4.3) that X^* is the set of fixed points of S_τ . Let $[N]$ denote the set of integers in the interval $[1, N]$, i.e., the set $\{1, \dots, N\}$. For $j \in [N]$, by $J_{\tau,j}(\cdot) := (J_\tau(\cdot))_j$, $G_{\tau,j}(\cdot) := (G_\tau(\cdot))_j$, and $S_{\tau,j}(\cdot) := (S_\tau(\cdot))_j$, we denote the row-wise operators associated with J_τ , G_τ , and S_τ , respectively, and similarly for

$\mathbf{x} \in \mathcal{H}$ let $\mathbf{x}_j$ denote the j -th row of $\mathbf{x}$. In the subsequent part, we will derive the following formulations for G_τ and J_τ :

$$G_\tau(\mathbf{x}) = \mathbf{x} - \tau \mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{u}), \quad (4.5)$$

$$J_{\tau,j}(\mathbf{x}) = J_\tau(\mathbf{x}_j) = \frac{\mathbf{x}_j}{\|\mathbf{x}_j\|_2} \cdot \max\{\|\mathbf{x}_j\|_2 - \tau, 0\}, \quad 1 \leq j \leq N. \quad (4.6)$$

One can observe that the forward operator G_τ resembles a step of gradient descent algorithm with stepsize τ for minimizing ϕ_2 . The backward operator J_τ , on the other hand, is a soft thresholding step associated with proximal point method. As a result, algorithm (4.4) can be considered as an instance of *proximal-gradient* method.

Our analysis is conducted under the following assumption, which states that $\mathbf{A}^*\mathbf{A}$ (the Hessian of ϕ_2 , see Section 4.1.1) has bounded spectral norm, and that the step size τ is chosen appropriately with respect to its spectral radius.

Assumption 4.1. *Let $\mathbf{H} := \mathbf{A}^*\mathbf{A}$, then $\|\mathbf{H}\|_2 < +\infty$, and the step size τ in (4.4) satisfies $0 < \tau < 2/\|\mathbf{H}\|_2$.*

Under this assumption, we establish some nonexpansive properties for G_τ and J_τ , which are essential for our convergence proofs. In particular, G_τ is nonexpansive, i.e.,

$$\|G_\tau(\mathbf{v}) - G_\tau(\mathbf{w})\|_{2,2} \leq \|\mathbf{v} - \mathbf{w}\|_{2,2} \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{H}, \quad (4.7)$$

and J_τ is row-wise firmly nonexpansive, i.e.,

$$\begin{aligned} \|J_\tau(\mathbf{v}_j) - J_\tau(\mathbf{w}_j)\|_2^2 &\leq \|\mathbf{v}_j - \mathbf{w}_j\|_2^2 - \|(I - J_\tau)\mathbf{v}_j - (I - J_\tau)\mathbf{w}_j\|_2^2, \\ &\quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{H}, \forall j \in [N]. \end{aligned} \quad (4.8)$$

It is clear that the firmly nonexpansive property implies nonexpansiveness, though further properties can also be shown, see, e.g., [10, Proposition 4.2 & Corollary 4.3]. In Section 4.1.3, we will see (4.8) also has an interesting geometric interpretation about the angle between elements $\mathbf{x}^* \in X^*$ and the iterates $\{\mathbf{x}^k\}$. This interpretation, in concert with the well-known weak convergence of the algorithm in the general setting of monotone inclusion problems [10, 36], are the key to our strong convergence result.

Remark 4.2 (Strong convergence under stronger assumptions than Assumption 4.1). *We remark that further assumptions can be imposed on $\mathbf{H}$ to make G_τ a contraction on the whole (or certain subspaces) of $\mathcal{H}$. In such setting, the desired strong convergence can be obtained via a routine arguments from classical convergence theory (see the discussion in Section 4.1.3 on linear convergence). On the other hand, it is also possible to prove strong convergence in joint-sparse recovery with nonexpansiveness properties of (4.7) and (4.8) by slightly extending the available arguments in single vector recovery, e.g., [67], given Ω being a finite set (and thus, $\mathbb{C}^{N \times \Omega}$ being locally compact). We stress that neither of these assumptions is required in our analysis below.*

Remark 4.3 (Steps towards strong convergence in the Hilbert-valued setting). *As mentioned in the introduction to Chapter 4, the following analysis has trivial extension to the setting*

of Hilbert-valued recovery. The Hilbert-valued versions of the operators G_τ and J_τ (and corresponding fixed-point iterations with the composition S_τ) are given by

$$G_\tau(\mathbf{x}) = \mathbf{x} - \tau \mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{u}), \quad (4.9)$$

$$J_{\tau,j}(\mathbf{x}) = J_\tau(\mathbf{x}_j) = \frac{\mathbf{x}_j}{\|\mathbf{x}_j\|_{\mathcal{V}}} \cdot \max\{\|\mathbf{x}_j\|_{\mathcal{V}} - \tau, 0\}, \quad 1 \leq j \leq N, \quad (4.10)$$

$$\mathbf{x}^{k+1} = S_\tau(\mathbf{x}^k) := J_\tau \circ G_\tau(\mathbf{x}^k), \quad k \in \mathbb{N}_0 \quad (4.11)$$

where (4.10) is understood to act component-wise on Hilbert-valued vectors $\mathbf{x} \in \mathcal{V}^N$, $\mathbf{x}^0 \in \mathcal{V}^N$ is a given initial guess, and the matrix-Hilbert valued vector products from (4.9) are given by

$$[\mathbf{A}\mathbf{x}]_i = \sum_{j=1}^N \mathbf{A}_{i,j} \mathbf{x}_j \in \mathcal{V}, \quad \forall i = 1, \dots, m. \quad (4.12)$$

Similar to (4.7) & (4.8) above, nonexpansive and firmly nonexpansive properties can be shown with respect to the $\|\cdot\|_{\mathcal{V},2}$ and $\|\cdot\|_{\mathcal{V}}$ -norms for these operators. Application of the proof techniques described in Section 4.1.3 for joint-sparse recovery carry over to Hilbert-valued recovery with these replacements, based on properties of the $\|\cdot\|_{\mathcal{V}}$ -norm and associated $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ inner product. The setting of sparse recovery using (4.9)-(4.11) will be discussed in more detail in Sections 4.2 & 4.3.

4.1.1 Gradient descent step

First, it is easy to show that ϕ_2 is twice Fréchet differentiable on $\mathcal{H}$ with

$$T_2 = \nabla \phi_2(\mathbf{x}) = \mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{u}), \quad \nabla^2 \phi_2(\mathbf{x}) = \mathbf{A}^* \mathbf{A}, \quad \forall \mathbf{x} \in \mathcal{H}, \quad (4.13)$$

see, e.g., [10, Example 2.48]. Substituting, we have

$$G_\tau(\mathbf{x}) = (I - \tau T_2)\mathbf{x} = \mathbf{x} - \tau \nabla \phi_2(\mathbf{x}) = \mathbf{x} - \tau \mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{u}), \quad \forall \mathbf{x} \in \mathcal{H},$$

as in (4.5). The below lemma establishes the nonexpansiveness of G_τ in mixed norm $\|\cdot\|_{2,2}$, given Assumption 4.1.

Lemma 4.4. *Let $\mathbf{H}$ and τ satisfy Assumption 4.1. Then G_τ is nonexpansive.*

Proof. Let $\mathbf{v}, \mathbf{w} \in \mathcal{H}$ be arbitrary. Since

$$\|T_2(\mathbf{v}) - T_2(\mathbf{w})\|_{2,2} = \|\mathbf{A}^* \mathbf{A}(\mathbf{v} - \mathbf{w})\|_{2,2} \leq \|\mathbf{H}\|_2 \|\mathbf{v} - \mathbf{w}\|_{2,2},$$

T_2 is Lipschitz continuous with constant $\|\mathbf{H}\|_2 = 1/\beta$, implying βT_2 is firmly nonexpansive, see [10, Definition 4.4 & Corollary 18.16], and $R := 2\beta T_2 - I$ is nonexpansive. We have

$T_2 = \frac{1}{2\beta}(I + R)$, then $G_\tau = I - \tau T_2 = (1 - \frac{\tau}{2\beta})I + \frac{\tau}{2\beta}(-R)$. There follows

$$\begin{aligned}\|G_\tau(\mathbf{v}) - G_\tau(\mathbf{w})\|_{2,2} &= \left\| \left(\left(1 - \frac{\tau}{2\beta}\right)I + \frac{\tau}{2\beta}(-R) \right) (\mathbf{v} - \mathbf{w}) \right\|_{2,2} \\ &\leq \left(1 - \frac{\tau}{2\beta}\right) \|\mathbf{v} - \mathbf{w}\|_{2,2} + \frac{\tau}{2\beta} \|R(\mathbf{v} - \mathbf{w})\|_{2,2} \leq \|\mathbf{v} - \mathbf{w}\|_{2,2},\end{aligned}$$

from which we conclude the nonexpansive property for G_τ . ■

4.1.2 Proximal point step

We denote $\mathcal{H}^0 = \{\mathbf{z} \in \mathbb{C}^\Omega : \|\mathbf{z}\|_2 < \infty\}$. To derive formulation (4.6) for J_τ , first we need the following lemma.

Lemma 4.5. *Let $\mathbf{z} \in \mathcal{H}^0$, then*

$$\partial\|\mathbf{z}\|_2 = \begin{cases} \{\mathbf{z}/\|\mathbf{z}\|_2\}, & \mathbf{z} \neq \mathbf{0}, \\ B_2(\mathbf{0}, 1), & \mathbf{z} = \mathbf{0}, \end{cases} \quad (4.14)$$

where $B_2(\mathbf{0}, 1)$ is the unit ball in $\mathcal{H}^0$.

Proof. First consider the case $\mathbf{z} = \mathbf{0}$. Let $\mathbf{v} \in \partial\|\mathbf{z}\|_2$, then for $\mathbf{z}' \in \mathcal{H}^0$ arbitrary, $\langle \mathbf{v}, \mathbf{z}' - \mathbf{z} \rangle_2 + \|\mathbf{z}\|_2 = \langle \mathbf{v}, \mathbf{z}' \rangle_2 \leq \|\mathbf{z}'\|_2$. Setting $\mathbf{z}' = \mathbf{v}$ gives $\|\mathbf{v}\|_2 \leq 1$, implying $\partial\|\mathbf{z}\|_2 \subseteq B_2(\mathbf{0}, 1)$. On the other hand, for $\mathbf{v} \in B_2(\mathbf{0}, 1)$, let $\mathbf{z}' \in \mathcal{H}^0$ arbitrary, we have $\langle \mathbf{v}, \mathbf{z}' \rangle_2 \leq \|\mathbf{v}\|_2 \|\mathbf{z}'\|_2 \leq \|\mathbf{z}'\|_2$. This implies $\mathbf{v} \in \partial\|\mathbf{z}\|_2$; hence, $B_2(\mathbf{0}, 1) = \partial\|\mathbf{z}\|_2$ when $\mathbf{z} = \mathbf{0}$.

Now suppose $\mathbf{z} \neq \mathbf{0}$. Let $\mathbf{v} \in \partial\|\mathbf{z}\|_2$, then $\langle \mathbf{v}, \mathbf{z}' - \mathbf{z} \rangle_2 + \|\mathbf{z}\|_2 \leq \|\mathbf{z}'\|_2$ for $\mathbf{z}' \in \mathcal{H}^0$ arbitrary. Choosing $\mathbf{z}' = \mathbf{0}$ and $\mathbf{z}' = 2\mathbf{z}$ gives $\|\mathbf{z}\|_2 = \langle \mathbf{v}, \mathbf{z} \rangle_2$, implying $\mathbf{v} = \mathbf{z}/\|\mathbf{z}\|_2$ and $\partial\|\mathbf{z}\|_2 = \{\mathbf{z}/\|\mathbf{z}\|_2\}$ when $\mathbf{z} \neq \mathbf{0}$. ■

With Lemma 4.5, for every $\mathbf{z} \in \mathcal{H}^0$, we can write

$$(I + \tau \partial\|\cdot\|_2)\mathbf{z} = \begin{cases} \{(1 + \tau/\|\mathbf{z}\|_2)\mathbf{z}\} & \mathbf{z} \neq \mathbf{0}, \\ B_2(\mathbf{0}, \tau) & \mathbf{z} = \mathbf{0}, \end{cases}$$

from which it is clear that for every $\mathbf{x} \in \mathcal{H}$, $j \in [N]$,

$$((I + \tau T_1)\mathbf{x})_j = \begin{cases} \{(1 + \tau/\|\mathbf{x}_j\|_2)\mathbf{x}_j\}, & \mathbf{x}_j \neq \mathbf{0}, \\ B_2(\mathbf{0}, \tau), & \mathbf{x}_j = \mathbf{0}. \end{cases}$$

A straightforward computation gives

$$\begin{aligned}J_{\tau,j}(\mathbf{x}) &= J_\tau(\mathbf{x}_j) = \begin{cases} (1 - \tau/\|\mathbf{x}_j\|_2)\mathbf{x}_j & \|\mathbf{x}_j\|_2 > \tau, \\ \mathbf{0} & \|\mathbf{x}_j\|_2 \leq \tau, \end{cases} \\ &\equiv \frac{\mathbf{x}_j}{\|\mathbf{x}_j\|_2} \cdot \max\{\|\mathbf{x}_j\|_2 - \tau, 0\}, \quad \forall \mathbf{x} \in \mathcal{H}, \end{aligned} \quad (4.15)$$

thus, J_τ is equivalent to row-wise soft thresholding on $\mathcal{H}$. Equation (4.15) can also be written

$$J_{\tau,j}(\mathbf{x}) = (I - \mathcal{P}_\tau)(\mathbf{x}_j), \quad (4.16)$$

where $\mathcal{P}_\tau$ is metric projection from $\mathcal{H}^0$ onto $B_2(\mathbf{0}, \tau)$, i.e., for $\mathbf{x}_j \in \mathcal{H}^0$,

$$\mathcal{P}_\tau(\mathbf{x}_j) = \begin{cases} \tau \frac{\mathbf{x}_j}{\|\mathbf{x}_j\|_2} & \mathbf{x}_j \notin B_2(\mathbf{0}, \tau), \\ \mathbf{x}_j & \mathbf{x}_j \in B_2(\mathbf{0}, \tau). \end{cases} \quad (4.17)$$

Our next result establishes that $J_{\tau,j}$ is firmly nonexpansive for each $j \in [N]$.

Lemma 4.6. *Let $\tau > 0$ be arbitrary, then $J_\tau : \mathcal{H} \rightarrow \mathcal{H}$ is row-wise firmly nonexpansive, i.e., satisfies (4.8).*

Proof. The proof follows arguments from [10]. We include it here for completeness. Since $B_2(\mathbf{0}, \tau)$ is a nonempty closed convex subset of $\mathcal{H}^0$, and $\mathcal{P}_\tau$ is a projection, we have $\langle \mathcal{P}_\tau \mathbf{v}_j - \mathcal{P}_\tau \mathbf{w}_j, \mathbf{v}_j - \mathcal{P}_\tau \mathbf{w}_j \rangle_2 \leq 0$ and $\langle \mathcal{P}_\tau \mathbf{w}_j - \mathcal{P}_\tau \mathbf{v}_j, \mathbf{v}_j - \mathcal{P}_\tau \mathbf{v}_j \rangle_2 \leq 0$ for every $\mathbf{v}_j, \mathbf{w}_j \in \mathcal{H}^0$. Adding, we obtain

$$\langle \mathcal{P}_\tau \mathbf{v}_j - \mathcal{P}_\tau \mathbf{w}_j, \mathbf{v}_j - \mathbf{w}_j \rangle_2 \geq \|\mathcal{P}_\tau \mathbf{v}_j - \mathcal{P}_\tau \mathbf{w}_j\|_2^2.$$

It follows that

$$\begin{aligned} & \|(I - \mathcal{P}_\tau) \mathbf{v}_j - (I - \mathcal{P}_\tau) \mathbf{w}_j\|_2^2 \\ &= \|\mathbf{v}_j - \mathbf{w}_j\|_2^2 + \|\mathcal{P}_\tau \mathbf{v}_j - \mathcal{P}_\tau \mathbf{w}_j\|_2^2 - 2\langle \mathbf{v}_j - \mathbf{w}_j, \mathcal{P}_\tau \mathbf{v}_j - \mathcal{P}_\tau \mathbf{w}_j \rangle_2 \\ &\leq \|\mathbf{v}_j - \mathbf{w}_j\|_2^2 - \|\mathcal{P}_\tau \mathbf{v}_j - \mathcal{P}_\tau \mathbf{w}_j\|_2^2, \end{aligned}$$

which, with (4.16), implies (4.8). ■

4.1.3 Convergence results

In this section, we present our main results showing that the sequence $\{\mathbf{x}^k\}$ obtained by iterating (4.4) converges strongly to an element $\mathbf{x}^* \in X^*$ from any initial guess $\mathbf{x}^0 \in \mathcal{H}$. These results apply for Ω being either a finite or infinite set. Our analysis relies on the following partition of the index set $[N]$, inspired by [67, Definition 4.3], for the joint-sparse recovery setting.

Definition 4.7. *For $\mathbf{x}^* \in X^*$, $j \in [N]$, let $\nabla \phi_{2,j}(\mathbf{x}^*) := (\nabla \phi_2(\mathbf{x}^*))_j$, we define*

$$L := \{j \in [N] : \|\nabla \phi_{2,j}(\mathbf{x}^*)\|_2 < 1\}, \quad E := \{j \in [N] : \|\nabla \phi_{2,j}(\mathbf{x}^*)\|_2 = 1\}, \quad (4.18)$$

$$\omega := \min\{\tau(1 - \|\nabla \phi_{2,j}(\mathbf{x}^*)\|_2) : j \in L\}. \quad (4.19)$$

The intuition for this partition can be derived as follows. It is easy to see from the definition of X^* and the subdifferential of ϕ_1 and Lemma 4.5 that

$$\mathbf{x}^* \in X^* \iff \nabla \phi_{2,j}(\mathbf{x}^*) \in \begin{cases} \{-\mathbf{x}_j^*/\|\mathbf{x}_j^*\|_2\}, & \mathbf{x}_j^* \neq \mathbf{0}, \\ B_2(\mathbf{0}, 1), & \mathbf{x}_j^* = \mathbf{0}, \end{cases} \quad \forall j \in [N],$$

implying that $\|\nabla\phi_{2,j}(\mathbf{x}^*)\|_2 \leq 1$ for all $j \in [N]$ and $\mathbf{x}^* \in X^*$. Hence $L \cup E = [N]$, and

$$\text{supp}(\mathbf{x}^*) \subseteq E, \quad \text{and} \quad L \subset (\text{supp}(\mathbf{x}^*))^c. \quad (4.20)$$

Furthermore, if $\mathbf{x}^* \in X^*$ and $j \in L$,

$$\tau(1 - \|\nabla\phi_{2,j}(\mathbf{x}^*)\|_2) = \tau - \|\mathbf{x}_j^* - \tau\nabla\phi_{2,j}(\mathbf{x}^*)\|_2 = \tau - \|G_{\tau,j}(\mathbf{x}^*)\|_2,$$

therefore

$$\min\{\tau - \|G_{\tau,j}(\mathbf{x}^*)\|_2 : j \in L\} = \omega > 0. \quad (4.21)$$

Using this partition of $[N]$, our strong convergence result is obtained in three steps:

1. Establish finite convergence on the set L , i.e., determine a bound on the number of iterations K such that $\mathbf{x}_j^k = \mathbf{x}_j^* = \mathbf{0}$ for each $j \in L$ whenever $k \geq K$. Therefore, in light of (4.20), the iterations (4.4) partially identify the support of an element $\mathbf{x}^* \in X^*$ after a finite number of iterations (Lemma 4.10),
2. Establish angular convergence: $\theta_j^k \rightarrow 0$ as $k \rightarrow \infty$, where, for $j \in \text{supp}(\mathbf{x}^*) \subseteq E$, θ_j^k is the angle between $\mathbf{x}_j^k$ and $\mathbf{x}_j^*$, and, for $j \in E \setminus \text{supp}(\mathbf{x}^*)$, θ_j^k is the angle between $\mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^k)$ and $\mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^*)$ (Theorem 4.11),
3. Combine the known weak convergence result, e.g., [36], with the angular convergence to obtain the convergence in norm on E (Theorem 4.13).

First, we show that the sets L and E defined above are invariant on X^* , thereby justifying the use of Definition 4.7 in studying the convergence of $\{\mathbf{x}^k\}$ to an arbitrary element $\mathbf{x}^* \in X^*$.

Lemma 4.8. *Under Assumption 4.1, for every $\mathbf{x}, \mathbf{x}' \in X^*$, $\mathbf{H}\mathbf{x} = \mathbf{H}\mathbf{x}'$, and hence $\nabla\phi_2(\mathbf{x}) = \nabla\phi_2(\mathbf{x}')$.*

Proof. Since J_τ and G_τ are nonexpansive, $\mathbf{x}, \mathbf{x}' \in X^*$ gives

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}'\|_{2,2} &= \|S_\tau(\mathbf{x}) - S_\tau(\mathbf{x}')\|_{2,2} \\ &\leq \|G_\tau(\mathbf{x}) - G_\tau(\mathbf{x}')\|_{2,2} \leq \|\mathbf{x} - \mathbf{x}'\|_{2,2}, \end{aligned} \quad (4.22)$$

and we may replace both inequalities with equalities. Letting $\mathbf{z} = \mathbf{x} - \mathbf{x}'$ and $\mathbf{v} = \mathbf{H}^{1/2}\mathbf{z}$, from (4.22) we have

$$\begin{aligned} \|G_\tau(\mathbf{x}) - G_\tau(\mathbf{x}')\|_{2,2} = \|\mathbf{z}\|_{2,2} &\iff \|(I - \tau\mathbf{H})\mathbf{z}\|_{2,2} = \|\mathbf{z}\|_{2,2} \\ &\iff -2\tau\langle \mathbf{z}, \mathbf{H}\mathbf{z} \rangle_{2,2} + \tau^2\langle \mathbf{H}\mathbf{z}, \mathbf{H}\mathbf{z} \rangle_{2,2} = 0 \\ &\iff \tau\langle \mathbf{v}, \mathbf{H}\mathbf{v} \rangle_{2,2} = 2\langle \mathbf{v}, \mathbf{v} \rangle_{2,2} \\ &\implies \tau \frac{\langle \mathbf{v}, \mathbf{H}\mathbf{v} \rangle_{2,2}}{\langle \mathbf{v}, \mathbf{v} \rangle_{2,2}} = 2, \quad \text{if } \mathbf{v} \neq 0, \end{aligned}$$

contradicting Assumption 4.1, since $\frac{\langle \mathbf{v}, \mathbf{H}\mathbf{v} \rangle_{2,2}}{\langle \mathbf{v}, \mathbf{v} \rangle_{2,2}} \leq \|\mathbf{H}\|_2 < \frac{2}{\tau}$. Hence, $\mathbf{v} = \mathbf{0}$ and (4.22) gives

$$\mathbf{H}(\mathbf{x} - \mathbf{x}') = \mathbf{H}^{1/2}\mathbf{v} = \mathbf{0}, \quad \forall \mathbf{x}, \mathbf{x}' \in X^*. \quad (4.23)$$

This yields $\mathbf{H}\mathbf{x} = \mathbf{H}\mathbf{x}'$, and since $\nabla\phi_2(\mathbf{x}) = \mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{u})$, $\forall \mathbf{x} \in \mathcal{H}$, the result follows. ■

Our next lemma concerns the weak convergence properties of the algorithm. Weak convergence of the forward-backward iterations has been well-established in the general case of monotone inclusion problems [36, 10]. Such problems include (3.76), and therefore the weak convergence holds in this setting as well.

Lemma 4.9. *Let Assumption 4.1 hold and $\{\mathbf{x}^k\}$ be generated by the forward-backward iterations (4.4) starting from any $\mathbf{x}^0 \in \mathcal{H}$. Then $\{\mathbf{x}^k\}$ converges weakly to some $\mathbf{x}^* \in X^*$.*

Proof. Since firmly nonexpansive operators are nonexpansive and G_τ is nonexpansive, it follows that the composition, S_τ , is also nonexpansive. The Browder-Göhde-Kirk Theorem and [10, Corollary 4.15] then imply that X^* , the set of fixed points of S_τ , is a nonempty, closed, and convex subset of $\mathcal{H}$. Therefore problem (3.76) and the algorithm (4.4) satisfy the assumptions of [36, Theorem 6.3 and Corollary 6.5], and the weak convergence of our forward-backward iterations follows. ■

Finite convergence in the support complement and angular convergence

In this section, we show two results key in proving strong convergence. The first says that the rows associated with the set $L \subset (\text{supp}(\mathbf{x}^*))^c$ of $\{\mathbf{x}^k\}$ converge to $\mathbf{0}$ in finitely many iterations, thus coinciding with $\mathbf{x}^*$.

Lemma 4.10. *Let Assumption 4.1 hold and $\{\mathbf{x}^k\}$ be generated by the forward-backward iterations (4.4) starting from any $\mathbf{x}^0 \in \mathcal{H}$. Then $\mathbf{x}_j^k = \mathbf{x}_j^* = \mathbf{0} \quad \forall j \in L$, for all but at most $\|\mathbf{x}^0 - \mathbf{x}^*\|_{2,2}^2/\omega^2$ iterations, with ω given in (4.19).*

Proof. Let $\mathbf{x}^* \in X^*$ and $j \in L$. Then $\mathbf{x}_j^* = \mathbf{0}$ by (4.20), so that $S_{\tau,j}(\mathbf{x}^*) = \mathbf{0}$. Suppose $\mathbf{x}_j^k \neq \mathbf{0}$ for some $k \in \mathbb{N}_0$. If $\|G_{\tau,j}(\mathbf{x}^k)\|_2 \leq \tau$, then the result follows after $k+1$ iterations, since $\mathbf{x}_j^{k+1} = S_{\tau,j}(\mathbf{x}^k) = J_{\tau,j} \circ G_{\tau,j}(\mathbf{x}^k) = \mathbf{0}$ by (4.15). Otherwise,

$$\|\mathbf{x}_j^{k+1}\|_2 = \left\| \left(1 - \frac{\tau}{\|G_{\tau,j}(\mathbf{x}^k)\|_2} \right) G_{\tau,j}(\mathbf{x}^k) \right\|_2 = \|G_{\tau,j}(\mathbf{x}^k)\|_2 - \tau > 0,$$

implying

$$\begin{aligned} \|\mathbf{x}_j^{k+1} - \mathbf{x}_j^*\|_2 &= \|G_{\tau,j}(\mathbf{x}^k) - G_{\tau,j}(\mathbf{x}^*) + G_{\tau,j}(\mathbf{x}^*)\|_2 - \tau \\ &\leq \|G_{\tau,j}(\mathbf{x}^k) - G_{\tau,j}(\mathbf{x}^*)\|_2 - (\tau - \|G_{\tau,j}(\mathbf{x}^*)\|_2). \end{aligned}$$

Hence, by (4.21),

$$\begin{aligned} \|\mathbf{x}_j^{k+1} - \mathbf{x}_j^*\|_2^2 &< \|G_{\tau,j}(\mathbf{x}^k) - G_{\tau,j}(\mathbf{x}^*)\|_2^2 - (\tau - \|G_{\tau,j}(\mathbf{x}^*)\|_2)^2 \\ &\leq \|G_{\tau,j}(\mathbf{x}^k) - G_{\tau,j}(\mathbf{x}^*)\|_2^2 - \omega^2. \end{aligned}$$

The row-wise nonexpansiveness of J_τ and the nonexpansiveness of G_τ implies

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{2,2}^2 \leq \|G_\tau(\mathbf{x}^k) - G_\tau(\mathbf{x}^*)\|_{2,2}^2 - \omega^2 \leq \|\mathbf{x}^k - \mathbf{x}^*\|_{2,2}^2 - \omega^2. \quad (4.24)$$

Applying (4.24) inductively gives $0 \leq \|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{2,2}^2 \leq \|\mathbf{x}^0 - \mathbf{x}^*\|_{2,2}^2 - k\omega^2$, showing that the number of iterations for which $\mathbf{x}_j^k \neq \mathbf{0}$ for $j \in L$ satisfies $k \leq \|\mathbf{x}^0 - \mathbf{x}^*\|_{2,2}^2 / \omega^2$. ■

Let $\angle(\mathbf{z}, \mathbf{z}')$ denote the angle between two nonzero vectors $\mathbf{z}, \mathbf{z}' \in \mathcal{H}^0$, i.e.,

$$\angle(\mathbf{z}, \mathbf{z}') := \cos^{-1} \left(\frac{\langle \mathbf{z}, \mathbf{z}' \rangle_2}{\|\mathbf{z}\|_2 \|\mathbf{z}'\|_2} \right),$$

our next result shows the angular convergence properties of the forward-backward algorithm using the firmly nonexpansive property. This is a generalization of [67, Lemma 5.2, part 2], which establishes that the signs of the components of gradient steps $G_\tau(\mathbf{x}^k)$ (defined in (4.5)) agree with those of $G_\tau(\mathbf{x}^*)$ for all but finitely many $k \in \mathbb{N}$, in case $\mathbf{x}^k, \mathbf{x}^* \in \mathbb{R}^N$. In essence, the sign function was used to partition $\mathbb{R}^N$; and the difference $\mathbf{x}^{k+1} - \mathbf{x}^*$ was proved to reduce a fixed amount from the previous step whenever the signs do not match, yielding that the mismatch can only hold for finitely many steps. Here, we derive a similar argument directly from the firmly nonexpansive property (4.8).

Theorem 4.11. *Let Assumption 4.1 hold, $\mathbf{x}^* \in X^*$, and $\{\mathbf{x}^k\}$ be generated by the forward-backward iterations (4.4) starting from any $\mathbf{x}^0 \in \mathcal{H}$. Then*

1. $\angle(\mathbf{x}_j^k, \mathbf{x}_j^*) \rightarrow 0$ as $k \rightarrow \infty$ for each $j \in \text{supp}(\mathbf{x}^*)$,
2. $\angle(\mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^k), \mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^*)) \rightarrow 0$ as $k \rightarrow \infty$ for each $j \in E \setminus \text{supp}(\mathbf{x}^*)$.

Proof. Let $\mathbf{z}, \mathbf{z}' \in \mathcal{H}$ and $j \in [N]$ be arbitrary. Then $\mathcal{P}_\tau(\mathbf{z}_j) = \kappa \mathbf{z}_j$ and $J_{\tau,j}(\mathbf{z}) = (1 - \kappa)\mathbf{z}_j$ where $\kappa := \min\{\tau/\|\mathbf{z}_j\|_2, 1\} \in (0, 1]$, from (4.15) and (4.17). Hence $\mathbf{z}_j$, $\mathcal{P}_\tau(\mathbf{z}_j)$, and $J_{\tau,j}(\mathbf{z})$ are all collinear. We observe that collinearity implies

$$\begin{aligned} \angle(\mathbf{z}_j, \mathbf{z}'_j) &= \angle(\mathcal{P}_\tau(\mathbf{z}_j), \mathcal{P}_\tau(\mathbf{z}'_j)), & \forall \mathbf{z}_j, \mathbf{z}'_j \neq \mathbf{0}, \\ \text{and } \angle(\mathbf{z}_j, \mathbf{z}'_j) &= \angle(J_\tau(\mathbf{z}_j), J_\tau(\mathbf{z}'_j)), & \forall \mathbf{z}_j, \mathbf{z}'_j \notin B_2(\mathbf{0}, \tau). \end{aligned} \quad (4.25)$$

Since J_τ is component-wise firmly nonexpansive

$$\begin{aligned} \|\mathbf{x}_j^{k+1} - \mathbf{x}_j^*\|_2^2 &= \|J_{\tau,j} \circ G_{\tau,j}(\mathbf{x}^k) - J_{\tau,j} \circ G_{\tau,j}(\mathbf{x}^*)\|_2^2 \\ &\leq \|G_{\tau,j}(\mathbf{x}^k) - G_{\tau,j}(\mathbf{x}^*)\|_2^2 \\ &\quad - \|(I - J_{\tau,j}) \circ G_{\tau,j}(\mathbf{x}^k) - (I - J_{\tau,j}) \circ G_{\tau,j}(\mathbf{x}^*)\|_2^2 \\ &= \|G_{\tau,j}(\mathbf{x}^k) - G_{\tau,j}(\mathbf{x}^*)\|_2^2 - \|\mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^k) - \mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^*)\|_2^2, \end{aligned} \quad (4.26)$$

for each $j \in [N]$ and $k \in \mathbb{N}_0$, where the last equality follows from (4.16). Let

$$c_j^k := \|\mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^k) - \mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^*)\|_2^2, \quad \bar{c}^k := \sum_{j \in [N]} c_j^k. \quad (4.27)$$

From the nonexpansiveness of G_τ , summing (4.26) and iterating, it follows

$$\begin{aligned} \|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{2,2}^2 &\leq \|G_\tau(\mathbf{x}^k) - G_\tau(\mathbf{x}^*)\|_{2,2}^2 - \bar{c}^k \leq \|\mathbf{x}^k - \mathbf{x}^*\|_{2,2}^2 - \bar{c}^k \\ &\leq \dots \leq \|\mathbf{x}^0 - \mathbf{x}^*\|_{2,2}^2 - \sum_{\ell=0}^k \bar{c}^\ell. \end{aligned} \quad (4.28)$$

Since $\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{2,2}^2 \geq 0$, it follows that $\sum_{\ell=0}^k \bar{c}^\ell \leq \|\mathbf{x}^0 - \mathbf{x}^*\|_{2,2}^2$. However, the right hand side is independent of k , so that taking limits gives $\sum_{\ell=0}^\infty \bar{c}^\ell \leq \|\mathbf{x}^0 - \mathbf{x}^*\|_{2,2}^2$. Hence $\bar{c}^k \rightarrow 0$ as $k \rightarrow \infty$, and, from (4.27), it follows that for each $j \in [N]$,

$$\mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^k) \rightarrow \mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^*), \quad \text{as } k \rightarrow \infty. \quad (4.29)$$

Now, let $j \in \text{supp}(\mathbf{x}^*)$. Then, from (4.15),

$$\mathbf{0} \neq \mathbf{x}_j^* = J_{\tau,j} \circ G_{\tau,j}(\mathbf{x}^*) = \frac{G_{\tau,j}(\mathbf{x}^*)}{\|G_{\tau,j}(\mathbf{x}^*)\|_2} \cdot \max\{\|G_{\tau,j}(\mathbf{x}^*)\|_2 - \tau, 0\},$$

implying $\|G_{\tau,j}(\mathbf{x}^*)\|_2 > \tau$, so that $\|\mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^*)\|_2 = \tau$. For $k \in \mathbb{N}_0$, observe that $\|G_{\tau,j}(\mathbf{x}^k)\|_2 \leq \tau$ yields $\mathbf{x}_j^{k+1} = J_{\tau,j} \circ G_{\tau,j}(\mathbf{x}^k) = \mathbf{0}$. Therefore, if $\|G_{\tau,j}(\mathbf{x}^k)\|_2 \leq \tau$ holds for infinitely many k , there exist infinitely many k such that $\mathbf{x}_j^{k+1} = \mathbf{0}$, contradicting the fact that $\mathbf{x}_j^k \rightarrow \mathbf{x}_j^* \neq \mathbf{0}$ (Lemma 4.9). This gives $\|G_{\tau,j}(\mathbf{x}^k)\|_2 > \tau$ for all but finitely many $k \in \mathbb{N}_0$. Let $K > 0$ be such that $\|G_{\tau,j}(\mathbf{x}^k)\|_2 > \tau$ and $\mathbf{x}_j^{k+1} \neq \mathbf{0}$, $\forall k \geq K$, we have from (4.25) that

$$\angle(\mathbf{x}_j^{k+1}, \mathbf{x}_j^*) = \angle(J_\tau \circ G_{\tau,j}(\mathbf{x}^k), J_\tau \circ G_{\tau,j}(\mathbf{x}^*)) = \angle(\mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^k), \mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^*)),$$

which, combined with $\mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^k) \rightarrow \mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^*)$, implies $\angle(\mathbf{x}_j^k, \mathbf{x}_j^*) \rightarrow 0$ as $k \rightarrow \infty$.

On the other hand, if $j \in E \setminus \text{supp}(\mathbf{x}^*)$, then $\mathbf{x}_j^* = \mathbf{0}$ and from (4.18),

$$\|G_{\tau,j}(\mathbf{x}^*)\|_2 = \|\mathbf{x}_j^* - \tau \nabla \phi_{2,j}(\mathbf{x}^*)\|_2 = \tau \|\nabla \phi_{2,j}(\mathbf{x}^*)\|_2 = \tau,$$

implying $\mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^*) = G_{\tau,j}(\mathbf{x}^*) \neq \mathbf{0}$. Then, from (4.29), one has $\angle(\mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^k), \mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^*)) \rightarrow 0$ as $k \rightarrow \infty$. ■

Strong convergence

In this section, we establish our main result on strong convergence without the strict convexity and compactness assumptions discussed in Remark 4.2. First, we require the following lemma, showing that in general, strong convergence can be implied by weak convergence plus angular convergence.

Lemma 4.12. *Let $\{\mathbf{z}^k\} \subset \mathcal{H}^0$ be such that $\mathbf{z}^k \rightharpoonup \mathbf{z}^*$ for some nonzero $\mathbf{z}^* \in \mathcal{H}^0$. Let $\theta^k = \angle(\mathbf{z}^k, \mathbf{z}^*)$. If $\theta^k \rightarrow 0$ as $k \rightarrow \infty$, then $\mathbf{z}^k \rightarrow \mathbf{z}^*$ as $k \rightarrow \infty$.*

Proof. It is enough to show $\|\mathbf{z}^k\|_2 \rightarrow \|\mathbf{z}^*\|_2$, since this fact, in combination with $\mathbf{z}^k \rightharpoonup \mathbf{z}^*$, implies

$$\|\mathbf{z}^k - \mathbf{z}^*\|_2^2 = \|\mathbf{z}^k\|_2^2 + \|\mathbf{z}^*\|_2^2 - 2\langle \mathbf{z}^k, \mathbf{z}^* \rangle_2 \rightarrow 0 \quad \text{as } k \rightarrow \infty.$$

First, the weak convergence gives

$$\|\mathbf{z}^k\|_2 \cos \theta^k = \langle \mathbf{z}^k, \mathbf{z}^* \rangle_2 / \|\mathbf{z}^*\|_2 \rightarrow \langle \mathbf{z}^*, \mathbf{z}^* \rangle_2 / \|\mathbf{z}^*\|_2 = \|\mathbf{z}^*\|_2, \quad k \rightarrow \infty,$$

and the angular convergence $\theta^k \rightarrow 0$ gives $\cos \theta^k \rightarrow 1$ as $k \rightarrow \infty$. On the other hand, the weak convergence also gives $\|\mathbf{z}^k\|_2 \leq M$, $\forall k \in \mathbb{N}$ for some $M > 0$. Therefore

$$\|\mathbf{z}^k\|_2 - \|\mathbf{z}^*\|_2 = \|\mathbf{z}^k\|_2(1 - \cos \theta^k) + (\|\mathbf{z}^k\|_2 \cos \theta^k - \|\mathbf{z}^*\|_2) \rightarrow 0$$

as $k \rightarrow \infty$, as desired. ■

With Lemma 4.12, together with the weak and angular convergence established for forward-backward splitting (4.4) in previous subsections, we are now ready to prove the strong convergence of the iterates $\{\mathbf{x}^k\}$.

Theorem 4.13. *Let Assumption 4.1 hold and $\{\mathbf{x}^k\}$ be generated by the forward-backward iterations (4.4) starting from any $\mathbf{x}^0 \in \mathcal{H}$. Then $\{\mathbf{x}^k\}$ converges strongly to some $\mathbf{x}^* \in X^*$.*

Proof. For $j \in L$, Lemma 4.10 shows that $\mathbf{x}_j^k \rightarrow \mathbf{x}_j^* = \mathbf{0}$ in a finite number of iterations. For $j \in \text{supp}(\mathbf{x}^*)$, Theorem 4.11 shows $\angle(\mathbf{x}_j^k, \mathbf{x}_j^*) \rightarrow 0$ as $k \rightarrow \infty$. Combining with the weak convergence in Lemma 4.9 and the sufficient condition for strong convergence of Lemma 4.12, this yields $\mathbf{x}_j^k \rightarrow \mathbf{x}_j^* \forall j \in L$.

It remains to consider the case $j \in E \setminus \text{supp}(\mathbf{x}^*)$. First, we have $\mathbf{x}_j^* = \mathbf{0}$ and $\|G_{\tau,j}(\mathbf{x}^*)\|_2 = \|\mathbf{x}_j^* - \tau \nabla \phi_{2,j}(\mathbf{x}^*)\|_2 = \tau$, implying $\|\mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^*)\|_2 = \tau$. Let $\{\mathbf{x}_j^{k_n}\}$ be the subsequence of all nonzero elements of $\{\mathbf{x}_j^k\}$, it is enough to show $\mathbf{x}_j^{k_n} \rightarrow \mathbf{0}$. Since $\mathbf{x}_j^{k_n} \neq \mathbf{0} \forall n$, $\theta^n := \angle(\mathbf{x}_j^{k_n}, \mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^*))$ is well-defined, and from Theorem 4.11,

$$\theta^n = \angle(\mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^{k_n-1}), \mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^*)) \rightarrow 0, \quad \text{as } n \rightarrow \infty,$$

implying $\cos \theta^n \rightarrow 1$. From the weak convergence property of $\mathbf{x}_j^k$,

$$\|\mathbf{x}_j^{k_n}\|_2 \cos \theta^n = \langle \mathbf{x}_j^{k_n}, \mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^*) \rangle_2 / \tau \rightarrow \langle \mathbf{x}_j^*, \mathcal{P}_\tau \circ G_{\tau,j}(\mathbf{x}^*) \rangle_2 / \tau = 0.$$

Since $\cos \theta^n \rightarrow 1$, this gives $\|\mathbf{x}_j^{k_n}\|_2 \rightarrow 0$, concluding the proof. ■

Linear convergence

In this last subsection, we discuss a sufficient condition to establish linear convergence for the forward-backward splitting method in joint sparse recovery. Recall that the sequence $\{\|\mathbf{x}^k - \mathbf{x}^*\|_{2,2}\}$ converges to zero q -linearly if its q_1 -factor satisfies

$$q_1 := \limsup_{k \rightarrow \infty} \frac{\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{2,2}}{\|\mathbf{x}^k - \mathbf{x}^*\|_{2,2}} < 1.$$

In [67, Section 4.2], under an additional assumption imposing the well-conditioning of a “reduced” Hessian of ϕ_2 , q -linear convergence of the forward-backward splitting method was shown for the single vector recovery problem. With the same assumption (stated in Theorem 4.14 below), we are able to establish the linear convergence for our considered problem. This assumption aims to make the forward operator G_τ a contraction. However, as one already has the finite convergence on L , roughly speaking, this operator only needs to be a contraction on E , explaining why the well-conditioning of a submatrix of $\mathbf{H}$ associated with E is sufficient. The discussion below also demonstrates that given a contractive-type property, the path to acquire strong convergence is quite routine.

Let us define

$$\tau(\lambda) := \frac{\gamma(\lambda)}{\gamma(\lambda) + 1} \frac{2}{\lambda_{\max}(\mathbf{H})}, \quad \gamma(\lambda) := \frac{\lambda_{\max}(\mathbf{H})}{\lambda}. \quad (4.30)$$

Then $\tau(\lambda) \in (0, 2/\lambda_{\max}(\mathbf{H}))$, since $\gamma(\lambda) > 0$. By $\mathbf{H}_{EE}$ we define the submatrix of $\mathbf{H}$ formed by the rows and columns associated with E .

Theorem 4.14. *Let Assumption 4.1 hold, and further assume that*

$$\lambda_{\min}^E := \lambda_{\min}(\mathbf{H}_{EE}) > 0. \quad (4.31)$$

Then the sequence $\{\mathbf{x}^k\}$ generated by the fixed-point iterations (4.4) converges to $\mathbf{x}^* \in X^*$ q -linearly. Moreover, if τ is chosen as in (4.30) with $\lambda = \lambda_{\min}^E$, then the q_1 -factor satisfies

$$q_1 \leq \frac{\gamma(\lambda_{\min}^E) - 1}{\gamma(\lambda_{\min}^E) + 1}. \quad (4.32)$$

Proof. We follow the arguments in [67, Theorem 4.10]. First, without loss of generality, assume that all iteration counts k are large enough to ensure that $\mathbf{x}_j^k = \mathbf{x}_j^* = \mathbf{0}$ for all $j \in L$, and that the spectrum of $\mathbf{H}_{EE}$ lies in the interval $[\lambda_{\min}^E - \varepsilon, \lambda_{\max}(\mathbf{H})]$ where $\varepsilon > 0$ is arbitrary. Then

$$\begin{aligned} G_{\tau,E}(\mathbf{x}^k) - G_{\tau,E}(\mathbf{x}^*) &= \mathbf{x}_E^k - \mathbf{x}_E^* - \tau(\nabla\phi_{2,E}(\mathbf{x}^k) - \nabla\phi_{2,E}(\mathbf{x}^*)) \\ &= (I - \tau\mathbf{H}_{EE})(\mathbf{x}_E^k - \mathbf{x}_E^*). \end{aligned}$$

It follows

$$\begin{aligned} \|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{2,2} &\equiv \|\mathbf{x}_E^{k+1} - \mathbf{x}_E^*\|_{2,2} \leq \|G_{\tau,E}(\mathbf{x}^k) - G_{\tau,E}(\mathbf{x}^*)\|_{2,2} \\ &\leq \|I - \tau\mathbf{H}_{EE}\|_{2,2} \|\mathbf{x}_E^k - \mathbf{x}_E^*\|_{2,2} \\ &\leq \max \{ |1 - \tau\lambda_{\max}(\mathbf{H})|, |1 - \tau\lambda_{\min}^E| + \tau\varepsilon \} \|\mathbf{x}_E^k - \mathbf{x}_E^*\|_{2,2} \\ &\equiv \max \{ |1 - \tau\lambda_{\max}(\mathbf{H})|, |1 - \tau\lambda_{\min}^E| + \tau\varepsilon \} \|\mathbf{x}^k - \mathbf{x}^*\|_{2,2} \end{aligned} \quad (4.33)$$

For any $\tau \in (0, 2/\|\mathbf{H}\|_2)$ it is clear that $\max\{|1 - \tau\lambda_{\max}(\mathbf{H})|, |1 - \tau\lambda_{\min}^E| + \tau\varepsilon\} < 1$ when ε is sufficiently small, implying the linear convergence of the method. Moreover

$$\begin{aligned} q_1 &= \limsup_{k \rightarrow \infty} \frac{\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{2,2}}{\|\mathbf{x}^k - \mathbf{x}^*\|_{2,2}} \leq \max\{|1 - \tau\lambda_{\max}(\mathbf{H})|, |1 - \tau\lambda_{\min}^E| + \tau\varepsilon\} \\ &= \max\left\{\left|\frac{1 - \gamma(\lambda_{\min}^E)}{1 + \gamma(\lambda_{\min}^E)}\right|, \left|\frac{\gamma(\lambda_{\min}^E) - 1}{\gamma(\lambda_{\min}^E) + 1}\right| + \frac{2\varepsilon}{\lambda_{\max}(\mathbf{H}) + \lambda_{\min}^E}\right\} \end{aligned}$$

when $\tau = \tau(\lambda_{\min}^E)$. Since ε is arbitrary, (4.32) follows. ■

4.2 Bregman iterations

The iterative method of finding common points of convex sets in convex programming problems, referred to as *Bregman iterations*, first appeared in the 1967 work by Lev M. Bregman [16]. Since then, Bregman iterations have seen numerous applications in *total variation* (TV)-based image reconstruction and denoising problems [93], hyperspectral image enhancement and demixing problems [66, 110], nonlocal variational methods in image and data processing [130], and the compressive-sensing contexts of ℓ_1 -minimization [129]. Other works have detailed the impressive *error-forgetting* properties of Bregman iterations [128], and proposed linearizations [127] and splitting methods [63] which seek to make Bregman iterations more efficient in the above contexts.

In this section, we discuss an extension of Bregman iterations for solving the single-sparse vector recovery BPDN problem (3.60) from [129] to the Hilbert-valued recovery setting of Section 3.6. As mentioned in the introduction to Chapter 4, for the rest of this chapter we will focus only on the Hilbert-valued setting, noting that such developments can easily be applied to joint-sparse recovery of Section 4.1. We therefore, with an abuse of notation, redefine the Hilbert space $\mathcal{H}$ here (and for the remainder of this work) as $\mathcal{H} := \{\mathbf{x} \in \mathcal{V}^N : \|\mathbf{x}\|_{\mathcal{V},2} < \infty\}$. We recall discussion from [129, Section 3] in extending the Bregman iterative regularization algorithm to the considered setting and analyzing its convergence properties.

Given a convex functional $J : \mathcal{H} \rightarrow \mathbb{R}$, the Bregman distance based on J between points $\mathbf{u}$ and $\mathbf{v}$ in $\mathcal{H}$ is defined as

$$D_J^{\mathbf{p}}(\mathbf{u}, \mathbf{v}) = J(\mathbf{u}) - J(\mathbf{v}) - \langle \mathbf{p}, \mathbf{u} - \mathbf{v} \rangle_{\mathcal{V},2}. \quad (4.34)$$

Here $\mathbf{p} \in \partial J(\mathbf{v})$ is an arbitrary element of the subdifferential of J at the point $\mathbf{v}$ (see (4.2) for the definition of the subdifferential). When J is continuously differentiable on $\mathcal{H}$, for every $\mathbf{v} \in \mathcal{H}$ there exists a unique element $\mathbf{p} \in \partial J(\mathbf{v})$ making the Bregman distance (4.34) unique. In such a case, the Bregman distance is simply the difference at $\mathbf{u}$ between J and its first-order Taylor series approximation at the point $\mathbf{v}$. Furthermore, if J is strictly convex, then $D_J^{\mathbf{p}}(\mathbf{u}, \mathbf{v})$ is also strictly convex in $\mathbf{u}$ for each fixed $\mathbf{v}$, and hence $D_J^{\mathbf{p}}(\mathbf{u}, \mathbf{v}) = 0$ if and only if $\mathbf{u} = \mathbf{v}$.

However, for nonsmooth and not strictly convex functionals J , e.g., the BV seminorm considered in [19, 129] or the functional $J(\mathbf{z}) = \|\mathbf{z}\|_{\mathcal{V},1}$ from problem (3.70), the corresponding definition of the Bregman distance (4.34) may be multivalued or the empty

set depending on ∂J at a particular point. It is important to note that (4.34) does not induce a metric on $\mathcal{H}$, even when J is continuously differentiable and strictly convex, due to its lack of symmetry and failure to satisfy the triangle inequality. On the other hand, it does satisfy non-negativity, i.e., $D_J^p(\mathbf{u}, \mathbf{v}) \geq 0 \forall \mathbf{u}, \mathbf{v} \in \mathcal{H}$, and it measures distance in the sense that $D_J^p(\mathbf{u}, \mathbf{v}) \geq D_J^p(\mathbf{w}, \mathbf{v})$ for all points $\mathbf{w}$ on the line connecting $\mathbf{u}$ and $\mathbf{v}$ in $\mathcal{H}$, i.e., convex combinations $\mathbf{w} = \lambda \mathbf{u} + (1 - \lambda) \mathbf{v}$ for $\lambda \in [0, 1]$.

Bregman iterative regularization may be applied to solve minimization problems of the form

$$\text{minimize}_{\mathbf{z} \in \mathcal{H}} J(\mathbf{z}) + H(\mathbf{z}). \quad (4.35)$$

For this problem, the general form of the Bregman iterative regularization algorithm is given in Algorithm 1.

Algorithm 1 Bregman iterative regularization

Require: $J, H : \mathcal{H} \rightarrow \mathbb{R}$

- 1: Initialize: $k = 0, \mathbf{z}^0 = \mathbf{p}^0 = \mathbf{0} \in \mathcal{H}$
 - 2: **while** “not converged” **do**
 - 3: $\mathbf{z}^{k+1} \leftarrow \arg \min_{\mathbf{z} \in \mathcal{H}} D_J^{\mathbf{p}^k}(\mathbf{z}, \mathbf{z}^k) + H(\mathbf{z})$
 - 4: $\mathbf{p}^{k+1} \leftarrow \mathbf{p}^k - \nabla H(\mathbf{z}^{k+1}) \in \partial J(\mathbf{z}^{k+1})$
 - 5: $k \leftarrow k + 1$
 - 6: **end while**
-

Under additional assumptions on J and H it is possible to prove the following convergence results from [129] in the general setting of optimization over $\mathcal{H}$ through problem (4.35).

Assumption 4.15. *J is convex, H is convex and differentiable, and the solutions $\mathbf{z}^{k+1}$ in step 3 of Algorithm 1 exist.*

These assumptions are readily verified for the unconstrained minimization problem (3.70) with $J(\mathbf{z}) := \|\mathbf{z}\|_{\mathcal{V},1}$ and $H(\mathbf{z}) := \frac{\mu}{2} \|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2}^2$. Therefore, the following properties hold for Bregman iterations in that setting also.

Proposition 4.16. *Under Assumption 4.15, the iterate sequence $\{\mathbf{z}^k\}$ obtained by the Bregman iterative regularization of Algorithm 1 satisfies the following:*

1. *Monotonic decrease in H : $H(\mathbf{z}^{k+1}) \leq H(\mathbf{z}^{k+1}) + D_J^{\mathbf{p}^k}(\mathbf{z}^{k+1}, \mathbf{z}^k) \leq H(\mathbf{z}^k)$.*
2. *Convergence to the original in H with exact data: If $\tilde{\mathbf{z}}$ minimizes H and $J(\tilde{\mathbf{z}}) < \infty$, then $H(\mathbf{z}^k) \leq H(\tilde{\mathbf{z}}) + J(\tilde{\mathbf{z}})/k$.*

Moreover, in the case that $J(\mathbf{z}) := \|\mathbf{z}\|_{\mathcal{V},1}$ and $H(\mathbf{z}) := \frac{\mu}{2} \|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2}^2$, the sequence $\{\mathbf{z}^k\}$ has a weak-* convergent subsequence in the Banach space $(\mathcal{V}^N, \|\cdot\|_{\mathcal{V},1})$, and the limit of each weak-* convergent subsequence is a solution of $\mathbf{A}\mathbf{z} = \mathbf{u}$.

The proof of Proposition 4.16 trivially follows from that of [129, Proposition 3.2 & Theorem 3.3] and the connecting discussions of those results. Since the extension to the considered setting does not add significant insight to the analysis of Algorithm 1, it is included only for completeness in Appendix A.

4.2.1 Bregman iterations for Hilbert-valued recovery

In this section we discuss particular details of the application of Algorithm 1 to the setting of problem (3.70). As in Section 4.1, we assume here for simplicity of presentation that $\mu = 1$. In this setting, for $\mathbf{z} \in \mathcal{H}$ with

$$J(\mathbf{z}) := \|\mathbf{z}\|_{\mathcal{V},1}, \quad H(\mathbf{z}) := \frac{1}{2}\|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2}^2, \quad (4.36)$$

the Bregman iterative regularization procedure can be rewritten:

Version 1:

$$\mathbf{z}^0 \leftarrow \mathbf{0}, \mathbf{p}^0 \leftarrow 0, \quad (4.37)$$

For $k = 0, 1, \dots$ do

$$\mathbf{z}^{k+1} \leftarrow \arg \min_{\mathbf{z} \in \mathcal{V}^N} D_J^{\mathbf{p}^k}(\mathbf{z}, \mathbf{z}^k) + \frac{1}{2}\|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2}^2, \quad (4.38)$$

$$\mathbf{p}^{k+1} \leftarrow \mathbf{p}^k - \mathbf{A}^*(\mathbf{A}\mathbf{z}^{k+1} - \mathbf{u}); \quad (4.39)$$

Version 2:

$$\mathbf{u}^0 \leftarrow \mathbf{0}, \mathbf{z}^0 \leftarrow 0, \quad (4.40)$$

For $k = 0, 1, \dots$ do

$$\mathbf{u}^{k+1} \leftarrow \mathbf{u} + (\mathbf{u}^k - \mathbf{A}\mathbf{z}^k), \quad (4.41)$$

$$\mathbf{z}^{k+1} \leftarrow \arg \min_{\mathbf{z} \in \mathcal{V}^N} J(\mathbf{z}) + \frac{1}{2}\|\mathbf{A}\mathbf{z} - \mathbf{u}^{k+1}\|_{\mathcal{V},2}^2. \quad (4.42)$$

Here $D_J^{\mathbf{p}^{k+1}}(\mathbf{z}, \mathbf{z}^{k+1})$ is well defined since, given $\mathbf{z}^k$ and $\mathbf{p}^k$ from Version 1, $\mathbf{z}^{k+1}$ satisfies the first-order optimality condition:

$$\mathbf{0} \in \partial J(\mathbf{z}^{k+1}) - \mathbf{p}^k + \mathbf{A}^*(\mathbf{A}\mathbf{z}^{k+1} - \mathbf{u}), \quad (4.43)$$

which, from (4.39), implies $\mathbf{p}^{k+1} \in \partial J(\mathbf{z}^{k+1})$. The equivalence of Versions 1 & 2 of the Bregman iterative procedure was shown in [129, Theorem 3.1]. We include the result here for completeness, noting that the proof in Appendix A is essentially unchanged.

Theorem 4.17. *The Bregman iterative procedure Version 1 (4.37)-(4.39) and Version 2 (4.40)-(4.42) are equivalent in the sense that (4.38) and (4.42) have the same objective functions (up to a constant) for all k . Moreover, if $\mathbf{z}^k$ and $\bar{\mathbf{z}}^k$ are the solutions to Version 1 and 2, respectively, then $\mathbf{A}^*(\mathbf{A}\mathbf{z}^k - \mathbf{u}) = \mathbf{A}^*(\mathbf{A}\bar{\mathbf{z}}^k - \mathbf{u})$ and $\mathbf{p}^k = \mathbf{A}^*(\mathbf{u}^k - \mathbf{A}\bar{\mathbf{z}}^k)$.*

In [129, Section 3.4] it was noted that the Bregman iterative method is equivalent to the well-known augmented Lagrangian method (also referred to as the method of multipliers) under a certain initial parameterization. They also show that if, at the k -th iteration, $\mathbf{z}^k$ satisfies the linear constraints $\mathbf{A}\mathbf{z}^k = \mathbf{u}$, then it also minimizes $J(\cdot) = \|\cdot\|_1$, making it a solution of the basis pursuit problem, i.e., (3.60) with $\tilde{\eta} = 0$. The same result holds in the

considered setting of problem (3.70). Therefore, we include it here for completeness with proof reserved for Appendix A.

Theorem 4.18. *Suppose an iterate $\mathbf{z}^k$ from (4.38) satisfies $\mathbf{A}\mathbf{z}^k = \mathbf{u}$; then $\mathbf{z}^k$ is a solution of the HV basis pursuit problem:*

$$\text{minimize}_{\mathbf{z} \in \mathcal{V}^N} \|\mathbf{z}\|_{\mathcal{V},1} \quad \text{subject to} \quad \mathbf{A}\mathbf{z} = \mathbf{u}. \quad (4.44)$$

Finally, we are able to show the following result on the finite convergence properties of the Bregman iterations.

Theorem 4.19. *Let $\{\mathbf{z}^k\}$ be the sequence defined by the iterations (4.38). Then $\mathbf{z}^k \in U^k$, for which $\min\{\frac{1}{2}\|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2}^2 : \mathbf{z} \in U^k\} = 0$, for all but finitely many iterations.*

Proof. Let $\mathbf{v} \in \mathcal{V}$ be such that $\mathbf{v} \neq \mathbf{0}$, and define $\mathcal{F}(\cdot) := \langle \cdot, \mathbf{v} \rangle_{\mathcal{V}}$. Then if (I_+^j, I_-^j, E^j) is an arbitrary partition of $\mathcal{J}$, we define

$$\begin{aligned} U^j &:= U(I_+^j, I_-^j, E^j) \\ &= \{\mathbf{z} \in \mathcal{V}^N : \mathcal{F}(\mathbf{z}_{\nu}) \geq 0, i \in I_+^j; \mathcal{F}(\mathbf{z}_{\nu}) \leq 0, i \in I_-^j; \mathcal{F}(\mathbf{z}_{\nu}) = 0, i \in E^j\}, \\ H^j &:= \min \left\{ \frac{1}{2} \|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2}^2 : \mathbf{z} \in U^j \right\}. \end{aligned}$$

There are a finite number of distinct partitions (I_+^j, I_-^j, E^j) , and the union of all possible sets U^j is $\mathcal{V}^N$. At iteration k , let (I_+^k, I_-^k, E^k) be defined as follows:

$$\begin{aligned} I_+^k &= \{\nu \in \mathcal{J} : \mathcal{F}(\mathbf{p}_{\nu}^k) \geq 0, \|\mathbf{p}_{\nu}^k\|_{\mathcal{V}} = 1\}, & I_-^k &= \{\nu \in \mathcal{J} : \mathcal{F}(\mathbf{p}_{\nu}^k) \leq 0, \|\mathbf{p}_{\nu}^k\|_{\mathcal{V}} = 1\}, \\ E^k &= \{\nu \in \mathcal{J} : \|\mathbf{p}_{\nu}^k\|_{\mathcal{V}} < 1\}. \end{aligned}$$

If $\nu \in I_+^k$, then since $\mathbf{p}_{\nu}^k \in \partial(\|\mathbf{z}^k\|_{\mathcal{V},1})$

$$0 \leq \mathcal{F}(\mathbf{p}_{\nu}^k) = \langle \mathbf{p}_{\nu}^k, \mathbf{v} \rangle_{\mathcal{V}} = \left\langle \frac{\mathbf{z}_{\nu}^k}{\|\mathbf{z}_{\nu}^k\|_{\mathcal{V}}}, \mathbf{v} \right\rangle_{\mathcal{V}} = \frac{\langle \mathbf{z}_{\nu}^k, \mathbf{v} \rangle_{\mathcal{V}}}{\|\mathbf{z}_{\nu}^k\|_{\mathcal{V}}}$$

implying $\mathcal{F}(\mathbf{z}_{\nu}^k) = \langle \mathbf{z}_{\nu}^k, \mathbf{v} \rangle_{\mathcal{V}} \geq 0$ since $1/\|\mathbf{z}_{\nu}^k\|_{\mathcal{V}} > 0$. Similarly, if $\nu \in I_-^k$, then

$$0 \geq \mathcal{F}(\mathbf{p}_{\nu}^k) = \langle \mathbf{p}_{\nu}^k, \mathbf{v} \rangle_{\mathcal{V}} = \left\langle \frac{\mathbf{z}_{\nu}^k}{\|\mathbf{z}_{\nu}^k\|_{\mathcal{V}}}, \mathbf{v} \right\rangle_{\mathcal{V}} = \frac{\langle \mathbf{z}_{\nu}^k, \mathbf{v} \rangle_{\mathcal{V}}}{\|\mathbf{z}_{\nu}^k\|_{\mathcal{V}}}$$

implying $\mathcal{F}(\mathbf{z}_{\nu}^k) = \langle \mathbf{z}_{\nu}^k, \mathbf{v} \rangle_{\mathcal{V}} \leq 0$. Also, if $\nu \in E^k$, then $\|\mathbf{p}_{\nu}^k\|_{\mathcal{V}} < 1$ so that $\mathbf{p}_{\nu}^k \in B_{\mathcal{V}}(\mathbf{0}, 1)$, implying $\mathbf{z}_{\nu}^k = \mathbf{0}$ and

$$\mathcal{F}(\mathbf{z}_{\nu}^k) = \langle \mathbf{z}_{\nu}^k, \mathbf{v} \rangle_{\mathcal{V}} = 0.$$

Hence $\mathbf{z}^k \in U^k$ for every k when U^k is defined in terms of $\mathbf{p}^k$ and (I_+^k, I_-^k, E^k) as above. We apply Proposition 4.16, with $\tilde{\mathbf{z}}$ satisfying $H(\tilde{\mathbf{z}}) = \frac{1}{2}\|\mathbf{A}\tilde{\mathbf{z}} - \mathbf{u}\|_{\mathcal{V},2}^2 = 0$. With this choice of $\tilde{\mathbf{z}}$, in statement 2 of Proposition 4.16, we see that for each j with $H^j > 0$ there is a sufficiently

large K_j such that $H(\mathbf{z}^k) < H^j$ if $k \geq K_j$, implying $\mathbf{z}^k$ is not in U^j for $k \geq K_j$. Therefore, letting $K := \max_j \{K_j : H^j > 0\}$, we have that $\mathbf{z}^k$ is not in U^j for any j such that $H^j > 0$ whenever $k \geq K$. Hence, if $k \geq K$, then $\mathbf{z}^k$ must belong to one of the sets U^k for which $H^k = 0$, which was to be shown. ■

Remark 4.20 (Conclusions drawn from Theorem 4.19). *We remark that the conclusion of the previous theorem does not imply that the iterations $\{\mathbf{z}^k\}$ obtained from (4.38) converge to a minimizer after a finite number of iterations. In [129, Theorem 5.1] a set of four conditions is presented which imply the finite convergence of Algorithm 1 in the setting of the infinite-dimensional optimization problem (4.35). The “missing ingredient” to show finite convergence with that result is condition 4: “there exists a rule to associate each $\mathbf{z}^k$ with a $U^{j_k} \ni \mathbf{z}^k$ so that if $H^{j_k} = 0$, then $D_J^{\mathbf{p}^k}(\mathbf{z}, \mathbf{z}^k) = 0 \iff \mathbf{z} \in U^{j_k}$.” While the above proof provides a rule to guarantee $\mathbf{z}^k \in U^k$ when U^k is defined in terms of $\mathbf{p}^k$ and (I_+^k, I_-^k, E^k) as above, a path for showing $D_J^{\mathbf{p}^k}(\mathbf{z}, \mathbf{z}^k) = 0 \iff \mathbf{z} \in U^{j_k}$ whenever $H^{j_k} = \min\{\frac{1}{2}\|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2}^2 : \mathbf{z} \in U^{j_k}\} = 0$ is unclear. Nonetheless, the above result can be interpreted in the following way. After a finite number of iterations, the iterates $\{\mathbf{z}^k\}$ are confined to a union of sets $\bar{U}$ for which $\min\{\frac{1}{2}\|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2}^2 : \mathbf{z} \in \bar{U}\} = 0$.*

4.3 Fixed-point continuation for improving forward-backward splitting algorithm performance

In this section, we discuss the *fixed-point continuation* (FPC) strategy from [67, 68], showing how it can be applied to improve the performance of the Hilbert-valued formulation of the forward-backward iterations from Remark 4.3, i.e., in the context of the sparse HV recovery problems of Section 3.6. We also detail the combination of FPC with the Bregman iterative regularization presented in Section 4.2, discussing issues related to parameterizing both methods to yield accurate approximations to problem (3.70). The combined FPC-Bregman iterations enjoy improved robustness to noise and decreased overall runtime compared to standard forward-backward iterations. We describe our approach as follows.

In Section 4.3.1, we give a description of the FPC strategy, which seeks to reduce the total number of forward-backward iterations $\mathbf{x}^{k+1} = J_\tau \circ G_\tau(\mathbf{x}^k)$ from (4.11) required to solve problem (3.70) to a prescribed tolerance. It achieves this through a continuation strategy on the step-size and residual weight parameters τ and μ . More specifically, FPC solves a sequence of subproblems of the form (3.70), associated with a sequence of relaxed data-fidelity weight parameters $\{\mu^\ell\}$ increasing towards a user-supplied desired final residual weight $\bar{\mu}$. The solution of each subproblem is then provided as initial guess for the next subproblem with $\mu_{\ell+1} = \min\{\beta^\ell \mu_0, \bar{\mu}\}$ for suitable $\mu_0, \beta > 0$, stopping once $\mu_\ell \geq \bar{\mu}$ for some $\ell \in \mathbb{N}$. The FPC step size continuation strategy applies only to the soft-thresholding step J_τ , replacing τ in (4.10) with $\nu_\ell = \tau/\mu_\ell$, thereby taking larger steps initially, decreasing in size as $\mu_\ell \rightarrow \bar{\mu}$. This final value of $\bar{\mu}$ is, of course, vital in parameterizing (3.70) to yield solutions to the HV-BPDN problem (3.69) obeying the residual constraint.

In Section 4.3.2 we discuss combining FPC with Bregman iterations, following guidelines from [68, 129] in parameterizing both solvers. We note, in light of version 2 (4.40)-(4.42) of Algorithm 1, that the combined Bregman-FPC iterations also solve a sequence

of subproblems in step (4.42) with updated data term $\mathbf{u}^k = \mathbf{u} + (\mathbf{u}^{k-1} - \mathbf{A}\mathbf{z}^{k-1})$ at the k -th Bregman iteration. This update can be interpreted as adding in the residual error $\mathbf{u} - \mathbf{A}\mathbf{z}^{k-1}$ at each iteration, and gives the Bregman iterative procedure favorable “error forgetting” properties, see [128]. When combined with certain solvers, including forward-backward iterations with FPC, the method also enjoys a non-accumulating “error cancellation” property. This property allows the FPC subproblems to be solved to a lesser accuracy of $\mathcal{O}(10^{-6})$, without affecting the overall error of approximation obtained with the combined solver. The resulting FPC-Bregman algorithm possesses superior convergence properties over standard forward-backward iterations with FPC, enabling efficient solution of (3.69) and hence sparse numerical approximations of solutions to parameterized PDEs.

4.3.1 Fixed point continuation

We recall the discussion of [68, Section 3.2] and the convergence results for forward-backward splitting in the finite-dimensional case of $\mathbb{R}^N$ from [67] in deriving the fixed-point continuation strategy for sparse HV recovery. In both works, a modification of the standard forward-backward iterations (4.11) was proposed, replacing the step size parameter τ in (4.10) with $\nu = \tau/\mu$. The analysis presented in those works suggests that larger values of ν (smaller μ) may improve observed convergence rates, with bounds on the number of iterations for convergence inversely proportional to ν^2 . This suggests to use a strategy of continuation on μ , choosing smaller initial values corresponding to larger ν , and increasing μ towards a final value of $\bar{\mu}$. We develop this strategy as follows.

With an abuse of notation, we define

$$X^* = X^*(\mu) := \left\{ \mathbf{x} \in \mathcal{H} : \mathbf{0} \in \partial \|\mathbf{x}\|_{\mathcal{V},1} + \frac{\mu}{2} \nabla \|\mathbf{A}\mathbf{x} - \mathbf{u}\|_{\mathcal{V},2}^2 \right\}, \quad (4.45)$$

the set of solutions to (3.70) for a particular $\mu > 0$. We begin by noting $\mathbf{0} \in X^*$ when

$$\mu \leq \frac{1}{\|\mathbf{A}^*(\mathbf{A}\mathbf{0} - \mathbf{u})\|_{\mathcal{V},\infty}} = \frac{1}{\|\mathbf{A}^*\mathbf{u}\|_{\mathcal{V},\infty}}, \quad (4.46)$$

since

$$\mathbf{x} \in X^* \iff \mu(\mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{u}))_{\nu} \in \begin{cases} \{-\mathbf{x}_{\nu}/\|\mathbf{x}_{\nu}\|_{\mathcal{V}}\} & \mathbf{x}_{\nu} \neq \mathbf{0}, \\ B_{\mathcal{V}}(\mathbf{0}, 1) & \mathbf{x}_{\nu} = \mathbf{0}, \end{cases} \quad \forall \nu \in \mathcal{J}, \quad (4.47)$$

implying $\mathbf{0} \in X^* \iff \mu\|\mathbf{A}^*\mathbf{u}\|_{\mathcal{V},\infty} \leq 1$. One can also interpret from (4.47) that smaller values of μ in (3.70) lead to more sparse solutions, while larger values result in less sparse solutions. As noted in the discussion preceding Lemma 4.8 in Section 4.1, the sets L and E from (4.18) are invariant on the solution set for the joint-sparse minimization problem (3.76). The same can be said for their reformulated versions in the context of problem (3.70) with $\mu > 0$ arbitrary and X^* as above.

Definition 4.21. With H defined by $H(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{u}\|_{\mathcal{V},2}^2$ for $\mathbf{x} \in \mathcal{H}$ as in (4.36), for $\mathbf{x}^* \in X^*$ and $\boldsymbol{\alpha} \in \mathcal{J}$, let $\nabla H_{\boldsymbol{\alpha}}(\mathbf{x}^*) := (\nabla H(\mathbf{x}^*))_{\boldsymbol{\alpha}} = (\mathbf{A}^*(\mathbf{A}\mathbf{x}^* - \mathbf{u}))_{\boldsymbol{\alpha}}$ and

$$L := \{\boldsymbol{\alpha} \in \mathcal{J} : \mu\|\nabla H_{\boldsymbol{\alpha}}(\mathbf{x}^*)\|_{\mathcal{V}} < 1\}, \quad E := \{\boldsymbol{\alpha} \in \mathcal{J} : \mu\|\nabla H_{\boldsymbol{\alpha}}(\mathbf{x}^*)\|_{\mathcal{V}} = 1\}, \quad (4.48)$$

$$\omega := \min\{\nu(1 - \mu\|\nabla H_{\boldsymbol{\alpha}}(\mathbf{x}^*)\|_{\mathcal{V}}) : \boldsymbol{\alpha} \in L\}. \quad (4.49)$$

For L and E from (4.48), equivalent statements to (4.20) can be made, i.e., it is trivial to show that $L \cup E = \mathcal{J}$, $\text{supp}(\mathbf{x}^*) \subseteq E$, and $L \subset (\text{supp}(\mathbf{x}^*))^c \forall \mathbf{x}^* \in X^*$. Moreover, for $\mathbf{x}^* \in X^*$ and $\boldsymbol{\alpha} \in L$ arbitrary

$$\nu(1 - \mu\|\nabla H_{\boldsymbol{\alpha}}(\mathbf{x}^*)\|_{\mathcal{V}}) = \nu - \tau\|\nabla H_{\boldsymbol{\alpha}}(\mathbf{x}^*)\|_{\mathcal{V}} = \nu - \|\mathbf{x}_{\boldsymbol{\alpha}}^* - \tau\nabla H_{\boldsymbol{\alpha}}(\mathbf{x}^*)\|_{\mathcal{V}} = \nu - \|G_{\tau,\boldsymbol{\alpha}}(\mathbf{x}^*)\|_{\mathcal{V}},$$

since $\nu\mu = \tau$, implying

$$\min\{\nu - \|G_{\tau,\boldsymbol{\alpha}}(\mathbf{x}^*)\|_{\mathcal{V}} : \boldsymbol{\alpha} \in L\} = \omega > 0, \quad \forall \mathbf{x}^* \in X^*. \quad (4.50)$$

By examining the linear convergence results of [67, Proposition 4.9 & Theorem 4.10] and its joint-sparse recovery version in Theorem 4.14 from Section 4.1, we see that the q_1 -factor improves as the cardinality of the set E , and hence number of nonzero elements of the solution $\mathbf{x}^* \in X^*$, decreases. Indeed, adapting the (worst-case) analysis of Theorem 4.14 to the sparse HV recovery setting, one can show that the convergence rate is controlled by the quantity $\|\mathbf{I} - \tau\mathbf{H}_{EE}\|_2$, with $\mathbf{H} = \mathbf{A}^*\mathbf{A}$ the Hessian of H above, which can be much smaller than $\|\mathbf{I} - \tau\mathbf{H}\|_2$ when $\#(E) \ll N$. Moreover, both [67, Theorem 4.5] and Lemma 4.10 imply that the number of iterations $k \in \mathbb{N}$ for which $\mathbf{x}_{\boldsymbol{\alpha}}^k \neq \mathbf{x}_{\boldsymbol{\alpha}}^* = \mathbf{0}$ for $\boldsymbol{\alpha} \in L$ is inversely proportional to ω^2 , which, from (4.50), satisfies $0 < \omega^2 \leq \nu^2$. In general, ω depends on $\max\{\|G_{\tau,\boldsymbol{\alpha}}(\mathbf{x}^*)\|_{\mathcal{V}} : \boldsymbol{\alpha} \in L\}$, which depends on τ , μ , $\mathbf{u}$, and $\mathbf{A}$, and is therefore problem-dependent. Nonetheless, as recognized in [68, Section 3.2], the above facts suggest a link between smaller values of μ (hence larger ν) and faster convergence of the iterates $\{\mathbf{x}^k\}$ towards an element $\mathbf{x}^* \in X^*$.

Therefore, we consider the solution of a finite sequence of subproblems of the form (3.70), depending on finite sequences of residual weightes $\{\mu_{\ell}\}$ satisfying $\mu_0 < \mu_1 < \dots \leq \bar{\mu}$ and step sizes $\{\nu_{\ell}\}$ satisfying $\nu_0 > \nu_1 > \dots \geq \bar{\nu}$, with the goal of reducing the total number of forward-backward iterations (4.11). The modified iterations

$$\mathbf{x}^{k+1} := J_{\nu_{\ell}} \circ G_{\tau}(\mathbf{x}^k), \quad (4.51)$$

with $J_{\nu_{\ell}}$ as in (4.10) (τ replaced with $\nu_{\ell} = \tau/\mu_{\ell}$) and G_{τ} as in (4.9), are applied to solve

$$\text{minimize}_{\mathbf{z} \in \mathcal{V}^N} \|\mathbf{z}\|_{\mathcal{V},1} + \frac{\mu_{\ell}}{2}\|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2}^2, \quad (4.52)$$

i.e., the ℓ -th subproblem, quickly identifying the support of an element $\mathbf{x}^{\ell,*} = \mathbf{x}^*(\mu_{\ell}) \in X^*(\mu_{\ell})$ as a result of Lemma 4.10. We then provide the approximate solution of the ℓ -th subproblem as initial guess for the $(\ell + 1)$ -th subproblem, continuing until the final value of $\bar{\mu}$ has been reached.

The FPC iterations begin from an initial guess of $\mathbf{x}^0 = \tau\mathbf{A}^*\mathbf{u}$, which provides problem-specific information and is easy to compute. As in [68], given a constant $\beta > 1$, we apply a

growth rule of

$$\mu_{\ell+1} = \min\{\beta^\ell \mu_0, \bar{\mu}\}, \quad (4.53)$$

increasing up to the final value of $\mu_L = \bar{\mu}$ for some $L \in \mathbb{N}$. The iterations terminate after an approximate solution is found to the L -th subproblem (4.52). We note that the choice of $\beta = 4$ proposed in [68, Section 4.2.1] works well in practice. We also follow the proposed initialization of μ_0 and ν_0 from [69, Section 7.2.1], setting

$$\mu_0 = \frac{\tau}{\gamma \|\mathbf{x}^0\|_{\nu, \infty}} \quad \nu_0 = \tau / \mu_0, \quad (4.54)$$

where $0 < \gamma < 1$ is a user-supplied constant which ensures the soft-thresholding operator J_{ν_0} reduces most components of G_τ to $\mathbf{0}$, leaving only those larger than $\gamma \|\mathbf{x}^0\|_{\nu, \infty}$ to refine. Our numerical experiments agree with the conclusions from [69] that γ and β should be adjusted to keep the total number of parameter updates approximately constant, therefore we use the proposed choice of

$$\gamma = 0.99, \quad \beta = 4. \quad (4.55)$$

Recalling Assumption 4.1, for the convergence of the forward-backward iterations (4.51) in the setting of sparse HV recovery we require $\tau \in (0, 2/\hat{\lambda}_{\max})$, where $\hat{\lambda}_{\max} = \lambda_{\max}(\mathbf{H})$. For fixed τ it was recommended in [69] to choose $\tau \in [1/\hat{\lambda}_{\max}, 2/\hat{\lambda}_{\max})$ over $\tau \in (0, 1/\hat{\lambda}_{\max})$ for improved performance. This can be simplified to choosing $\tau \in [1, 2)$ when $\mathbf{A}$ and $\mathbf{u}$ from (3.68) are rescaled to

$$\mathbf{A} \leftarrow \frac{\mathbf{A}}{\sqrt{\hat{\lambda}_{\max}}} \quad \mathbf{u} \leftarrow \frac{\mathbf{u}}{\sqrt{\hat{\lambda}_{\max}}}. \quad (4.56)$$

Under this rescaling, the authors of [69] recommend setting

$$\tau = \min\{1 + 1.665(1 - m/N), 1.999\}, \quad (4.57)$$

which is constant and equal to 1.999 for $m/N \leq 0.4$, decreasing to 1 as $m/N \rightarrow 1$. This choice works well in the context of unstructured random matrices, such as the Gaussian matrices considered in that work. However for smaller values of m/N , e.g., $1/8$ or $1/4$, our experiments with the structured random matrices considered in (3.68) defined in terms of the orthonormal Legendre system $(\Psi_\alpha)_{\alpha \in \mathcal{J}}$ exhibit poor convergence under (4.57). Therefore, we elect to use the simple and conservative choice of

$$\tau = 1 \quad (4.58)$$

when $\mathbf{A}$ and $\mathbf{u}$ are rescaled as in (4.56). However, we find that for larger values of m/N , the choice of τ given in (4.57) works well. We leave a more detailed study on optimal step size parameterization strategies, e.g., the application of BB steps and line search methods from [68, Section 3.3] to sparse HV recovery, for a future work.

The residual weight parameter μ_ℓ and step size parameter ν_ℓ are updated once the k -th iteration satisfies

$$\frac{\|\mathbf{x}^k - \mathbf{x}^{k-1}\|_{\mathcal{V},2}}{\max\{\|\mathbf{x}^{k-1}\|_{\mathcal{V},2}, 1\}} < \sqrt{\frac{\bar{\mu}}{\mu_\ell}} x_{\text{tol}} \quad \text{and} \quad \mu_\ell \|\mathbf{A}^*(\mathbf{A}\mathbf{x}^k - \mathbf{u})\|_{\mathcal{V},\infty} - 1 < g_{\text{tol}}, \quad (4.59)$$

where $x_{\text{tol}}, g_{\text{tol}} > 0$ are user-supplied tolerance constants. As noted in [68], the first condition in (4.59) ensures the last step is small relative to the previous iteration, while the second condition checks for complementarity in (4.47) at the current iteration. The authors further note that the second condition greatly improves accuracy, however g_{tol} should be large to ensure faster convergence, recommending the parameterization $x_{\text{tol}} = 10^{-4}$ and $g_{\text{tol}} = 0.2$. Through extensive numerical testing, our results suggest that x_{tol} is not as important when using Bregman iterations combined with FPC for sparse HV recovery, and we observe that updating the parameters μ_ℓ and ν_ℓ more often can improve convergence. In practice, we find that the choice of

$$x_{\text{tol}} = 1, \quad g_{\text{tol}} = 0.1, \quad (4.60)$$

works well for the problems considered in Section 6.3, while smaller values of x_{tol} require more forward-backward iterations, but do not improve overall errors. An explanation of this phenomenon may be found in the “error-forgetting” and “error-cancellation” properties of the Bregman algorithm when combined with FPC [128], though further testing is needed to verify our observations. We leave a more detailed investigation on best practices in parameterizing the combined solver for sparse HV recovery problems to a future work.

4.3.2 Combined Bregman-FPC algorithm

In this section we describe the combination of Bregman iterative regularization from Section 4.2 with FPC in the context of sparse HV recovery. The combined Bregman-FPC algorithm involves the solution of a sequence of subproblems of the form (3.70). In essence, the outer Bregman iterations supply an updated data term $\mathbf{u}^k$ to the residual, as in (4.42), while the inner FPC iterations supply an updated data fidelity weight μ_ℓ on the residual, as in (4.52), updating the step size ν_ℓ in the forward-backward iterations (4.51) accordingly. The iterations continue until

$$\|\mathbf{A}\mathbf{z}^k - \mathbf{u}\|_{\mathcal{V},2} < b_{\text{tol}}, \quad (4.61)$$

for some $k \in \mathbb{N}$ and $b_{\text{tol}} > 0$. In practice we want the output of the Bregman-FPC algorithm to satisfy the residual constraint of problem (3.69). Therefore, we choose b_{tol} proportional to $\eta/\sqrt{m}$, i.e., taking

$$b_{\text{tol}} = 1.2 \cdot \|\mathbf{A}\mathbf{c}^* - \mathbf{u}\|_{\mathcal{V},2}, \quad (4.62)$$

with $\mathbf{c}^* = (\mathbf{c}_\nu^*)_{\nu \in \mathcal{J}}$ the coefficients of the Galerkin projection of the solution u of problem (2.2) onto $\mathcal{V}^N$ from (3.6) as in Section 3.2.1. This choice is sufficient to approximately satisfy the residual constraint of (3.69) up to a small multiplicative constant. In actual applications,

one can take b_{tol} proportional to known a-priori error estimates for the Galerkin projection. A more detailed discussion on such results can be found in Section 5.3.

The combined Bregman-FPC algorithm is given in Algorithm 2. Line 8 checks to see if $\mathbf{0}$ is a solution of (3.70) when $\mu = \bar{\mu}$ and $\mathbf{u} = \mathbf{u}^k$, see (4.47). In which case lines 9-10 skip solving the k -th FPC subproblem and return $\mathbf{0}$. Line 16 checks if the current FPC iterate satisfies the conditions of (4.59) for updating μ_ℓ and ν_ℓ and Line 17 checks for convergence of the k -th FPC solve. Finally, line 4 checks for convergence of the Bregman iterations for solving (3.70) with $\bar{\mu}$ and $\mathbf{u}$.

Algorithm 2 Bregman-FPC Algorithm

Input: $\mathbf{A}$, $\mathbf{u}$, $\bar{\mu}$, τ , γ , β , x_{tol} , g_{tol} , b_{tol} , $I_{\text{Max-FPC}}$, $I_{\text{Max-Bregman}}$

Output: $\mathbf{z}^k$, the approximate solution to (3.70) with $\bar{\mu}$

```
1: Compute:
    $\hat{\lambda}_{\max} \leftarrow \lambda_{\max}(\mathbf{A}\mathbf{A}^*), \hat{\lambda}_{\min} \leftarrow \lambda_{\min}(\mathbf{A}\mathbf{A}^*)$ 
2: Rescale according to  $\hat{\lambda}_{\max}$ :
    $\mathbf{A} \leftarrow \mathbf{A}/\sqrt{\hat{\lambda}_{\max}}, \mathbf{u} \leftarrow \mathbf{u}/\sqrt{\hat{\lambda}_{\max}}$ 
3: Initialize parameters for Bregman iterations:
    $k \leftarrow 0, \mathbf{u}^0 \leftarrow \mathbf{0}, \mathbf{z}^0 \leftarrow \mathbf{0}$ 
4: while  $k < I_{\text{Max-Bregman}}$  and  $\|\mathbf{A}\mathbf{z}^k - \mathbf{u}\|_{\mathcal{V},2} > b_{\text{tol}}$  do
5:    $k \leftarrow k + 1$ 
6:    $\mathbf{u}^k \leftarrow \mathbf{u} + (\mathbf{u}^{k-1} - \mathbf{A}\mathbf{z}^{k-1})$ ,
7:   Initialize parameters for the  $k$ -th FPC subproblem:
      $j \leftarrow 0, \mathbf{x}^0 \leftarrow \tau \mathbf{A}^* \mathbf{u}^k, \ell \leftarrow 0, \mu_0 \leftarrow \frac{1}{\gamma \|\mathbf{x}^0\|_{\mathcal{V},\infty}}, \nu_0 \leftarrow \tau/\mu_0$ ,
8:   if  $\bar{\mu} \|\mathbf{A}^* \mathbf{u}^k\|_{\mathcal{V},\infty} \leq 1$  then
9:      $\mathbf{z}^k \leftarrow \mathbf{0}$ 
10:   end while
11:   else
12:     while  $j < I_{\text{Max-FPC}}$  and  $\mu_\ell \leq \bar{\mu}$  do
13:        $j \leftarrow j + 1$ 
14:        $\mathbf{v}^j \leftarrow \mathbf{x}^{j-1} - \tau \mathbf{A}^*(\mathbf{A}\mathbf{x}^{j-1} - \mathbf{u}^k)$ 
15:       for each  $\nu \in \mathcal{J}$ :  $\mathbf{x}_\nu^j \leftarrow \frac{\mathbf{v}_\nu^j}{\|\mathbf{v}_\nu^j\|_{\mathcal{V}}} \cdot \max\{\|\mathbf{v}_\nu^j\|_{\mathcal{V}} - \nu_\ell, 0\}$ 
16:       if  $\frac{\|\mathbf{x}^j - \mathbf{x}^{j-1}\|_{\mathcal{V},2}}{\max\{\|\mathbf{x}^{j-1}\|_{\mathcal{V},2}, 1\}} < \sqrt{\frac{\bar{\mu}}{\mu_\ell}} x_{\text{tol}}$  and  $\mu_\ell \|\mathbf{A}^*(\mathbf{A}\mathbf{x}^j - \mathbf{u}^k)\|_{\mathcal{V},\infty} - 1 < g_{\text{tol}}$  then
17:         if  $\mu_\ell = \bar{\mu}$  then
18:            $\mathbf{z}^k \leftarrow \mathbf{x}^j$ 
19:         end while
20:       else
21:          $\ell \leftarrow \ell + 1, \mu_\ell \leftarrow \min\{\mu_0 \beta^\ell, \bar{\mu}\}, \nu_\ell \leftarrow \tau/\mu_\ell$ 
22:       end if
23:     end if
24:   end while
25: end if
26: end while
```

Choice of residual weight parameter

We conclude this chapter by examining the choice of parameter $\bar{\mu}$ in solving problem (3.70). As previously noted, this parameter is key to ensuring the residual satisfies the constraint of problem (3.69). However, the running time of Algorithm 2 in solving the FPC subproblems is also heavily dependent on the choice of $\bar{\mu}$. The works [68, 129] recommend various strategies for setting this parameter, under differing assumptions on the underlying measurement system. Neither of these settings focus on applications to compressed-sensing-based polynomial approximation methods, which of course is the setting relevant to our sparse Hilbert-valued vector recovery techniques of Section 3.6. Therefore we present both parameterization strategies, and their necessary modifications to obtain the numerical results of Section 6.3

In [68], the forward-backward iterations are presented in the context of solving under-determined linear systems $\mathbf{A}\mathbf{c} = \mathbf{u}$ using the ℓ_1 -regularized weighted least-squares problem formulation

$$\text{minimize}_{\mathbf{z} \in \mathbb{R}^N} \|\mathbf{z}\|_1 + \frac{\bar{\mu}}{2} \|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathbf{M}}^2, \quad (4.63)$$

for $\mathbf{A} \in \mathbb{R}^{m \times N}$, $m \ll N$, and $\|\mathbf{p}\|_{\mathbf{M}}^2 = \mathbf{p}^* \mathbf{M} \mathbf{p}$ for $\mathbf{M} \in \mathbb{R}^{m \times m}$ positive definite. This setting is relevant to a number of measurement problems, including that of compressed sensing in the presence of Gaussian noise. More specifically, under the assumption that the measurement model satisfies

$$\mathbf{u} = \mathbf{A}(\mathbf{c} + \mathbf{e}_1) + \mathbf{e}_2, \quad (4.64)$$

for $\mathbf{e}_j \in \mathbb{R}^N$ following distribution $\mathcal{N}(0, \sigma_j^2)$, i.e., $\mathbf{e}_j$ a mean zero Gaussian random vector with variance $\sigma_j^2 > 0$ for $j = 1, 2$, conditions on $\mathbf{M}$ and μ are derived in [68, Section 3.1] to optimally weight problem (4.63) in search of s -sparse vectors $\mathbf{c} \in \mathbb{R}^N$ consistent with (4.64). Setting

$$\mathbf{M} = (\sigma_1^2 \mathbf{A}\mathbf{A}^* + \sigma_2^2 \mathbf{I})^{-1} \quad \bar{\mu} = \frac{1}{\underline{\sigma}} \sqrt{\frac{N}{\chi_{1-\alpha, m}^2}}, \quad (4.65)$$

where $\underline{\sigma}^2 := \lambda_{\min}(\mathbf{M}^{1/2} \mathbf{A}\mathbf{A}^* \mathbf{M}^{1/2})$ and $\chi_{1-\alpha, m}^2$ is the $1-\alpha$ critical value of the χ^2 distribution with m degrees of freedom, yields the desired result.

On the other hand, the work [129] on Bregman iterations in the context of ℓ_1 -minimization and compressed sensing considered a slight modification to (4.63), weighting the ℓ_1 part of the objective function as

$$\text{minimize}_{\mathbf{z} \in \mathbb{R}^N} \tilde{\mu} \|\mathbf{z}\|_1 + \frac{1}{2} \|\mathbf{A}\mathbf{z} - \mathbf{u}\|_2^2. \quad (4.66)$$

When $\mathbf{M} = \mathbf{I}$, both problems (4.63) & (4.66) are equivalent after the change of parameter $\bar{\mu} \longleftrightarrow 1/\tilde{\mu}$. Noting that the total number of iterations depend heavily on $\tilde{\mu}$, with $\mathbf{c}$ is the

s -sparse solution of (4.66), they elect for a moderate value of

$$\tilde{\mu} = \xi / \sqrt{\|\mathbf{c}\|_0}, \quad (4.67)$$

for $\xi = 0.02$, in solving the subproblems (4.66). This choice can be justified by noting, as in [129, Section 4], that $\tilde{\mu}$ bounds the maximum residual $\|\mathbf{A}\mathbf{c} - \mathbf{u}\|_2$ up to a constant factor and any minimizer $\mathbf{c}_{\text{opt}} \approx \mathbf{c}$ of the unconstrained subproblem (4.66) satisfies

$$\|\mathbf{A}\mathbf{c}_{\text{opt}} - \mathbf{u}\|_2^2 = \|\mathbf{A}^*(\mathbf{A}\mathbf{c}_{\text{opt}} - \mathbf{u})\|_2^2 \lesssim \mathcal{O}(\|\mathbf{c}\|_0) \|\mathbf{A}^*(\mathbf{A}\mathbf{c}_{\text{opt}} - \mathbf{u})\|_\infty^2 \leq \mathcal{O}(\|\mathbf{c}\|_0) \tilde{\mu}^2. \quad (4.68)$$

Here the first equality follows when $\mathbf{A}\mathbf{A}^* = \mathbf{I}$, i.e., the setting relevant to orthogonalized Gaussian and partial DCT matrices, and the constant scaled inequality is an improvement over the standard estimate $\|\cdot\|_2 \leq \sqrt{N} \|\cdot\|_\infty$ in $\mathbb{R}^N$, see, e.g., [59, Appendix A.1]. The authors of [129] note that the FPC method is efficient at solving the subproblems (4.66) for the given $\tilde{\mu}$ in (4.67) and choice of $\xi = 0.02$. However, they also note that the resulting error of these “FPC only” approximations is much larger than that of the combined FPC-Bregman iterations when using a stopping condition for the Bregman iterations of

$$\frac{\|\mathbf{A}\mathbf{z}^k - \mathbf{u}\|_2}{\|\mathbf{u}\|_2} < 10^{-5}, \quad (4.69)$$

which is sufficient to give a small relative error $\|\mathbf{z}^k - \mathbf{c}\|_2 / \|\mathbf{c}\|_2$ of $\mathcal{O}(10^{-6})$. The work [128] proved “error-forgetting” and “error-cancelation” properties of the Bregman-FPC algorithm, which allow the subproblems of (4.41) to be solved at a lower accuracy, obtaining a highly-accurate result. Therefore selecting $\xi = 0.02$ is sufficient to guarantee the total accumulation of subproblem error does not exceed $\mathcal{O}(10^{-6})$ in that context.

In the considered setting of Section 3.6, our measurement model (see (3.68) and subsequent discussion) satisfies

$$\mathbf{A}\mathbf{c} + \mathbf{e} = \mathbf{u},$$

where $\mathbf{e} = \mathbf{u} - \mathbf{A}\mathbf{c}$ satisfies the tail bound $\|\mathbf{e}\|_{\mathcal{V},2} \leq \eta / \sqrt{m}$ for some $\eta > 0$, i.e., similar to (4.64) when $\mathbf{e}_2 = \mathbf{e}$ and $\mathbf{e}_1 = \mathbf{0}$. When $\mathbf{A}$ and $\mathbf{u}$ are rescaled with respect to $\hat{\lambda}_{\max}$, as in (4.56), we define

$$\hat{\lambda}_{\min} = \lambda_{\min}(\mathbf{A}\mathbf{A}^*). \quad (4.70)$$

For $m \ll N$ when $\mathbf{A}$ is given by (3.68) with the orthonormal Legendre system, we note that $\lambda_{\min}(\mathbf{H}) = 0$ (up to machine epsilon), while $\hat{\lambda}_{\min}$ is close but not equal to 0. Therefore, with the value of $\bar{\mu}$ from (4.65) in mind, we elect to choose $\mathbf{M} = \mathbf{I}$ and set

$$\bar{\mu} := \sqrt{\frac{N}{\xi}} \cdot \sqrt{\frac{\hat{\lambda}_{\max}}{\hat{\lambda}_{\min}}}, \quad \xi = 10^{-5}. \quad (4.71)$$

Table 4.1: Parameters used in the Bregman-FPC Algorithm 2 for the numerical experiments on sparse HV recovery in Section 6.3. The values of $I_{\text{Max-Bregman}}$ & $I_{\text{Max-FPC}}$ are selected based on run-time considerations.

Parameter	Value
γ (4.55)	0.99
β (4.55)	4
τ (4.58)	1
x_{tol} (4.60)	1
g_{tol} (4.60)	0.1
b_{tol} (4.62)	$1.2 \cdot \ \mathbf{A}\mathbf{c}^* - \mathbf{u}\ _{\mathcal{V},2}$
$\bar{\mu}$ (4.71)	$\sqrt{\frac{N}{\xi}} \cdot \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$
ξ (4.71)	10^{-5}
$I_{\text{Max-Bregman}}$	100
$I_{\text{Max-FPC}}$	1000

Our numerical experiments show that this choice of $\bar{\mu}$ is sufficient to yield accurate solutions to the Bregman iteration subproblems

$$\text{minimize}_{\mathbf{z} \in \mathcal{V}^N} \|\mathbf{z}\|_{\mathcal{V},1} + \frac{\bar{\mu}}{2} \|\mathbf{A}\mathbf{z} - \mathbf{u}^k\|_{\mathcal{V},2}^2,$$

within a reasonable amount of FPC iterations. For example, we find the number of FPC iterations j on lines 7 through 24 of Algorithm 2 is typically $\mathcal{O}(10^2)$. We conclude this chapter by summarizing our choice of input parameters for Algorithm 2. These parameters, included in Table 4.1, are supplied to the Bregman-FPC algorithm in the numerical experiments presented in Section 6.3. We find that these parameters work well for approximating the solution of the parameterized elliptic PDE problem (2.2) under the assumptions presented in Section 2.1.

Chapter 5

Computational complexity

In this chapter, we present results on the computational complexity of the methods presented in Chapter 3. Section 5.1 develops a thorough analysis of the computational complexity of the SG method. Under a certain set of assumptions, we derive bounds on the complexity of matrix-vector products with the SG matrix from (3.21) when discretizing problem (2.2) using the affine, polynomial, and transcendental functions from Examples 2.1-2.3. With these results, an ε -complexity analysis is then conducted to derive asymptotic cost estimates for solving the resulting fully-coupled systems of FE equations to as the tolerance $\varepsilon \rightarrow 0$. Section 5.2 briefly discusses similar ε -complexity results from [60] on solving problem (2.2) with the stochastic collocation method, providing a comparison to the results of Section 5.1. Finally, Section 5.3 presents results relating the sample complexity results from Sections 3.5 & 3.6 to error estimates for HV recovery and the necessary sparsity results for approximation of sparse polynomial expansions of solutions to parameterized PDEs by solving (3.69).

5.1 Complexity of stochastic Galerkin method

The primary goal of this section is to estimate the algorithmic complexity required by the SG method from Section 3.2 with FE physical discretization from Section 3.1 to construct an approximation to the solution u of (2.2) within a prescribed tolerance $\varepsilon > 0$. We assume $a(x, \mathbf{y})$ is a general coefficient satisfying assumptions (A1) & (A2). Under these assumptions, we examine the complexity of solving (3.14) in each of the cases of Examples 2.1-2.3, taking advantage of separable representations of the coefficient described in Section 3.2.2 in applying the matrix-free methods of solution described in Remark 3.3. We describe our approach and the necessary assumptions in our analysis as follows.

To simplify the analysis and provide a baseline for counting, we assume that the solution u is expanded in an orthonormal basis of a total degree polynomial space of order p , i.e., $P_{\mathcal{J}_p^{\text{TD}}}(\mathcal{U})$ with $\mathcal{J}_p^{\text{TD}}$ from (3.9). We also assume that the system $(\Psi_\nu)_{\nu \in \mathcal{J}_p^{\text{TD}}}$ used in the expansion of u is orthonormal with respect to a product of even weight functions ϱ_i , $i = 1, \dots, d$. In the case of the affine coefficient from Example 2.1, we rely on results from [55] on the sparsity of the SG systems in conducting our analysis. For the more general cases of Examples 2.2 & 2.3, we assume that all separable representations of the coefficient have non-trivial terms coinciding with a total degree set of order r , with $r \leq 2p$. Hence, in the

case of Example 2.2, we assume that $a(x, \mathbf{y})$ has terms $c_\alpha(x) \neq 0$ for each $\alpha \in \mathcal{J}_r^{\text{TD}}$. In the case of Example 2.3, where a may not be separable in x and $\mathbf{y}$, we apply the strategy of truncating the orthonormal expansion of a , obtaining an approximation a^r to a as in (3.16) on $P_{\mathcal{J}_r^{\text{TD}}}(\mathcal{U})$ in the same orthonormal system (Ψ_ν) as u at all relevant points $x \in D$. Recalling the discussion of Remark 3.4, we further assume that r satisfies $\tilde{r} \leq r$, with $\tilde{r}$ given in (3.22), so that $a^r(x, \mathbf{y})$ also satisfies (A1) & (A2). We then substitute the approximation a^r in the SG equations as in (3.21), counting the number of nonzero elements of the matrices $(\mathbf{G}_\alpha)_{\alpha \in \mathcal{J}_r^{\text{TD}}}$ under the general assumption $\tilde{r} \leq r \leq 2p$ and $p \in \mathbb{N}$.

In Theorem 5.1, we prove sharp estimates on $\text{nnz}(\mathbf{G}_\alpha)$ for each $\alpha \in \mathcal{J}_r^{\text{TD}}$, using these estimates to bound the number of nonzero elements of the matrices $\mathbf{K}_r$ (up to a multiplicative constant $\mathcal{O}(K_h)$), and hence FLOPS involved in matrix-vector products. These estimates can also be used to bound the number of nonzero elements of $\mathbf{K}$ from (3.14), and therefore are valid for any SG system resulting from the discretization of a parameterized elliptic PDE (or a parameterized PDE containing the elliptic operator) under the above assumptions. These results are presented in the context of solving the linear system (3.21) with a PCG method when a zero initial vector is used to seed the solver. The results, however, can be generalized to other methods, such as preconditioned GMRES and other Krylov methods.

The results are organized as follows. In Section 5.1.1 we present a cost metric in terms of FLOPs for solving the unpreconditioned and preconditioned linear systems (3.21). In Section 5.1.2 we discuss the overall complexity of the matrix-vector products associated with solving (3.21) when using the SG matrix $\mathbf{K}_r = \sum_{\alpha \in \mathcal{J}_r^{\text{TD}}} \mathbf{G}_\alpha \otimes \mathbf{B}_\alpha$. In particular, for $\mathbf{G}_\alpha$ given in (3.20), we show that $\text{nnz}(\mathbf{G}_\alpha) = \mathcal{O}(\min\{2^{|\alpha|}, N_{\lceil |\alpha|/2 \rceil}\} N_{p - \lceil r/2 \rceil})$ for every $\alpha \in \mathcal{J}_r^{\text{TD}}$, where $N_r = \#(\mathcal{J}_r^{\text{TD}}) = \binom{d+r}{d}$ for $r \in \mathbb{N}_0$, when solving (3.21), so that the total complexity of the matrix-vector products with the Galerkin system $\mathbf{K}_r$ is $\mathcal{O}(K_h N_p N_r \min\{2^r, N_{\lceil r/2 \rceil}\})$. In Section 5.1.3, we perform an ε -complexity analysis to derive the explicit cost bounds of the SG method with FE using PCG to solve, in terms of total number of FLOPs, as the tolerance $\varepsilon \rightarrow 0$. We conclude this section in Section 5.1.4 with a discussion of issues related to the conditioning of the SG system.

5.1.1 Cost of solving the generalized SG system

To solve the stochastic Galerkin system (3.21) for $\mathbf{u}_{h,p}^r \in \mathbb{R}^{K_h N_p}$, we use the preconditioned conjugate gradient (PCG) method. For the unpreconditioned CG method, we have the estimate

$$\|\mathbf{u}_{h,p}^r - \mathbf{u}_{h,p}^{r,k}\|_{\mathbf{K}_r} \leq 2 \left(\frac{\sqrt{\kappa_r} - 1}{\sqrt{\kappa_r} + 1} \right)^k \|\mathbf{u}_{h,p}^r - \mathbf{u}_{h,p}^{r,0}\|_{\mathbf{K}_r}. \quad (5.1)$$

Here κ_r is the condition number of $\mathbf{K}_r$, $\mathbf{u}_{h,p}^{r,0}$ is the vector of the initial guess, and $\mathbf{u}_{h,p}^{r,k}$ is the output of the k -th iteration of the CG solver. From (5.1), we see that the cost of the CG method is highly dependent on the conditioning of the system, and when κ_r is large, the number of iterations needed to reduce the error in $\mathbf{u}_{h,p}^{r,k}$ will also be large. Hence we introduce the mean-based block-diagonal preconditioner (see, e.g., [94, 99]),

$$\mathbf{P} := \mathbf{G}_0 \otimes \mathbf{B}_0, \quad (5.2)$$

with $\mathbf{B}_0$ and $\mathbf{G}_0$ the matrices defined in (3.18) and (3.20) for $\alpha = 0$, respectively. We note that the orthonormality of the system (Ψ_ν) implies that $\mathbf{G}_0 = \mathbf{I}$, the identity.

At every iteration of the CG method with $\mathbf{K}_r$, or any iterative approach, each nonzero entry in the matrices $(\mathbf{G}_\alpha)$ implies a matrix-vector product of the form $\langle \Psi_\nu \Psi_\mu \Psi_\alpha \rangle \mathbf{B}_\alpha \mathbf{u}_{h,\mu}^{r,k}$, where $\langle \Psi_\nu \Psi_\mu \Psi_\alpha \rangle$ is a scalar quantity. Therefore, letting $\text{nnz}(\mathbf{A})$ denote the number of nonzeros of a matrix $\mathbf{A}$, to account for the complexity of matrix-vector products with $\mathbf{K}_r$ we define

$$\mathcal{M}(p, r) = \sum_{\alpha \in \mathcal{J}_r^{\text{TD}}} \text{nnz}(\mathbf{G}_\alpha) \quad (5.3)$$

to be the total number of nonzeros in all of the matrices $(\mathbf{G}_\alpha)_{\alpha \in \mathcal{J}_r^{\text{TD}}}$ when u is expanded in the orthonormal system of total degree order p . With this in mind, an approximate upper bound for the work in terms of the total number of *floating point operations* (FLOPs) of solving (3.21) with the CG method is given by

$$W^{\text{SG}} \approx \mathcal{O}(K_h) * \mathcal{M}(p, r) * k_{\text{iter}}^{\text{SG}}. \quad (5.4)$$

Here the term $\mathcal{O}(K_h)$ corresponds to the cost of a single finite element matrix-vector product, and $k_{\text{iter}}^{\text{SG}}$ is the number of iterations of the CG solver without a preconditioner.

When applying a preconditioner in hopes to minimize $k_{\text{iter}}^{\text{SG}}$, we must also account for the added cost of preconditioning at each iteration. With the mean-based preconditioner from (5.2), at each iteration we multiply an additional matrix of size $K_h N_p \times K_h N_p$, consisting of only N_p blocks on the diagonal. Here we assume that in finding the inverse of the preconditioning matrix $\mathbf{P}$, a sparse approximation to $\mathbf{P}^{-1}$ is used, which we will denote $\tilde{\mathbf{P}}^{-1}$. Such a decomposition can be found from, e.g., incomplete LU or incomplete Cholesky factorizations. Hence, for each iteration we require N_p additional matrix-vector products of the size and complexity of the original finite element system, so the work estimate in FLOPs in this case is given by

$$W^{\text{pSG}} \approx \mathcal{O}(K_h) * (N_p + \mathcal{M}(p, r)) * k_{\text{iter}}^{\text{pSG}}, \quad (5.5)$$

where $k_{\text{iter}}^{\text{pSG}}$ is the number of iterations needed by the PCG method. Other preconditioners, such as the Kronecker product preconditioner suggested in [116], would require a different form of (5.5) corresponding to counting the number of FLOPs of applying the new choice.

Figure 5.1 illustrates the importance of accurately counting $\mathcal{M}(p, r)$ in estimating the cost of matrix-vector products involving $\mathbf{K}_r$ by displaying a sequence of sparsity plots of $\sum_{\alpha \in \mathcal{J}_r^{\text{TD}}} \mathbf{G}_\alpha \in \mathbb{R}^{N_p \times N_p}$. The progression from left to right shows the effect of fixing the total degree expansion order p of the solution u and increasing the order r of the projection a^r of the coefficient a in the elliptic operator, after discretization by (3.20). Such situations arise when balancing the projection orders p and r in approximating highly nonlinear coefficients a and resulting solutions u . Therefore these sparsity plots reflect the importance of considering $\mathcal{M}(p, r)$, with respect to both p and r , in the computational cost of the SG method.

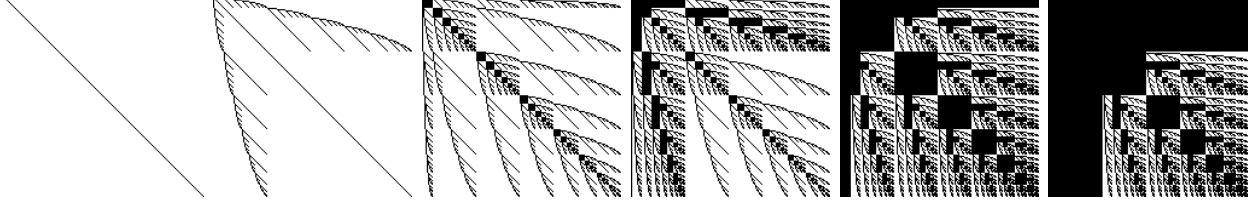


Figure 5.1: Visualization of the number of nonzeros of $\sum_{\alpha \in \mathcal{J}_r^{\text{TD}}} \mathbf{G}_\alpha \in \mathbb{R}^{N_p \times N_p}$ with $\mathbf{G}_\alpha$ from (3.20), when p and $\mathcal{J}_p^{\text{TD}}$ are fixed and r and $\mathcal{J}_r^{\text{TD}}$ are allowed to grow. Each pixel implies a nonzero block $\langle \Psi_\nu \Psi_\mu \Psi_\alpha \rangle * \mathbf{B}_\alpha \in \mathbb{R}^{K_h \times K_h}$ in the matrix $\mathbf{K}_r$. Here $d = 8$ and $p = 3$ so that $N_p = \#(\mathcal{J}_p^{\text{TD}}) = 165$. Moving left to right r is increasing by one unit from 0 to 5. At $r = 6$, the matrix is entirely block-dense.

5.1.2 Complexity of SG matrix-vector products

In this section we provide rigorous bounds on the sparsity of the SG matrix $\mathbf{K}_r$ from (3.21), for arbitrary $0 \leq r \leq 2p$, and $p \in \mathbb{N}_0$. Our main result, given by Theorem 5.1, provides an exact count for $\text{nnz}(\mathbf{G}_\alpha)$ in the general case $|\alpha| \in \mathbb{N}_0$ and $d \in \mathbb{N}$ when the integrals $[\mathbf{G}_\alpha]_{\nu, \mu} = \langle \Psi_\nu \Psi_\mu \Psi_\alpha \rangle$ are defined in terms of products of even weight functions ϱ_i , $i = 1, \dots, d$. This result can be seen as an extension of estimates from [55], where bounds on the sparsity of $\mathbf{G}_\alpha$ were shown in the cases (i) $|\alpha| = 1$ and $d \in \mathbb{N}$ and (ii) $d = 1$ and $\alpha = r \in \mathbb{N}_0$. We then provide upper bounds on the total number of nonzero blocks $\mathcal{M}(p, r) = \sum_{\alpha \in \mathcal{J}_r^{\text{TD}}} \text{nnz}(\mathbf{G}_\alpha)$ of the matrix $\mathbf{K}_r$ from (3.21), both in the cases that $a(x, \mathbf{y})$ is a finite order polynomial as in Example 2.2 and the case that $a(x, \mathbf{y})$ is a transcendental function of the random variables, as in Example 2.3. Our first major result is summarized in the following Theorem:

Theorem 5.1. *Let $\mathcal{J}_p^{\text{TD}}$ and $\mathcal{J}_r^{\text{TD}}$ be the isotropic total degree index sets corresponding to the solution and coefficient, respectively, with $p, r \in \mathbb{N}_0$, and $0 \leq r \leq 2p$. If $\alpha \in \mathcal{J}_r^{\text{TD}}$, and ϱ_i are even for all $i = 1, \dots, d$, then for the matrix $\mathbf{G}_\alpha$ from (3.20) we have*

$$\text{nnz}(\mathbf{G}_\alpha) = \sum_{\ell=\lceil |\alpha|/2 \rceil}^{|\alpha|} c(\alpha, \ell) \binom{d+p-\ell}{p-\ell}, \quad (5.6)$$

$$c(\alpha, \ell) = \begin{cases} \#\mathbf{S}(\alpha, \ell) & |\alpha| \text{ even}, \ell = |\alpha|/2, \\ 2\#\mathbf{S}(\alpha, \ell) & \text{otherwise}, \end{cases} \quad (5.7)$$

with $\mathbf{S}(\alpha, \ell) = \{\mathbf{s} \in \mathbb{N}_0^d : |\mathbf{s}| = \ell, \mathbf{s} \leq \alpha\}$, so that $\#\mathbf{S}(\alpha, \ell)$ is equal to the coefficient of t^ℓ in the polynomial $p_\alpha(t) = \prod_{i=1}^d \sum_{j=0}^{r_i} t^j$. Moreover, we have the following bound for $\text{nnz}(\mathbf{G}_\alpha)$, i.e.,

$$\text{nnz}(\mathbf{G}_\alpha) \leq 2 \min \left\{ 2^{|\alpha|}, \binom{d + \lceil |\alpha|/2 \rceil}{d} \right\} \binom{d+p - \lceil |\alpha|/2 \rceil}{d}, \quad (5.8)$$

so that

$$\mathcal{M}(p, r) \leq 2 \sum_{j=0}^r \min \left\{ 2^j, \binom{d + \lceil j/2 \rceil}{d} \right\} \binom{d-1+j}{d-1} \binom{d+p-\lceil j/2 \rceil}{d}. \quad (5.9)$$

Proof. For a given $\alpha \in \mathcal{J}_r^{\text{TD}}$, we estimate the number of pairs $(\nu, \mu) \in \mathcal{J}_p^{\text{TD}} \times \mathcal{J}_p^{\text{TD}}$ such that $\langle \Psi_\alpha \Psi_\nu \Psi_\mu \rangle = \prod_{i=1}^d \langle \psi_{r_i} \psi_{p_i} \psi_{q_i} \rangle \neq 0$. To do this, we extend the result of [55, Lemma 28] to a general matrix $\mathbf{G}_\alpha$ with $|\alpha| \in \mathbb{N}$. Since $(\Psi_\alpha)_{\alpha \in \mathcal{J}_r^{\text{TD}}}$ are orthonormal with respect to the even weight function $\varrho(\mathbf{y}) = \prod_{i=1}^d \varrho_i(y_i)$, we see that $\langle \Psi_\alpha \Psi_\nu \Psi_\mu \rangle \neq 0$ if and only if $(\nu, \mu) \in \Theta_\alpha$, where

$$\begin{aligned} \Theta_\alpha = \{(\nu, \mu) \in \mathcal{J}_p^{\text{TD}} \times \mathcal{J}_p^{\text{TD}} : & |p_i - q_i| \leq r_i \leq p_i + q_i, \\ & \text{and } p_i + q_i + r_i \text{ is even } \forall i = 1, \dots, d\}. \end{aligned}$$

Therefore, to estimate the number of nonzeros in the matrix $\mathbf{G}_\alpha$, we must estimate $\#\Theta_\alpha$. However, Θ_α is different for each $\alpha \in \mathcal{J}_r^{\text{TD}}$. Even when $\alpha_1, \alpha_2 \in \mathcal{J}_r^{\text{TD}}$ are such that $|\alpha_1| = |\alpha_2|$, in general we do not have that $\#\Theta_{\alpha_1} = \#\Theta_{\alpha_2}$. On the other hand, if α_2 is a permutation of α_1 , then it is easy to see that $\#\Theta_{\alpha_1} = \#\Theta_{\alpha_2}$ since $\mathcal{J}_p^{\text{TD}}$ is the isotropic total degree set. Also, by symmetry if $(\nu, \mu) \in \Theta_\alpha$ then $(\mu, \nu) \in \Theta_\alpha$ as well. Note that Θ_α can be rewritten

$$\begin{aligned} \Theta_\alpha = \{(\nu, \mu) \in \mathcal{J}_p^{\text{TD}} \times \mathcal{J}_p^{\text{TD}} : & |p_i - q_i| \leq r_i \leq p_i + q_i, \\ & \text{and } |p_i - q_i| + r_i \text{ is even } \forall i = 1, \dots, d\}. \end{aligned}$$

Hence if $(\nu, \mu) \in \Theta_\alpha$, then we see that (ν, μ) must satisfy

- (i) $|p_i - q_i| \leq r_i$ for all $i = 1, \dots, d$,
- (ii) $r_i \leq p_i + q_i$ for all $i = 1, \dots, d$, and
- (iii) $|p_i - q_i| + r_i$ is even for all $i = 1, \dots, d$.

Note that when $r_i = 0$, we see that $p_i = q_i \leq p$, and when $r_i > 0$ we see that (i) and (iii) imply $|p_i - q_i| \in \{0, 2, 4, \dots, r_i\}$ for r_i even, and $|p_i - q_i| \in \{1, 3, 5, \dots, r_i\}$ for r_i odd. For each $i = 1, \dots, d$, let $\{k_i^{(n)}\}_{n=0}^{\lfloor r_i/2 \rfloor}$ be the sequence defined by

$$k_i^{(n)} = \begin{cases} 2n + 1 & r_i \text{ odd,} \\ 2n & r_i \text{ even,} \end{cases}$$

so that fixing $|p_i - q_i| = k_i^{(n)}$ implies that conditions (i) and (iii) are met.

To satisfy (ii) we must have $r_i \leq p_i + q_i$ and to satisfy (i) and (iii) we must have $|p_i - q_i| = k_i^{(n)}$. To avoid overcounting due to symmetry, we first fix possible values of p_i and consider what q_i must be. Let $\{s_i^{(n)}\}_{n=0}^{\lfloor r_i/2 \rfloor}$ be the sequence defined by

$$s_i^{(n)} = \frac{r_i + k_i^{(n)}}{2},$$

which we will refer to as the sequence of *starting points* for p_i corresponding to $k_i^{(n)}$. Note that the starting points $\{s_i^{(n)}\}_{n=0}^{\lfloor r_i/2 \rfloor}$ enumerate the integers between $\lceil r_i/2 \rceil$ and r_i . Picking $p_i \in \{s_i^{(n)}, s_i^{(n)} + 1, \dots, p\}$ and $q_i = p_i - k_i^{(n)}$ we have

$$p_i + q_i = 2p_i - k_i^{(n)} \geq 2s_i^{(n)} - k_i^{(n)} = 2 \left(\frac{r_i + k_i^{(n)}}{2} \right) - k_i^{(n)} = r_i$$

or $p_i + q_i \geq r_i$, so that (ii) is satisfied.

Since (i), (ii), and (iii) are satisfied by setting $p_i \in \{s_i^{(n)}, s_i^{(n)} + 1, \dots, p\}$ and $q_i = p_i - k_i^{(n)}$ for a fixed $0 \leq n \leq \lfloor r_i/2 \rfloor$, we count the number of admissible pairs for these choices. In $d-1$ variables, the number of polynomials of total degree less than or equal to $p - p_i$ is given by

$$\binom{d-1+p-p_i}{p-p_i},$$

where $\binom{n}{k} = 0$ if $n < k$ or $n, k < 0$. To simplify notation, pick $s_i = s_i^{(n)}$ (one of the starting points in the i -th direction) and $k_i = k_i^{(n)}$ (its associated distance), where $0 \leq n \leq \lfloor r_i/2 \rfloor$ is fixed. To count the number of admissible pairs associated with the difference k_i and starting point s_i , we compute

$$\sum_{p_i=s_i}^p \binom{d-1+p-p_i}{p-p_i} = \sum_{j=0}^{p-s_i} \binom{d-1+j}{j} = \binom{d+p-s_i}{p-s_i}.$$

Define $\mathbf{s} \in \mathbb{N}_0^d$ with the s_i as above, then $\mathbf{s}$ corresponds to a possible combination of starting points in each direction. To estimate the number of polynomials associated with the starting point $\mathbf{s}$, we compute

$$\sum_{p_1=s_1}^p \sum_{p_2=s_2}^{p-p_1} \cdots \sum_{p_d=s_d}^{p-p_1-\cdots-p_{d-1}} \binom{p-p_1-\cdots-p_d}{p-p_1-\cdots-p_d} = \binom{d+p-|\mathbf{s}|}{p-|\mathbf{s}|}, \quad (5.10)$$

where the sum easily follows by an induction argument and Pascal's rule.

Enumerating all of the pairs $(\boldsymbol{\nu}, \boldsymbol{\mu}) \in \Theta_{\boldsymbol{\alpha}}$ thus reduces to counting the number of possible combinations of starting points. Hence, in d dimensions we consider all such multi-indices of the $\{s_i^{(n)}\}_{n=0}^{\lfloor r_i/2 \rfloor}$ whose components sum to some integer $\lceil |\boldsymbol{\alpha}|/2 \rceil \leq \ell \leq |\boldsymbol{\alpha}|$. For two multi-indices $\mathbf{s}, \boldsymbol{\alpha} \in \mathbb{N}_0^d$, we say $\mathbf{s} \leq \boldsymbol{\alpha}$ if and only if $s_i \leq r_i$ for all $i = 1, \dots, d$. Define the set $\mathbf{S}(\boldsymbol{\alpha}, \ell) = \{\mathbf{s} \in \mathbb{N}_0^d : |\mathbf{s}| = \ell, \mathbf{s} \leq \boldsymbol{\alpha}\}$, which corresponds to a particular slice of the desired set of starting points. To estimate $\#\mathbf{S}(\boldsymbol{\alpha}, \ell)$, we consider the familiar counting argument of placing d bars among ℓ stars with the added restriction that the number of stars in the i -th bin not exceed r_i . Such a problem can be reframed in terms of finding the coefficient $c(\boldsymbol{\alpha}, \ell)$ of t^ℓ in the generating function $p_{\boldsymbol{\alpha}}(t) = \prod_{i=1}^d \sum_{j=0}^{r_i} t^j$. Combining (5.10) and summing over ℓ between $\lceil |\boldsymbol{\alpha}|/2 \rceil \leq \ell \leq |\boldsymbol{\alpha}|$ we arrive at (5.6), where the coefficients $c(\boldsymbol{\alpha}, \ell) = \#\mathbf{S}(\boldsymbol{\alpha}, \ell)$ when $|\boldsymbol{\alpha}|$ is even and $\ell = |\boldsymbol{\alpha}|/2$ (in this case the roles of p_i and q_i can not be reversed) and $c(\boldsymbol{\alpha}, \ell) = 2\#\mathbf{S}(\boldsymbol{\alpha}, \ell)$ otherwise.

Noting that $\cup_{\ell=\lceil |\alpha|/2 \rceil}^{|\alpha|} \mathbf{S}(\alpha, \ell)$ is a change of coordinates of a total degree index set of order $\lceil |\alpha|/2 \rceil$ intersected with the hyperrectangle $\{\mathbf{s} \in \mathbb{N}_0^d : \mathbf{s} \leq \alpha\}$ yields the bound

$$\sum_{\ell=\lceil |\alpha|/2 \rceil}^{|\alpha|} c(\alpha, \ell) \leq 2 \sum_{\ell=\lceil |\alpha|/2 \rceil}^{|\alpha|} \#\mathbf{S}(\alpha, \ell) \leq 2 \binom{d + \lceil |\alpha|/2 \rceil}{d}.$$

On the other hand, from the generating function $p_\alpha(t)$ we see that $c(\alpha, \ell)$ is bounded by $\binom{|\alpha|}{\ell}$ when $|\alpha|$ is even and $\ell = |\alpha|/2$ and $2\binom{|\alpha|}{\ell}$ otherwise. This follows from the fact that when $\mathbf{k}$ is the multi-index having $|\alpha|$ ones and the rest zeros, since $\ell \leq |\alpha|$, we have that $\#\mathbf{S}(\alpha, \ell) \leq \#\mathbf{T}(\mathbf{k}, \ell)$ where $\mathbf{T}(\mathbf{k}, \ell) = \{\mathbf{s} \in \mathbb{N}_0^d : |\mathbf{s}| = \ell, \mathbf{s} \leq \mathbf{k}\}$ and $\#\mathbf{T}(\mathbf{k}, \ell)$ is given by the coefficient of t^ℓ in $p_\mathbf{k}(t) = (1+t)^{|\alpha|} = \sum_{\ell=0}^{|\alpha|} \binom{|\alpha|}{\ell} t^\ell$ from the binomial theorem. Then

$$\sum_{\ell=\lceil |\alpha|/2 \rceil}^{|\alpha|} c(\alpha, \ell) \leq 2 \sum_{\ell=0}^{|\alpha|} c(\alpha, \ell) = 2^{|\alpha|+1},$$

so that

$$\text{nnz}(\mathbf{G}_\alpha) = \sum_{\ell=\lceil |\alpha|/2 \rceil}^{|\alpha|} c(\alpha, \ell) \binom{d+p-\ell}{p-\ell} \leq 2 \min \left\{ 2^{|\alpha|}, \binom{d + \lceil |\alpha|/2 \rceil}{d} \right\} \binom{d+p-\lceil |\alpha|/2 \rceil}{d},$$

showing (5.8). Substituting (5.8) into (5.3) shows the bound of $\mathcal{M}(p, r)$ from (5.9). ■

We note that the bound of $\mathcal{M}(p, r)$ from (5.9) is an overestimate due to the particular form of (5.6), which is different for each $\alpha \in \mathcal{J}_r^{\text{TD}}$. As a consequence, we see that $\text{nnz}(\mathbf{G}_\alpha) = \mathcal{O}(\min\{2^{|\alpha|}, N_{\lceil |\alpha|/2 \rceil}\} N_{p-\lceil |\alpha|/2 \rceil})$ for $\alpha \in \mathcal{J}_r^{\text{TD}}$. For large d and small r , 2^r is smaller than $N_{\lceil r/2 \rceil}$, however, as $r \rightarrow \infty$ the bound $N_{\lceil r/2 \rceil}$ is sharper. For the ε -complexity analysis in the next section, we note that

$$\begin{aligned} \mathcal{M}(p, r) &= \sum_{\alpha \in \mathcal{J}_r^{\text{TD}}} \text{nnz}(\mathbf{G}_\alpha) \leq \sum_{\alpha \in \mathcal{J}_r^{\text{TD}}} 2 \min \left\{ 2^{|\alpha|}, \binom{d + \lceil |\alpha|/2 \rceil}{d} \right\} \binom{d+p-\lceil |\alpha|/2 \rceil}{d} \\ &= \sum_{j=0}^r 2 \min \left\{ 2^j, \binom{d + \lceil j/2 \rceil}{d} \right\} \binom{d-1+j}{d-1} \binom{d+p-\lceil j/2 \rceil}{d} \\ &\leq 2 \min \left\{ 2^r, \binom{d + \lceil r/2 \rceil}{d} \right\} \binom{d+p}{d} \sum_{j=0}^r \binom{d-1+j}{d-1} \\ &= 2 \min \left\{ 2^r, \binom{d + \lceil r/2 \rceil}{d} \right\} \binom{d+p}{d} \binom{d+r}{d}. \end{aligned} \quad (5.11)$$

Figure 5.2 plots how $\mathcal{M}(p, r)$ and its bounds (5.9) and (5.11) scale when r in Theorem 5.1 is fixed at the values 1, 2, and 3 and p is increasing. This setting of fixed r is relevant to the affine and polynomial coefficients in Examples 2.1 & 2.2, and illustrates why the bound (5.9) is used Theorem 5.9 and its proof for polynomial coefficients. Figure 5.3 plots how sharply $\mathcal{M}(p, r)$ is bounded by (5.9) and (5.11) in the case that $r = p$, i.e., increasing p and

r together. There it can be seen that in the case $r = p$, the bound (5.9) is a much better estimate of the actual sparsity $\mathcal{M}(p, r)$ than the bound (5.11). We are also able to show that Theorem 5.1 yields an exact result in the case $|\alpha| = 1$.

Corollary 5.2. *Under the same conditions in Theorem 5.1, when $\alpha \in \mathcal{J}_r^{\text{TD}}$ is such that $|\alpha| = 1$, we have*

$$\text{nnz}(\mathbf{G}_\alpha) = \sum_{\ell=\lceil |\alpha|/2 \rceil}^{|\alpha|} c(\alpha, \ell) \binom{d+p-\ell}{p-\ell} = 2 \binom{d+p-1}{p-1}. \quad (5.12)$$

Corollary 5.2 is the result of [55, Lemma 28], and follows from the application of the exact formula for $\text{nnz}(\mathbf{G}_\alpha)$ from (5.6). Here $|\alpha| = \lceil |\alpha|/2 \rceil = 1$ is odd and $\mathcal{S}(\alpha, 1)$ has only one element $\mathcal{S}(\alpha, 1) = \{\mathbf{s} \in \mathbb{N}_0^d : |\mathbf{s}| = 1, \mathbf{s} \leq \alpha\} = \{\alpha\}$. Hence $c(\alpha, 1) = 2\#\mathcal{S}(\alpha, 1) = 2$, and (5.12) is shown. We are also able to show that the formula for $\text{nnz}(\mathbf{G}_\alpha)$ from (5.6) yields a result that is exact in the case $d = 1$.

Corollary 5.3. *Under the same conditions in Theorem 5.1, when $d = 1$ and $\alpha = r \in \mathbb{N}_0$, we have*

(a) in case $r = 2k$, $k \in \mathbb{N}_0$,

$$\text{nnz}(\mathbf{G}_r) = \begin{cases} (p-r+1)(r+1) + k^2, & 0 \leq r \leq p, \\ (p-k+1)^2, & p+1 \leq r \leq 2p, \\ 0, & r > 2p. \end{cases} \quad (5.13)$$

(b) in case $r = 2k+1$, $k \in \mathbb{N}_0$,

$$\text{nnz}(\mathbf{G}_r) = \begin{cases} (p-r+1)(r+1) + k^2 + k, & 0 \leq r \leq p, \\ (p-k+1)(p-k), & p+1 \leq r \leq 2p, \\ 0, & r > 2p. \end{cases} \quad (5.14)$$

Corollary 5.3 is the result of [55, Lemma 25], and its proof using Theorem 5.1 is included in Appendix B. In the remarks that follow, we make a distinction between the cases that $a(x, \mathbf{y})$ is a polynomial of fixed degree $\bar{r} < \infty$, e.g., the coefficients from Examples 2.1 and 2.2, and that $a(x, \mathbf{y})$ is a transcendental function of the random variables, e.g., the coefficient from Example 2.3.

Remark 5.4. (Complexity of matrix-vector products for polynomial coefficients, see e.g., Examples 2.1 and 2.2) *From Corollary 5.2 and the work estimate (5.5) when using (5.2) as a preconditioner, we see that for coefficients that are affine functions of the random variables, e.g., Example 2.1, the complexity of a single PCG iteration is of the order $\mathcal{O}(K_h(2N_p + 2NN_{p-1})) = \mathcal{O}(K_h N_p)$, where $N_p = \#\mathcal{J}_p^{\text{TD}} = \binom{d+p}{d}$. On the other hand, when the coefficient $a(x, \mathbf{y})$ is a polynomial function of the random variables, e.g., Example 2.2, having fixed order $\bar{r} \in \mathbb{N}$, $\bar{r} < \infty$, we use Theorem 5.1 to obtain a different estimate. Since $(\Psi_\alpha)_{\alpha \in \mathcal{J}_{\bar{r}}^{\text{TD}}}$ is a basis for the space $P_{\mathcal{J}_{\bar{r}}^{\text{TD}}}(\mathcal{U})$, there exists coefficients $(a_\alpha(x))_{\alpha \in \mathcal{J}_{\bar{r}}^{\text{TD}}}$ such that $a(x, \mathbf{y}) =$*

$a^{\bar{r}}(x, \mathbf{y}) = \sum_{\alpha \in \mathcal{J}_{\bar{r}}^{\text{TD}}} a_{\alpha}(x) \Psi_{\alpha}(\mathbf{y})$. With this representation, it is clear to see that substituting $a(x, \mathbf{y})$ into (3.17) yields $\mathbf{K}_{\bar{r}}$ from (3.21), and $\mathbf{K}_{\bar{r}} = \mathbf{K}$ from (3.14). However, it is not clear how many of the coefficients $a_{\alpha}(x)$ are identically zero. In this case, we can provide an upper bound on the block-sparsity of $\mathbf{K}_{\bar{r}}$ under the assumption that $a_{\alpha}(x) \not\equiv 0 \forall \alpha \in \mathcal{J}_{\bar{r}}^{\text{TD}}$. Using the bound of (5.11), the complexity of a single matrix-vector product of $\mathbf{K}_{\bar{r}}$ is of the order $\mathcal{O}(K_h N_p N_{\bar{r}} \min\{2^{\bar{r}}, N_{\lceil \bar{r}/2 \rceil}\})$. Thus, when $\bar{r}$ is fixed, $N_{\bar{r}} \min\{2^{\bar{r}}, N_{\lceil \bar{r}/2 \rceil}\}$ is a constant, and this estimate has the same asymptotic complexity as $\mathcal{O}(K_h N_p)$.

Remark 5.5. (Complexity of matrix-vector products in the transcendental case, see e.g., Example 2.3) We recall the discussion of [116, Section 3.4]. There, the complexity of matrix-vector products with the SG system was estimated when a full orthogonal expansion is substituted into the SG discretization. This case corresponds to fixing the expansion order $r = 2p$ following Remark 3.4. Assuming that $\text{nnz}(\mathbf{G}_{\alpha}) = \mathcal{O}(N_p)$ or $\mathcal{O}(N_p^2)$, it was estimated in [117] that the cost of matrix-vector products involving $\mathbf{K}_r$ is between $\mathcal{O}(K_h N_p^2)$ and $\mathcal{O}(K_h N_p^3)$. However, the use of Theorem 5.1 allows us to consider the complexity in the case of truncating the expansion, where a sharper estimate can be obtained. Let $T_r := \prod_{k=\lceil r/2 \rceil+1}^r \frac{d+k}{k} \ll N_{\lceil r/2 \rceil} = \binom{N+\lceil r/2 \rceil}{\lceil r/2 \rceil}$, which is bounded independent of r , i.e.,

$$T_r \leq \left(\frac{d + \lceil r/2 \rceil + 1}{\lceil r/2 \rceil + 1} \right)^{\lceil r/2 \rceil + 1} \rightarrow e^d \quad \text{as } r \rightarrow \infty,$$

so that $N_r = T_r N_{\lceil r/2 \rceil} \leq e^d N_{\lceil r/2 \rceil}$. From (5.11), we see that $\mathcal{M}(p, r)$ is of the order $\mathcal{O}(N_p N_r N_{\lceil r/2 \rceil})$ as $p, r \rightarrow \infty$, since $\min\{2^r, N_{\lceil r/2 \rceil}\} \rightarrow N_{\lceil r/2 \rceil}$ as $r \rightarrow \infty$. When $r = 2p$, this implies the complexity of matrix-vector multiplications involving $\mathbf{K}_r$ is of the order $\mathcal{O}(K_h N_p N_r N_{\lceil r/2 \rceil}) = \mathcal{O}(K_h N_p T_r N_{\lceil r/2 \rceil}^2) = \mathcal{O}(K_h N_p^3)$. On the other hand, when $r = p$, we see that the complexity of matrix-vector products with $\mathbf{K}_r$ is order $\mathcal{O}(K_h N_p N_r N_{\lceil r/2 \rceil}) = \mathcal{O}(K_h T_r^2 N_{\lceil p/2 \rceil}^3) = \mathcal{O}(K_h N_{\lceil p/2 \rceil}^3)$.

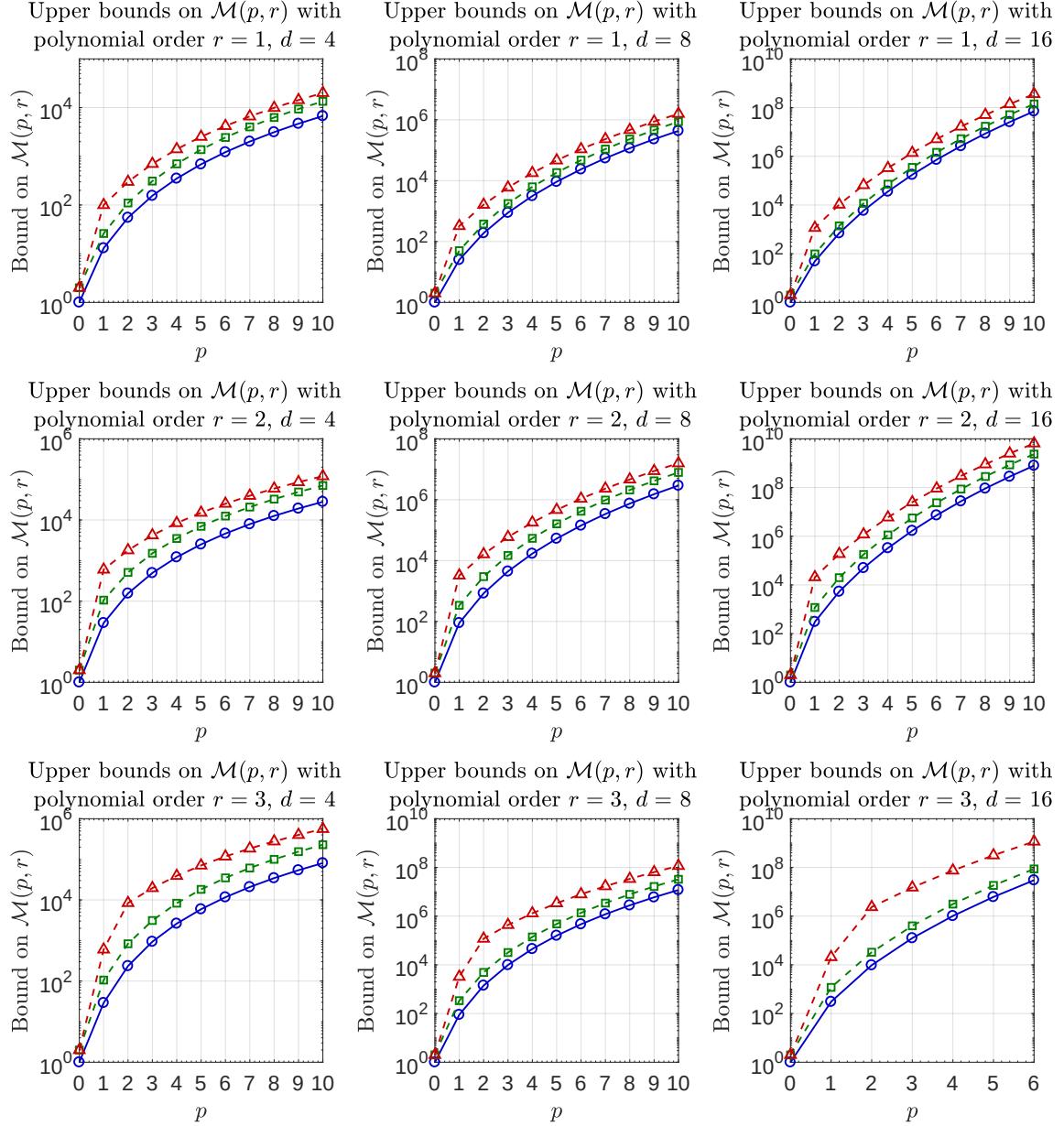


Figure 5.2: For a range of p values, we plot for **(top)** $r = 1$, **(middle)** $r = 2$, and **(bottom)** $r = 3$, the actual sparsity $\mathcal{M}(p, r)$ given by (5.3) of the Galerkin system $\mathbf{K}_r$ from (3.21) (blue), the bound on the sparsity from (5.9) (green), and the bound on the sparsity from (5.11) (red) in the cases $d = 4, 8, 16$ moving left to right.

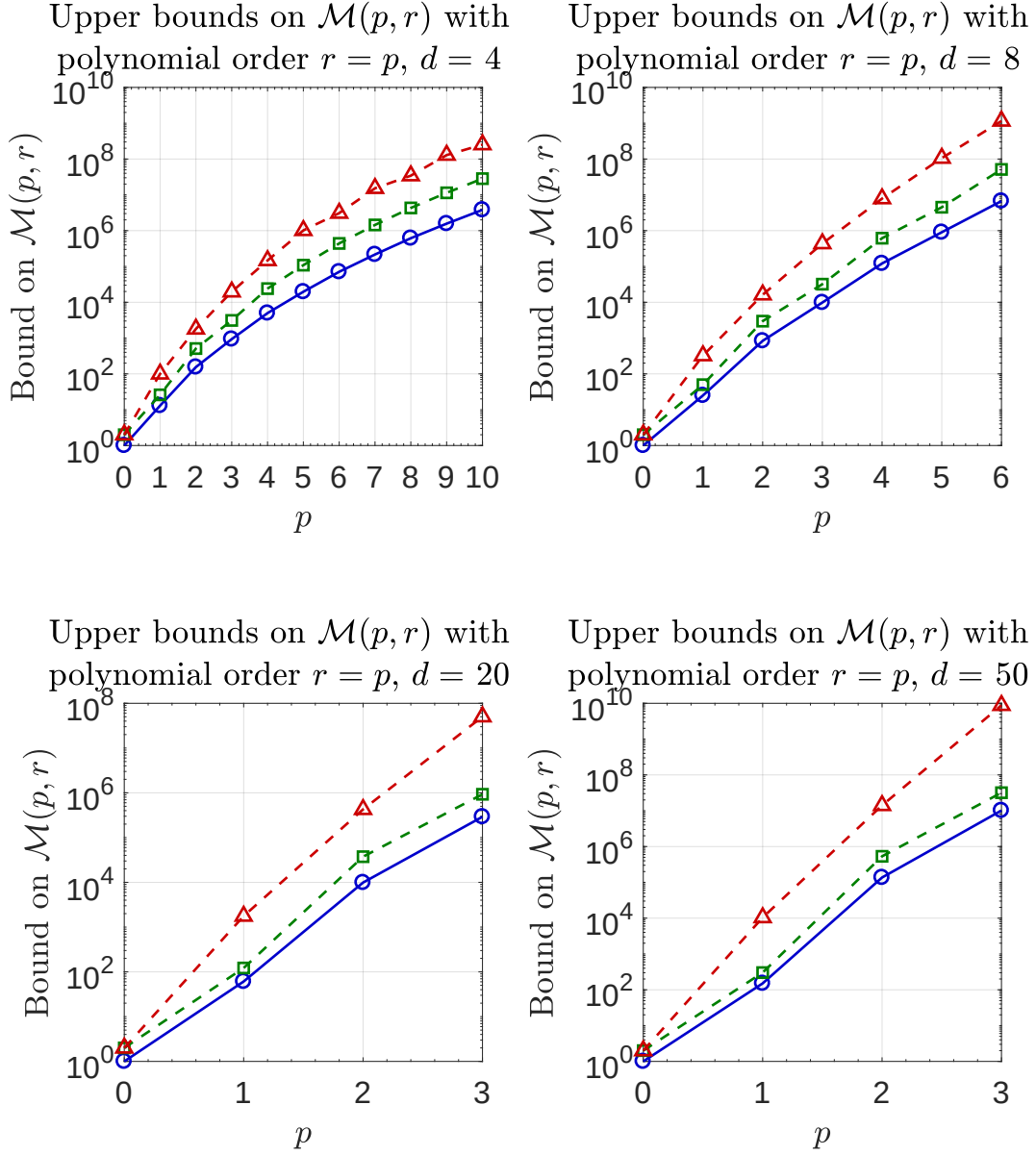


Figure 5.3: For $r = p$ with a range of p values, we plot for **(top)** $d = 4$ (left) and $d = 8$ (right) and **(bottom)** $d = 20$ (left) and $d = 50$ (right), the actual sparsity $\mathcal{M}(p, r)$ given by (5.3) of the Galerkin system $\mathbf{K}_r$ from (3.21) (blue), the bound on the sparsity from (5.9) (green), and the bound on the sparsity from (5.11) (red).

5.1.3 ε -complexity analysis of the SG method

An estimate of the total complexity to obtain a fully discrete approximation of tolerance $\varepsilon > 0$ with the SG and FE methods and PCG solver can be shown in four steps:

1. Estimate the maximum mesh size $h_{\max}$ and minimum polynomial order $p_{\min}$ necessary in the finite element and SG discretizations, respectively,
2. If projection of the coefficient is necessary, estimate the minimum projection order $r_{\min}$, otherwise set $r_{\min} = \bar{r}$ where $\bar{r} < \infty$ is the order of the coefficient,
3. Estimate the minimum number of iterations $k_{\min}$ needed by the PCG solver,
4. Substitute $h_{\max}$, $p_{\min}$, $r_{\min}$, and $k_{\min}$ into the cost (5.5) for h , p , r , and $k_{\text{iter}}^{\text{PSG}}$, respectively.

We proceed to estimate these parameters as follows. Denote by u^r the corresponding solution of (2.2) when $a^r(x, \mathbf{y})$ is substituted in place of $a(x, \mathbf{y})$, and let $\tilde{u}_{h,p}^r$ be the approximation to $u_{h,p}^r$ found by PCG. Then the total error for the SG approximation satisfies the following bound:

$$\|u - \tilde{u}_{h,p}^r\|_{\mathcal{H}_e^2} \leq \underbrace{\|u - u^r\|_{\mathcal{H}_e^2}}_{\text{SG(I)}} + \underbrace{\|u^r - u_h^r\|_{\mathcal{H}_e^2}}_{\text{SG(II)}} + \underbrace{\|u_h^r - u_{h,p}^r\|_{\mathcal{H}_e^2}}_{\text{SG(III)}} + \underbrace{\|u_{h,p}^r - \tilde{u}_{h,p}^r\|_{\mathcal{H}_e^2}}_{\text{SG(IV)}}. \quad (5.15)$$

In this setting SG(I) is the approximation error using a truncated expansion of $a(x, \mathbf{y})$, SG(II) is the discretization error induced by the finite element method, SG(III) is the SG error coming from the orthogonal expansion (3.10), and SG(IV) is the solver error resulting from the PCG method. We note that when the projection of the coefficient is exact, as discussed in Remark 3.4, the approximation error SG(I) is no longer present and $u_{h,p}^r \equiv u_{h,p}$.

We start with bounding SG(III). Without loss of generality, it is reasonable to assume that since u^r has a holomorphic dependence on $\mathbf{z} \in \mathbb{C}^d$ in an open neighborhood of the polyellipse $\mathcal{E}_\gamma$ from Theorem 2.4, then u_h^r does as well. Then, the following result, whose proof is found in [112], and immediately follows from classical spectral convergence results [40, 109], describes the convergence rate of the fully discrete solutions obtained by the SG method using a total degree approximation in $P_{\mathcal{J}_p^{\text{TD}}}(\mathcal{U})$:

Proposition 5.6 (Convergence rate for the SG method). *If Theorem 2.4 holds for the solution u_h^r to (3.1) with coefficient $a^r(x, \mathbf{y})$, and $u_{h,p}^r$ is the solution to (3.11) with $\mathcal{J}_p^{\text{TD}}$ the order p total degree index set, then*

$$\|u_h^r - u_{h,p}^r\|_{\mathcal{H}_e^\infty} \leq C_1 \exp(-C_2 p) \quad \forall p \in \mathbb{N},$$

for some constants $C_1, C_2 > 0$ independent of p .

To investigate the error in SG(I), we note that since $a(x, \mathbf{y})$ satisfies assumption (A2), the projection error in $P_{\mathcal{J}_r^{\text{TD}}}(\mathcal{U})$ can be similarly estimated as

$$\|a - a^r\|_{L_e^2(\mathcal{U}; L^\infty(D))} \leq C_3 \exp(-C_4 r) \quad \forall r \in \mathbb{N}, \quad (5.16)$$

for some constants $C_3, C_4 > 0$ independent of r . Hence, $\forall r \in \mathbb{N}$,

$$\|u - u^r\|_{\mathcal{H}_\varepsilon^2} \leq \frac{\|f\|_{H^{-1}}}{a_{\min}^2} \|a - a^r\|_{L_\varepsilon^2(\mathcal{U}; L^\infty(D))} \leq \frac{\|f\|_{H^{-1}}}{a_{\min}^2} C_3 \exp(-C_4 r) \quad (5.17)$$

providing a bound for SG(I). For a bound of SG(II), we present the following convergence result regarding solutions to the parameterized finite element problem, whose proof can be found in a number of standard texts on the theory of finite element methods, e.g., [3, 81]:

Lemma 5.7. *Let $\mathcal{T}_h$ be a uniform finite element mesh over D with $K_h = \mathcal{O}(h^{-n})$ degrees of freedom and $h > 0$. For the elliptic PDE (2.2) and $\mathbf{y} \in \mathcal{U}$, when $u^r(\mathbf{y}) \in H_0^1(D) \cap H^{s+1}(D)$, the error from the finite element approximation is bounded by*

$$\|u^r(\mathbf{y}) - u_h^r(\mathbf{y})\|_{H_0^1(D)} \leq C_{\text{FEM}} h^s,$$

where the constant $C_{\text{FEM}} > 0$ is independent of h and $\mathbf{y}$.

For the treatment of SG(IV), we begin by defining $\mathcal{B}^r(\mathbf{y})$ to be the corresponding bilinear operator in (2.10) with $a(x, \mathbf{y})$ replaced with $a^r(x, \mathbf{y})$. Since both $\mathcal{B}(\mathbf{y})$ and $\mathcal{B}^r(\mathbf{y})$ are symmetric, uniformly coercive and continuous bilinear operators on $H_0^1(D)$, there exist $\alpha, \beta > 0$ independent of $\mathbf{y}$ such that for every $u, v \in H_0^1(D)$

$$\begin{aligned} |\mathcal{B}^r[u, v](\mathbf{y})| &= \left| \int_D a^r(x, \mathbf{y}) \nabla u \cdot \nabla v dx \right| \leq \alpha \|u\|_{H_0^1(D)} \|v\|_{H_0^1(D)}, \quad \text{and} \\ \beta \|u\|_{H_0^1(D)}^2 &\leq \int_D a^r(x, \mathbf{y}) |\nabla u|^2 dx = \|u\|_{\mathcal{B}^r(\mathbf{y})}^2, \end{aligned}$$

and similarly for $\mathcal{B}_r(\mathbf{y})$ with the same α, β , e.g., taking α to be the maximum and β to be the minimum in each case. Recall $\mathbf{u}_{h, \nu}^r = [u_{1, \nu}^r, \dots, u_{K_h, \nu}^r]^T$, the vector of nodal values of the finite element solution corresponding to the ν -th stochastic mode of $u_{h, p}^r$, and $\mathbf{u}_{h, p}^r = [\mathbf{u}_{h, \nu}^r]_{\nu \in \mathcal{J}_p^{\text{TD}}}^T$. Then we have the following estimates expressing

$$\text{Continuity:} \quad \|\mathbf{u}_{h, p}^r\|_{\mathbf{K}_r} = \|u_{h, p}^r\|_{\mathbb{E}[\mathcal{B}^r(\mathbf{y})]} \leq \sqrt{\alpha} \|u_{h, p}^r\|_{\mathcal{H}_\varepsilon^2}, \quad \text{and} \quad (5.18)$$

$$\text{Ellipticity:} \quad \sqrt{\beta} \|u_{h, p}^r\|_{\mathcal{H}_\varepsilon^2} \leq \|u_{h, p}^r\|_{\mathbb{E}[\mathcal{B}^r(\mathbf{y})]} = \|\mathbf{u}_{h, p}^r\|_{\mathbf{K}_r}, \quad (5.19)$$

where $\|\mathbf{u}\|_{\mathbf{K}_r}^2 = \mathbf{u}^T \mathbf{K}_r \mathbf{u}$ is the $\mathbf{K}_r$ matrix norm, and $\|u\|_{\mathbb{E}[\mathcal{B}^r(\mathbf{y})]}$ is the expectation of the energy norm (2.11). Given Proposition 5.6, Lemma 5.7, and the estimates from (5.17), (5.18), and (5.19), we can now provide the minimal projection orders $p, r \in \mathbb{N}$ for the SG approximation (3.10) and the coefficient (3.16), respectively, the maximum mesh size h for finite element method, and the minimum number of PCG iterations k necessary to ensure that the error in the SGFEM solution $\tilde{u}_{h, p}^r$ is less than the tolerance $\varepsilon > 0$.

Lemma 5.8. *Let $u \in L_\varepsilon^2(\mathcal{U}; H_0^1(D) \cap H^{s+1}(D))$ be the solution to (2.2), $u_{h, p}^r$ be the solution to (3.11) with the coefficient $a^r(x, \mathbf{y})$, and $\tilde{u}_{h, p}^r$ be the approximation of $u_{h, p}^r$ found by PCG with a zero initial guess. Then, for $\varepsilon > 0$, to ensure that $\|u - \tilde{u}_{h, p}^r\|_{\mathcal{H}_\varepsilon^2} \leq \varepsilon$ we must choose*

$h \leq h_{\max}$, $r \geq r_{\min}$, $p \geq p_{\min}$, and $k \geq k_{\min}$, where:

$$\begin{aligned} h_{\max} &= \left(\frac{\varepsilon}{4C_{\text{FEM}}} \right)^{\frac{1}{s}}, & r_{\min} &= \log \left[\left(\frac{4C_5}{\varepsilon} \right)^{\frac{1}{C_4}} \right], \\ p_{\min} &= \log \left[\left(\frac{4C_1}{\varepsilon} \right)^{\frac{1}{C_2}} \right], & k_{\min} &= \frac{\log \left(\frac{4C_6}{\varepsilon} \right)}{\log \left(\frac{\sqrt{\tilde{\kappa}_r} + 1}{\sqrt{\tilde{\kappa}_r} - 1} \right)}, \end{aligned}$$

with $C_{\text{FEM}} > 0$ the constant from Lemma 5.7, $C_1, C_2, C_3, C_4 > 0$ the constants from Proposition 5.6 and (5.16), and, for $\alpha, \beta > 0$ from (5.18) and (5.19)

$$C_5 = C_3 \frac{\|f\|_{H^{-1}}}{a_{\min}^2}, \quad C_6 = 2\sqrt{\frac{\alpha}{\beta}} \|u_{h,p}^r\|_{\mathcal{H}_\varrho^2},$$

with $\tilde{\kappa}_r$ is the condition number of $\tilde{\mathbf{P}}^{-1}\mathbf{K}_r$ with $\mathbf{P}$ the mean-based preconditioner from (5.2).

Proof. Without loss of generality, we seek to bound the quantities SG(I)-SG(IV) from (5.15) each by $\varepsilon/4$. For the error SG(I) we recall estimate (5.17) and solve for r . From Lemma 5.7, when $u \in L_\varrho^2(\mathcal{U}; H_0^1(D) \cap H^{s+1}(D))$ we have that $\|u^r - u_h^r\|_{\mathcal{H}_\varrho^2} \leq C_{\text{FEM}} h^s \forall h > 0$, and from Proposition 5.6 we have that $\|u_h^r - u_{h,p}^r\|_{\mathcal{H}_\varrho^\infty} \leq C_1 \exp(-C_2 p) \forall p \in \mathbb{N}$, so that solving for h and p gives the desired maximum mesh size $h_{\max}$ and minimum polynomial order $p_{\min}$ to bound SG(II) and SG(III) by $\varepsilon/4$. Let $\mathbf{u}_{h,p}^r$ and $\mathbf{u}_{h,p}^{r,(k)}$ be the coefficients of the exact SG solution $u_{h,p}^r$ and the approximate SG solution $\tilde{u}_{h,p}^r$ after k PCG iterations, respectively. Then from (5.1) and (5.19) we see that

$$\|u_{h,p}^r - \tilde{u}_{h,p}^r\|_{\mathcal{H}_\varrho^2} \leq \frac{1}{\sqrt{\beta}} \|\mathbf{u}_{h,p}^r - \mathbf{u}_{h,p}^{r,k}\|_{\mathbf{K}_r} \leq \frac{2}{\sqrt{\beta}} \left(\frac{\sqrt{\tilde{\kappa}_r} - 1}{\sqrt{\tilde{\kappa}_r} + 1} \right)^k \|\mathbf{u}_{h,p}^r - \mathbf{u}_{h,p}^{r,0}\|_{\mathbf{K}_r},$$

where $\mathbf{u}_{h,p}^{r,0}$ is the initial guess used in CG and $\tilde{\kappa}_r = \text{cond}(\tilde{\mathbf{P}}^{-1}\mathbf{K}_r)$ with mean based preconditioner $\mathbf{P}$ from (5.2). If we use the zero vector as the initial iteration in PCG, we have from (5.18)

$$\|u_{h,p}^r - \tilde{u}_{h,p}^r\|_{\mathcal{H}_\varrho^2} \leq \frac{2}{\sqrt{\beta}} \left(\frac{\sqrt{\tilde{\kappa}_r} - 1}{\sqrt{\tilde{\kappa}_r} + 1} \right)^k \|\mathbf{u}_{h,p}^r\|_{\mathbf{K}_r} \leq 2\sqrt{\frac{\alpha}{\beta}} \left(\frac{\sqrt{\tilde{\kappa}_r} - 1}{\sqrt{\tilde{\kappa}_r} + 1} \right)^k \|u_{h,p}^r\|_{\mathcal{H}_\varrho^2}. \quad (5.20)$$

Solving for k gives the minimum number of iterations $k_{\min}$ required to ensure SG(IV) is bounded by $\varepsilon/4$. ■

Given the necessary parameters from Lemma 5.8 to achieve $\|u - \tilde{u}_{h,p}^r\|_{\mathcal{H}_\varrho^2} \leq \varepsilon$, and the estimates on the computational complexity of one iteration in the PCG method from §5.1.2, we provide a bound on the minimal number of FLOPs required by the SG method with FE when approximating (2.2). We split these results into the cases that the stochastic coefficient $a(x, \mathbf{y})$ from (2.2) is:

- (i) an affine function of the random parameters, e.g., $a(x, \mathbf{y}) \in P_{\mathcal{J}_{\bar{r}}^{\text{TD}}}(\mathcal{U})$ with $\bar{r} = 1$, as in Example 2.1,

- (ii) a non-affine polynomial of the random parameters, e.g., $a(x, \mathbf{y}) \in P_{\mathcal{J}_{\bar{r}}^{\text{TD}}}(\mathcal{U})$ for some $1 < \bar{r} < \infty$, as in Example 2.2,
- (iii) a non-affine, transcendental function of the random parameters, e.g., $a(x, \mathbf{y}) \notin P_{\mathcal{J}_{\bar{r}}^{\text{TD}}}(\mathcal{U})$ for any $r \in \mathbb{N}$, as in Example 2.3, so that r must be chosen to satisfy $r \geq r_{\min}$ from Lemma 5.8.

The results are summarized in Theorems 5.9 and 5.10 next.

Theorem 5.9. *Let $u \in L^2_0(\mathcal{U}; H_0^1(D) \cap H^{s+1}(D))$ be the solution to (2.2), and $\bar{r}$ be the smallest natural number such that $a(x, \mathbf{y}) \in \mathcal{P}_{\mathcal{J}_{\bar{r}}^{\text{TD}}}(\mathcal{U})$. When $\bar{r} = 1$, the minimum work (5.5) of solving (3.21) with PCG to a tolerance $\varepsilon > 0$ can be bounded by*

$$W^{\text{pSG}} \leq C_7 \left(\frac{3C_{\text{FEM}}}{\varepsilon} \right)^{\frac{n}{s}} 2e^d(1+d) \left(1 + \log \left[\left(\frac{3C_1}{\varepsilon} \right)^{\frac{1}{C_2 d}} \right] \right)^d \left(\frac{\log \left(\frac{3C_6}{\varepsilon} \right)}{\log \left(\frac{\sqrt{\tilde{\kappa}}+1}{\sqrt{\tilde{\kappa}}-1} \right)} \right), \quad (5.21)$$

and when $\bar{r} > 1$, the minimum work (5.5) of solving (3.21) with PCG to a tolerance $\varepsilon > 0$ can be bounded by

$$\begin{aligned} W^{\text{pSG}} &\leq C_7 \left(\frac{3C_{\text{FEM}}}{\varepsilon} \right)^{\frac{n}{s}} \left(\frac{\log \left(\frac{3C_6}{\varepsilon} \right)}{\log \left(\frac{\sqrt{\tilde{\kappa}}+1}{\sqrt{\tilde{\kappa}}-1} \right)} \right) e^d \left[\left(1 + \log \left[\left(\frac{3C_1}{\varepsilon} \right)^{\frac{1}{C_2 d}} \right] \right)^d \right. \\ &\quad \left. + 2 \sum_{j=0}^{\bar{r}} \binom{d-1+j}{d-1} \min \left\{ 2^j, \binom{d+\lceil j/2 \rceil}{d} \right\} \left(1 - \frac{\lceil j/2 \rceil}{d} + \log \left[\left(\frac{3C_1}{\varepsilon} \right)^{\frac{1}{C_2 d}} \right] \right)^d \right], \end{aligned} \quad (5.22)$$

with $C_{\text{FEM}}, C_1, C_2, C_6$ as in Lemma 5.8, $C_7 > 0$ independent of ε , and $\tilde{\kappa}$ the condition number of the preconditioned system $\tilde{\mathbf{P}}^{-1}\mathbf{K}_{\bar{r}} = \tilde{\mathbf{P}}^{-1}\mathbf{K}$, using the mean-based preconditioner from (5.2).

Proof. When $a(x, \mathbf{y}) \in P_{\mathcal{J}_{\bar{r}}^{\text{TD}}}(\mathcal{U})$ we do not need to consider SG(I) from (5.15), and bound SG(II), SG(III), and SG(IV) by $\varepsilon/3$. Hence, to minimize the error of the SG discretization, we choose $p \geq p_{\min} = \log[(3C_1/\varepsilon)^{1/C_2}]$ which differs from the $p_{\min}$ stated in Lemma 5.8. For a uniform triangulation $\mathcal{T}_h$, $K_h = \mathcal{O}(h^{-n})$ so that

$$K_{h_{\max}} = C_7 \left[\left(\frac{\varepsilon}{3C_{\text{FEM}}} \right)^{\frac{1}{s}} \right]^{-n} = C_7 \left(\frac{3C_{\text{FEM}}}{\varepsilon} \right)^{\frac{n}{s}} \quad (5.23)$$

for some constant $C_7 > 0$ depending on the connectivity of the finite element mesh, but independent of ε . In the case that $\bar{r} = 1$, we substitute $p_{\min}$ into (5.6) for the matrices $\mathbf{G}_{\alpha}$

having $0 \leq |\alpha| \leq 1$, and apply Stirling's approximation to obtain

$$\begin{aligned} N_{p_{\min}} + \mathcal{M}(p_{\min}, 1) &= 2 \binom{d + p_{\min}}{d} + 2d \binom{d + p_{\min} - 1}{d} \\ &\leq 2e^d(1 + d) \left(1 + \log \left[\left(\frac{3C_1}{\varepsilon} \right)^{\frac{1}{C_2 d}} \right] \right)^d, \end{aligned}$$

Similarly, when $\bar{r} > 1$ we use the bound from (5.9) and Stirling's approximation to obtain

$$\begin{aligned} N_{p_{\min}} + \mathcal{M}(p_{\min}, \bar{r}) &\leq e^d \left[\left(1 + \log \left[\left(\frac{3C_1}{\varepsilon} \right)^{\frac{1}{C_2 d}} \right] \right)^d \right. \\ &\quad \left. + 2 \sum_{j=0}^{\bar{r}} \min \left\{ 2^j, \binom{d + \lceil j/2 \rceil}{d} \right\} \binom{d - 1 + j}{d - 1} \left(1 - \frac{\lceil j/2 \rceil}{d} + \log \left[\left(\frac{3C_1}{\varepsilon} \right)^{\frac{1}{C_2 d}} \right] \right)^d \right]. \end{aligned}$$

Substituting $K_{h_{\max}}$ for K_h , $k_{\min}$ for $k_{\text{iter}}^{\text{SG}}$ from Lemma 5.8, and the bounds for $N_{p_{\min}} + \mathcal{M}(p_{\min}, \bar{r})$ into the work estimate (5.5), in the cases $\bar{r} = 1$ and $\bar{r} > 1$ above, we obtain the desired results. ■

Theorem 5.10. *Let $u \in L^2_{\mathcal{U}}(\mathcal{U}; H^1_0(D) \cap H^{s+1}(D))$ be the solution to (2.2), and suppose that $a(x, \mathbf{y}) \notin P_{\mathcal{J}_r^{\text{TD}}}(\mathcal{U})$ for any $r \in \mathbb{N}$. In this case r must be chosen to satisfy $r \geq r_{\min}$ from Lemma 5.8. Then the minimum work (5.5) of solving (3.21) with PCG to a tolerance $\varepsilon > 0$ can be bounded by*

$$\begin{aligned} W^{\text{pSG}} &\leq C_7 \left(\frac{4C_{\text{FEM}}}{\varepsilon} \right)^{\frac{n}{s}} e^d \left(1 + \log \left[\left(\frac{4C_1}{\varepsilon} \right)^{\frac{1}{C_2 d}} \right] \right)^d \\ &\quad \left[1 + 2e^{2d} \left(1 + \log \left[\left(\frac{4C_5}{\varepsilon} \right)^{\frac{1}{C_4 d}} \right] \right)^d \left(1 + \frac{1}{d} + \log \left[\left(\frac{4C_5}{\varepsilon} \right)^{\frac{1}{2C_4 d}} \right] \right)^d \right] \left(\frac{\log \left(\frac{4C_6}{\varepsilon} \right)}{\log \left(\frac{\sqrt{\tilde{\kappa}_r} + 1}{\sqrt{\tilde{\kappa}_r} - 1} \right)} \right), \end{aligned} \quad (5.24)$$

with $C_{\text{FEM}}, C_1, C_2, C_4, C_5, C_6$ as in Lemma 5.8, $C_7 > 0$ independent of ε , and $\tilde{\kappa}_r$ the condition number of the preconditioned system $\tilde{\mathbf{P}}^{-1}\mathbf{K}_r$, using the mean-based preconditioner from (5.2).

Proof. In this setting r must be chosen to satisfy $r \geq r_{\min} = \log[(4C_5/\varepsilon)^{1/C_4}]$ from Lemma 5.8 and, therefore, we must bound the sum from (5.9) which now depends on $r_{\min}$, and hence on ε . Thus, we use the bound (5.11) for $\mathcal{M}(p, r)$, noting that as $\varepsilon \rightarrow 0$, $r_{\min} \rightarrow \infty$ so that

$$\min \left\{ 2^{r_{\min}}, \binom{d + \lceil r_{\min}/2 \rceil}{d} \right\} = \binom{d + \lceil r_{\min}/2 \rceil}{d}.$$

Substituting $p_{\min}$ and $r_{\min}$ from Lemma 5.8 into (5.11) and applying Stirling's approximation, we obtain

$$N_{p_{\min}} + \mathcal{M}(p_{\min}, r_{\min}) \leq e^d \left(1 + \log \left[\left(\frac{4C_1}{\varepsilon} \right)^{\frac{1}{C_2 d}} \right] \right)^d \left[1 + 2e^{2d} \left(1 + \log \left[\left(\frac{4C_5}{\varepsilon} \right)^{\frac{1}{C_4 d}} \right] \right)^d \left(1 + \frac{1}{d} + \log \left[\left(\frac{4C_5}{\varepsilon} \right)^{\frac{1}{2C_4 d}} \right] \right)^d \right].$$

As in the proof of Theorem 5.9, we substitute $K_{h_{\max}}$ for K_h from (5.23), $k_{\min}$ for $k_{\text{iter}}^{\text{SG}}$, and the bound for $N_{p_{\min}} + \mathcal{M}(p_{\min}, r_{\min})$ with $p_{\min}$ and $k_{\min}$ from Lemma 5.8 into the cost (5.5) to complete the proof. ■

Given Theorems 5.9 and 5.10 we see that the work of obtaining the fully discrete approximation using the SGFEM, with PCG as a solver, is asymptotically given by:

$$\underbrace{\mathcal{O} \left(\frac{1}{\varepsilon} \right)^{\frac{n}{s}}}_{(\text{SG.1})} \underbrace{\left[\log \left(\frac{1}{\varepsilon} \right) \right]^{g(d)}}_{(\text{SG.2})} \underbrace{\left(\frac{\log \left(\frac{1}{\varepsilon} \right)}{\log \left(\frac{\sqrt{\tilde{\kappa}_r} + 1}{\sqrt{\tilde{\kappa}_r} - 1} \right)} \right)}_{(\text{SG.3})}, \quad (5.25)$$

where $g(d) = d$ and $\tilde{\kappa}_r = \tilde{\kappa}$ if $a(x, \mathbf{y})$ is an affine or non-affine, polynomial function of the random parameters of fixed order $\bar{r} < \infty$, e.g., Examples 2.1 and 2.2, and $g(d) = 3d$ when $a(x, \mathbf{y})$ is a non-affine, transcendental function of the random parameters, e.g., Example 2.3, requiring a total degree orthogonal expansion of order $r \geq r_{\min}$ depending on ε . Here, (SG.1), (SG.2), and (SG.3) correspond to the work required by the finite element, SG, and PCG methods, respectively. In particular, (SG.2) corresponds to the estimates for the sparsity of the Galerkin system $\mathbf{K}_r$ from (3.21), and represents the number of coupled finite element systems that must be solved simultaneously by the PCG method. However, due to the bound (5.9), the asymptotic complexity in the cases that $a(x, \mathbf{y})$ is affine or polynomial in $\mathbf{y}$ are the same. This does not imply that there is no need to consider the work estimates in these cases separately. Indeed, if $a(x, \mathbf{y})$ is a polynomial having the representation $\sum_{\alpha \in \mathcal{J}_{\bar{r}}^{\text{TD}}} a_{\alpha}(x) \Psi_{\alpha}(\mathbf{y})$ where $a_{\alpha}(x) \neq 0$ for all $\alpha \in \mathcal{J}_{\bar{r}}^{\text{TD}}$, then the complexity of matrix-vector multiplications with $\mathbf{K}_r$ is of the order $\mathcal{O}(K_h N_p N_{\bar{r}} \min\{2^{\bar{r}}, N_{\lceil \bar{r}/2 \rceil}\})$. Here, the constant $N_{\bar{r}} \min\{2^{\bar{r}}, N_{\lceil \bar{r}/2 \rceil}\}$ grows rapidly with $\bar{r}$, suggesting that high order polynomial functions of $\mathbf{y}$ incur additional cost.

5.1.4 Conditioning aspects of the generalized SG system

In this section, we discuss issues related to the conditioning of the linear system that results from the SGFEM discretization. We first recall [55, Theorem 10]: the eigenvalues of the matrices $(\mathbf{G}_\alpha)_{\alpha \in \mathcal{J}_r^{\text{TD}}}$ from (3.20) lie in the interval $[\xi_\alpha, \Xi_\alpha]$, where

$$\xi_\alpha := \min\{\Psi_\alpha(\mathbf{y}) : \mathbf{y} \in \mathcal{G}^{m(\mathbf{l})}\}, \quad \Xi_\alpha := \max\{\Psi_\alpha(\mathbf{y}) : \mathbf{y} \in \mathcal{G}^{m(\mathbf{l})}\}, \quad (5.26)$$

$\mathcal{G}^{m(\mathbf{l})}$ is a tensor product grid of Gauss-Legendre quadrature points having $m(\mathbf{l}) = (m(l_1), \dots, m(l_n))$ points in each direction, and $\mathbf{l}$ is such that $m(l_n) := p + \lceil \frac{k_n+1}{2} \rceil$, $n = 1, \dots, d$. Since $a^r(x, \mathbf{y})$ satisfies (A1), the analysis of [99, Theorem 3.8] shows that the eigenvalues for the preconditioned system $\mathbf{P}^{-1}\mathbf{K}_r$ lie in the interval $[1 - \underline{\tau}_r, 1 + \bar{\tau}_r]$ where

$$\underline{\tau}_r = \frac{1}{a_{\min}} \sum_{\substack{\alpha \in \mathcal{J}_r^{\text{TD}} \\ |\alpha| \neq 0}} \xi_\alpha \|a_\alpha(x)\|_{L^\infty(D)}, \quad \bar{\tau}_r = \frac{1}{a_{\min}} \sum_{\substack{\alpha \in \mathcal{J}_r^{\text{TD}} \\ |\alpha| \neq 0}} \Xi_\alpha \|a_\alpha(x)\|_{L^\infty(D)}. \quad (5.27)$$

As a result of (5.27), we see that in the case that the projection order r of the coefficient $a^r(x, \mathbf{y})$ depends on ε , the condition number of the preconditioned system $\mathbf{P}^{-1}\mathbf{K}_r$ does as well through the number of terms in $\underline{\tau}_r$ and $\bar{\tau}_r$. This should come as no surprise since even in the case of the Karhunen-Loève expansion, the condition number of $\mathbf{P}^{-1}\mathbf{K}$ depends on the number of terms in the truncated Karhunen-Loève expansion which is chosen a-priori to minimize the error.

5.2 Complexity of stochastic collocation method

In this section, we present results from [60] on the computational complexity of solving (2.2) with the stochastic collocation method from Section 3.4. Similar to the previous section, these results are presented in terms of total number of FLOPs required to solve the decoupled FE systems at the required points $(\mathbf{y}_i)_{i=1}^{M_L}$. Section 5.2.1 develops the cost metric used to analyze computational complexity. Section 5.2.2 presents background results from approximation theory, comparing the degrees of freedom required to construct the SC and SG approximations through the Lebesgue constants associated with their construction. We conclude this section with a comparison of the ε -complexity results derived for both methods and a consideration of the impact of the conditioning in the SG systems in comparing cost.

5.2.1 Cost of solving the SC systems

To construct the fully discrete SC approximation with the FE method, we must solve M_L distinct decoupled FE systems, each dependent on a realization of the parameters $\mathbf{y}_i \in \mathcal{G}_L^{m,g}$ for $i = 1, \dots, M_L$. Similar to the SG method with FE, we can apply the PCG method to the solution of each system. Let $k_{\text{iter}}^{(i)}$ be the number of iterations required by the CG method to solve the finite element system at a particular $\mathbf{y}_i$ and $k_{\text{iter}}^{\text{p}(i)}$ be the corresponding number of iterations when a preconditioner is used. We are interested in choosing a suitable preconditioning strategy to decrease the total number of iterations $k_{\text{iter}}^{\text{pSC}} = \sum_{i=1}^{M_L} k_{\text{iter}}^{\text{p}(i)}$ required

to obtain the fully discrete approximation $u_{h,L}$. We present a preconditioning strategy of choosing

$$\mathbf{P}_0 := \mathbf{B}(\mathbf{y}_1), \quad (5.28)$$

with $\mathbf{B}(\mathbf{y})$ from (3.3) the finite element stiffness matrix corresponding to the sample point $\mathbf{y}_1 \in \mathcal{G}_L^{\mathbf{m},g}$, as the preconditioner for all of the individual finite element solutions. We refer to this choice of preconditioner as the level-zero preconditioner since it corresponds to the SC approximation at level $L = 0$.

Since we apply CG to the solution of each individual finite element system, the work in FLOPs required to obtain a fully discrete approximation with the SC and FE methods without a preconditioner is given by

$$W^{\text{SC}} \approx \mathcal{O}(K_h) * \sum_{i=1}^{M_L} k_{\text{iter}}^{(i)}. \quad (5.29)$$

On the other hand, the “level-zero” preconditioner requires an additional matrix-vector product with complexity $\mathcal{O}(K_h)$ FLOPs per iteration when a sparse factorization of $\mathbf{P}_0$ is used. Hence the work of solving (2.9) with PCG is given by

$$W^{\text{pSC}} \approx 2 * \mathcal{O}(K_h) * \sum_{i=1}^{M_L} k_{\text{iter}}^{\text{p}(i)}. \quad (5.30)$$

Here, the reduction in work due to preconditioning will be seen in the number of iterations saved in each individual count $k_{\text{iter}}^{\text{p}(i)}$ contributing to the sum.

5.2.2 Comparing the explicit cost bounds of the SG and SC methods

Given a particular index set $\mathcal{J}_p$, we can find increasing functions $\mathbf{m} : \mathbb{N}_+^d \rightarrow \mathbb{N}_+^d$ and $g : \mathbb{N}_+^d \rightarrow \mathbb{N}_+$, and $L \in \mathbb{N}$ such that $\mathcal{J}_p = \mathcal{J}_L^{\mathbf{m},g}$ from (3.24). In this setting, we can either use Galerkin projection or construct an interpolant to obtain an approximation to u in $P_{\mathcal{J}_p}(\mathcal{U})$. Let $u_{\mathcal{J}_p}$ denote the Galerkin projection of u onto the space $P_{\mathcal{J}_p}(\mathcal{U})$. Then we have the estimate

$$\|u - u_{\mathcal{J}_p}\|_{L_\varrho^2(\mathcal{U}; H_0^1(D))} \leq C_a \min_{v \in H_0^1(D) \otimes P_{\mathcal{J}_p}(\mathcal{U})} \|u - v\|_{L_\varrho^2(\mathcal{U}; H_0^1(D))}$$

where $C_a > 0$ depends on the coefficient $a(x, \mathbf{y})$ and the bounds from assumption (A1). This estimate expresses optimality in the $L_\varrho^2(\mathcal{U})$ error of the Galerkin projection since C_a does not grow with $\mathcal{J}_p$, and suggests that the Galerkin method is the best choice for approximating u in the space $P_{\mathcal{J}_p}(\mathcal{U})$. We can also define an interpolation operator $\mathcal{I}_L^{\mathbf{m},g} : C^0(\mathcal{U}) \rightarrow P_{\mathcal{J}_p}(\mathcal{U})$, and then we have the estimate

$$\begin{aligned} \|u - \mathcal{I}_L^{\mathbf{m},g}[u]\|_{L_\varrho^\infty(\mathcal{U}; H_0^1(D))} &\leq (1 + \mathbb{L}_{\mathcal{J}_L}) \min_{v \in H_0^1(D) \otimes P_{\mathcal{J}_p}(\mathcal{U})} \|u - v\|_{L_\varrho^\infty(\mathcal{U}; H_0^1(D))} \\ &= (1 + \mathbb{L}_{\mathcal{J}_L}) \|u - u_{\mathcal{J}_p}\|_{L_\varrho^\infty(\mathcal{U}; H_0^1(D))} \end{aligned} \quad (5.31)$$

where $\mathbb{L}_{\mathcal{J}_L}$ is the Lebesgue constant of $\mathcal{I}_L^{\mathbf{m},g}$. A good interpolant will be one for which $\mathbb{L}_{\mathcal{J}_L}$ grows moderately with $\#\mathcal{J}_L^{\mathbf{m},g}$. For example, it is known (see [51, 60]) that for a one-dimensional Lagrange interpolation operator using a Clenshaw-Curtis rule, the Lebesgue constant is bounded by $\frac{2}{\pi} \log(m-1) + 1$, where m is the number of points. For the SC method, we define SDOF to be the total number of points needed to construct the approximation. From (5.31), if we only consider the number of SDOF needed to represent the solution, we expect the error for the Galerkin approximation to be much lower than the error in the interpolant. Indeed, this is reflected in our numerical results in Figures 6.2 and 6.4, and has been observed in previous comparisons [7, 54].

However, if we are willing to change the space $\mathcal{J}_p$, e.g., adding more interpolation points to gain a more stable interpolant by changing $\mathbf{m}$ or changing which points are required by the set $\mathcal{J}_L^{\mathbf{m},g}$ through changing g , it might be possible to obtain an approximation with lower complexity to reach a given tolerance, despite having to solve more systems. Therefore, to properly compare the work involved in constructing both approximations $u_{\mathcal{J}_p}$ and $\mathcal{I}_L^{\mathbf{m},g}[u]$, we consider the computational complexity of both methods, not in terms of SDOF, but in terms of FLOPs. For a specific $\mathcal{J}_p$, this reduces to studying the complexity of the system resulting from Galerkin projections and the stability properties of the interpolant $\mathcal{I}_L^{\mathbf{m},g}$.

Let $\tilde{u}_{h,L}$ denote the numerical solution to the fully discrete approximation $u_{h,L}$ obtained with the SC and FE methods from (3.27) approximated by the PCG solver, and observe that we have a similar splitting to (5.15) for the error in the approximation

$$\|u - \tilde{u}_{h,L}\|_{\mathcal{H}_\epsilon^2} \leq \underbrace{\|u - u_h\|_{\mathcal{H}_\epsilon^2}}_{\text{SC(I)}} + \underbrace{\|u_h - u_{h,L}\|_{\mathcal{H}_\epsilon^2}}_{\text{SC(II)}} + \underbrace{\|u_{h,L} - \tilde{u}_{h,L}\|_{\mathcal{H}_\epsilon^2}}_{\text{SC(III)}} \quad (5.32)$$

Note that unlike in the case of the SG method with FE, the SC method does not require a further projection of the coefficient $a(x, \mathbf{y})$, so that we do not need to consider the error $\|u - u^r\|_{\mathcal{H}_\epsilon^2}$ from (5.15). In addition, we need not worry about well-posedness of the PDE after truncation, as discussed in Remark 3.4. Similar to the complexity analysis for the SG method with FE, we must choose $h \leq h_{\max}$ and $L \geq L_{\min}$ so that the errors $\|u - u_h\|_{\mathcal{H}_\epsilon^2}$ from the finite element discretization and $\|u_h - u_{h,L}\|_{\mathcal{H}_\epsilon^2}$ from the SC interpolation are both bounded by $\varepsilon/3$. From this, a minimum tolerance $\tau_{\min}$ for the PCG solver can be derived and the maximum number of PCG iterations, with a zero initial guess, can be estimated. In what follows, we present a result, whose proof can be found in [60, Theorem 7], bounding the number of PCG iterations in the context of the work estimate (5.30). Using this estimate we can compare the cost in FLOPs for both the SC and SG methods with the results from Theorems 5.9 and 5.10 in the previous section.

Theorem 5.11. *Let $u \in L_\epsilon^2(\mathcal{U}; H_0^1(D) \cap H^{s+1}(D))$ be the solution to (2.2). Then for $\varepsilon > 0$ arbitrary, the work of finding $\tilde{u}_{h,L}$, the approximation to the fully discrete SC solution $u_{h,L}$*

from (3.27) found by PCG, denoted by W^{pSC} , can be bounded by

$$W^{\text{pSC}} \leq 2C_7 \left(\frac{3C_{\text{FEM}}}{\varepsilon} \right)^{\frac{n}{s}} C_8 \left[\log \left(\frac{3C_{\text{SC}}}{\varepsilon} \right) \right]^d \left[C_9 + \frac{1}{\log 2} \log \log \left(\frac{3C_{\text{SC}}}{\varepsilon} \right) \right]^{d-1} \quad (5.33)$$

$$\times \frac{1}{\log \left(\frac{\sqrt{\bar{\kappa}}+1}{\sqrt{\bar{\kappa}}-1} \right)} \left\{ \log \left(\frac{C_{10}}{\varepsilon} \right) + C_{11} + 2d \log \log \left[\frac{1}{rd} \log \left(\frac{3C_{\text{SC}}}{\varepsilon} \right) \right] \right\}.$$

Here C_{FEM} from Lemma 5.7, C_7 from Theorem 5.9, and $C_8, C_9, C_{10}, C_{11}, C_{\text{SC}}$, and r from [60, Theorem 7] are positive constants independent of ε . Moreover, we define $\bar{\kappa} = \sup_{\mathbf{y} \in \mathcal{U}} \kappa(\mathbf{y})$ where $\kappa(\mathbf{y})$ is the condition number of the preconditioned system $\tilde{\mathbf{P}}_0^{-1} \mathbf{B}(\mathbf{y})$ with $\mathbf{P}_0$ from (5.28).

Theorem 5.11 follows from the fact that $k_{\text{iter}}^{\text{pSC}} = \sum_{i=1}^{M_L} k_{\text{iter}}^{\text{p}(i)} \leq k_{\text{zero}}$, where k_{zero} is the number of iterations needed by the SC method with a PCG and a zero vector initial guess. Substituting the bound on k_{zero} shown in [60, Theorem 7] into the work estimate (5.30) and using $K_{h_{\text{max}}} = C_7(3C_{\text{FEM}}/\varepsilon)^{n/s}$ as in Theorem 5.9, puts the result in terms of FLOPs. Given Theorem 5.11 we see that the work of obtaining the fully discrete approximation with the SC method with FE and the PCG solver is asymptotically bounded by:

$$\underbrace{\mathcal{O} \left(\frac{1}{\varepsilon} \right)^{\frac{n}{s}}}_{(\text{SC.1})} \underbrace{\left[\log \left(\frac{1}{\varepsilon} \right) \right]^d \left[\log \log \left(\frac{1}{\varepsilon} \right) \right]^{d-1}}_{(\text{SC.2})} \underbrace{\left(\frac{\log \left(\frac{1}{\varepsilon} \right)}{\log \left(\frac{\sqrt{\bar{\kappa}}+1}{\sqrt{\bar{\kappa}}-1} \right)} \right)}_{(\text{SC.3})} \quad (5.34)$$

where (SC.1), (SC.2), and (SC.3) correspond to the work required by the finite element, SC interpolant, and PCG methods, respectively. Since the costs associated with the finite element discretization are the same for both methods, we focus only on the costs associated with the SG projection and the SC interpolation, coupled with the costs of the PCG method. In particular, as the work required by the SG approximation from (SG.2) of (5.25) has different bounds depending on whether the coefficient is a fixed order polynomial or is given by a total degree orthogonal expansion having order depending on ε , we now provide a comparison in both of these cases.

Comparison in the affine and non-affine polynomial cases, e.g., Examples 2.1 and 2.2. The terms (SG.2) from (5.25) and (SC.2) from (5.34) are asymptotic estimates of the number of coupled and decoupled finite element systems that must be solved by the SG and SC methods to construct the stochastic approximation, respectively. For the coefficients from Examples 2.1 and 2.2, (SG.2) from (5.25) for the SG method with FE is $\mathcal{O}([\log(1/\varepsilon)]^d)$. Hence, our analysis shows that in these cases, the number of coupled finite element systems in the SG matrix (3.21) is smaller than the number of decoupled finite element systems required by the SC method by a factor of $\mathcal{O}([\log \log(1/\varepsilon)]^{d-1})$. This difference is enough to suggest that, if the condition numbers of both the preconditioned coupled system from the SG method and the preconditioned decoupled systems from the SC method are of the same order, then the SG method will outperform the SC method in terms of minimum work required to obtain a fully discrete approximation. In fact,

whenever $a(x, \mathbf{y})$ is a general non-affine, polynomial coefficient, e.g., Example 2.2, having fixed order $\bar{r} < \infty$, the complexity of matrix-vector products involving $\mathbf{K}$ from (3.14) is approximately $\mathcal{O}(K_h N_p N_{\bar{r}} \min\{2^{\bar{r}}, N_{\lceil \bar{r}/2 \rceil}\})$ from Remark 5.4. Hence, our analysis shows that when $\mathcal{O}(N_{\bar{r}} \min\{2^{\bar{r}}, N_{\lceil \bar{r}/2 \rceil}\}) < \mathcal{O}((\log \log(1/\varepsilon))^{d-1})$, which, in limit as $\varepsilon \rightarrow 0$, is always the case, and the condition numbers of both systems are of the same order, the SG method will outperform the SC method. However, in practical applications, it may require unrealistically small tolerance ε to see this when $\bar{r}$ is large.

Comparison in the non-affine, transcendental case, e.g., Example 2.3. In the case that $a(x, \mathbf{y})$ is a non-affine, transcendental function of the random parameters, e.g., Example 2.3, the estimate for (SG.2) is $\mathcal{O}([\log(1/\varepsilon)]^{3d})$. Here our analysis shows that the number of coupled finite element matrices present in the SG system $\mathbf{K}_r$ from (3.21) dominates the number of decoupled finite element systems needed by the SC method by a factor of $\mathcal{O}([\log(1/\varepsilon)]^{2d})$. Note that, as in the case of the SG method, the term (SC.3) has a dependence on the condition numbers of the preconditioned finite element systems through the bound $\bar{\kappa} = \sup_{\mathbf{y} \in \mathcal{U}} \kappa(\mathbf{y})$. For the unpreconditioned systems $\mathbf{B}(\mathbf{y})$, the condition numbers can be bounded by $\kappa(\mathbf{B}(\mathbf{y})) \leq (C_\kappa/h)^2$ for every $\mathbf{y} \in \mathcal{U}$, following from assumption (A1) and the quasi-uniformity of the mesh $\mathcal{T}_h$. However, if we use the exact inverse of $\mathbf{P}_0$ when preconditioning the FE systems for SC, the condition numbers are bounded independent of h and $\mathbf{y} \in \mathcal{U}$, since in this case $\bar{\kappa}$ is independent of mesh size h and level L . Hence the work required by PCG when solving the FE systems with the SC method is dependent on ε only through the term $\log(1/\varepsilon)$. On the other hand, if we use the exact inverse of $\mathbf{P}$ when preconditioning the SG system, the condition number $\tilde{\kappa}_r$ can be bounded by

$$\tilde{\kappa}_r \leq \frac{1 + \bar{\tau}_r}{1 - \underline{\tau}_r},$$

where $\underline{\tau}_r$ and $\bar{\tau}_r$ are defined in (5.27), hence depend on ε when r is chosen to satisfy $r \geq r_{\min}$ from Lemma 5.8. Figure 5.4 plots the condition numbers of both the unpreconditioned matrix $\mathbf{K}_r$ and the preconditioned matrix $\mathbf{P}^{-1}\mathbf{K}_r$ with decreasing finite element mesh parameter h for the coefficient $a(x, \mathbf{y})$ given in (6.7) from Section 6.1.3 with $d = 4$, $L_c = 1/2$, and letting $r = p$ with p increasing. There we see that the dependence on h has been removed by applying $\mathbf{P}^{-1}$, but as p increases, we see a corresponding increase in the condition number. Other preconditioners than $\mathbf{P}$ may be used to reduce the dependence on r , e.g. [116], but then their associated costs must be accounted for in the work estimate (5.5) as well. However, even if the condition numbers of both the preconditioned coupled SG system and the preconditioned decoupled SC systems are of the same order, the additional work required to solve the coupled systems induced by the nonlinearity of the coefficient makes it difficult to see how the SG method can compete with SC.

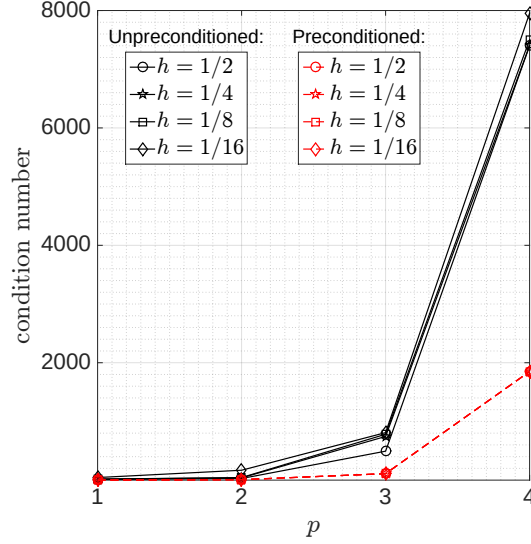


Figure 5.4: Condition numbers of both the unpreconditioned matrix $\mathbf{K}_r$ and the preconditioned matrix $\mathbf{P}^{-1}\mathbf{K}_r$ with decreasing finite element mesh parameter h and $r = p$ for the coefficient (6.7) from Section 6.1.3 with $d = 4$ and $L_c = 1/2$.

5.3 Complexity of the sparse Hilbert-valued recovery method

In this section, we provide a discussion of the complexity of the sparse Hilbert-valued recovery approach outlined in Section 3.6.1. To establish the results to follow, we provide a framework for the approximation of solutions to parameterized PDE problems. This framework relies on compressibility results, shown in prior works, for solutions of parameterized PDE problems of type (2.1) having parametric expansion as in (3.5). With these compressibility results, and *a priori* error estimates shown for quasi-optimal polynomial approximations in [113], we are able to derive convergence rates for such approximations, relevant to many practical finite-dimensional parametric PDE problems. The rest of this section is organized as follows. Section 5.3.1 discusses error estimates that have been shown for nonlinear best s -term and constructive quasi-optimal Galerkin approximations, and the types of bounds on the coefficients of (3.5) required for our analysis. Section 5.3.2 combines the quasi-optimal error estimates with the error estimates for the Hilbert-valued BPDN problem to establish the convergence with respect to number of samples for our approach.

5.3.1 Compressibility results for the parameterized PDE setting

As in the setting of standard compressive-sensing, we require sparsity (or compressibility) of the coefficients of the parametric expansion (3.5) for successful recovery. Of course, such sparsity results depend on the particular polynomial basis $(\Psi_\nu)_{\nu \in \mathcal{F}}$ and additional assumptions on the model (2.1). The landmark works [33, 34] established algebraic, dimension-independent, rates of convergence for best s -term Legendre and Taylor approximations of solutions to the parameterized elliptic PDE (2.2). Briefly, best s -term approximations seek to minimize the residual (3.8) by choosing the s largest coefficients $\mathbf{c}_\nu$ in norm $\|\cdot\|_\nu$. The analysis in those works required that the solution u depends affinely on countably many parameters $\mathbf{y}$ through the diffusion coefficient $a(x, \mathbf{y})$ in the elliptic operator. Subsequent studies have since extended that analysis to parameterized parabolic and hyperbolic PDEs, non-affine parameterizations, and also Chebyshev approximations of such systems [14, 25, 28, 71, 72, 75, 76, 77, 112]. In these works, the assumptions placed on the particular PDE guarantee that its solution is sufficiently regular with respect to the parameterization, enabling upper bounds and ℓ^p -summability results to be shown for the sequence $(\|\mathbf{c}_\nu\|_\nu)_{\nu \in \mathcal{F}}$. The Stechkin inequality (see, e.g., [43]) can then be applied to obtain estimates of the form

$$\|u - u_{\mathcal{J}_s}\|_{L^2_\theta(\mathcal{U}; \nu)} \leq \|(\|\mathbf{c}_\nu\|_\nu)_{\nu \in \mathcal{F}}\|_{\ell^p(\mathcal{F})} s^{1/2-1/p} \quad \forall s \in \mathbb{N}, \quad (5.35)$$

where $u_{\mathcal{J}_s}$ is the best s -term approximation to u and $0 < p < 1$ is such that $(\|\mathbf{c}_\nu\|_\nu)_{\nu \in \mathcal{F}}$ is ℓ^p -summable. In general, explicit estimates of the coefficient $\|(\|\mathbf{c}_\nu\|_\nu)_{\nu \in \mathcal{F}}\|_{\ell^p(\mathcal{F})}$ are unavailable without access to the sequence $(\mathbf{c}_\nu)_{\nu \in \mathcal{F}}$. Moreover, (5.35) often holds for a continuum of values of p , with stronger rates (smaller p) coupled with larger coefficients. Nonetheless, estimates of this form provide a useful theoretical benchmark of performance for best s -term approximations in the infinite-dimensional setting, and can be used to bound the overall approximation error of compressive recovery strategies for such problems. Recently, (5.35) was applied in [101] to the problems considered in [33, 34] to establish an error estimate, with explicit dependence on the number of samples $m \in \mathbb{N}$. More specifically, it was shown that

$$\|G - G_{\mathcal{J}}^\#\|_2 \leq C' \|\mathbf{g}\|_p s^{1/2-1/p} \leq C''_{\mathbf{g}} \left(\frac{\log^3(m) \log(N)}{m} \right)^{1/p-1/2}, \quad (5.36)$$

for compressive sensing Petrov-Galerkin approximations $G_{\mathcal{J}}^\#$ of the functionals $G(u(\mathbf{y}))$ discussed in Section 3.5.3, obtained by solving the single-sparse BPDN problem (3.60).

On the other hand, given bounds on the coefficients $\|\mathbf{c}_\nu\|_\nu$, one can construct the set $\mathcal{J}_s$ corresponding to the s largest of these bounds. The resulting approximations $u_{\mathcal{J}_s}$ are referred to as quasi-optimal approximations, see, e.g., [8, 9, 113]. Such approximations are relevant to a wide class of parameterized operator equations having finite-dimensional parametric dependence. In order to show fast rates of convergence of such approximations, the bounding functions must satisfy some additional assumptions. We repeat these assumptions here below. Consider a multi-indexed sequence of coefficient estimates on $(\|\mathbf{c}_\nu\|_\nu)_{\nu \in \mathcal{F}}$ written in

the form $(e^{-b(\boldsymbol{\nu})})_{\boldsymbol{\nu} \in \mathcal{F}}$, i.e., satisfying

$$\|\mathbf{c}_{\boldsymbol{\nu}}\|_{\mathcal{V}} \lesssim e^{-b(\boldsymbol{\nu})} \quad \forall \boldsymbol{\nu} \in \mathcal{F}. \quad (5.37)$$

We recall the following assumption from [113, Assumption 3] on the functions $b(\boldsymbol{\nu})$.

Assumption 5.12. *For all $\boldsymbol{\nu} \in \mathcal{F}$, define the map $H_{\boldsymbol{\nu}} : (0, \infty) \rightarrow \mathbb{R}$ as*

$$H_{\boldsymbol{\nu}}(\tau) = \frac{1}{\tau} b(\tau \boldsymbol{\nu}), \quad \forall \tau \in (0, \infty).$$

Then the maps $b : [0, \infty)^d \rightarrow \mathbb{R}$ and $H_{\boldsymbol{\nu}}$ satisfy

1. $b(\mathbf{0}) = 0$ and b is continuous on $\mathcal{F}$,
2. $H_{\boldsymbol{\nu}}$ is either increasing in $(0, \infty)$ for all $\boldsymbol{\nu} \in \mathcal{F}$ or decreasing in $(0, \infty)$ for all $\boldsymbol{\nu} \in \mathcal{F}$,
3. $b(\boldsymbol{\nu}) \in \Theta(|\boldsymbol{\nu}|)$. In other words, there exists $0 < c < C$ such that $c|\boldsymbol{\nu}| < b(\boldsymbol{\nu}) < C|\boldsymbol{\nu}|$ as $\boldsymbol{\nu} \rightarrow \infty$ in the Euclidean norm.

When b satisfies Assumption 5.12, and $\mathcal{J}_s$ is constructed from the s largest bounds of the sequence $(e^{-b(\boldsymbol{\nu})})_{\boldsymbol{\nu} \in \mathcal{F}}$, [113, Theorem 2] says that for any $\varepsilon > 0$ there exists an $s_{\varepsilon} > 0$ such that

$$\left\| u - \sum_{\boldsymbol{\nu} \in \mathcal{J}_s} \mathbf{c}_{\boldsymbol{\nu}} \Psi_{\boldsymbol{\nu}} \right\|_{L^2_{\theta}(\mathcal{U}; \mathcal{V})} \leq C_{\varepsilon} \sqrt{s} \exp \left[- \left(\frac{\kappa s}{(1 + \varepsilon)} \right)^{1/d} \right], \quad (5.38)$$

for every $s > s_{\varepsilon}$, where $\kappa, C_{\varepsilon} > 0$ are constants independent of s . Here κ corresponds to the volume of a limiting polytope $\mathcal{Q}$ bounding the sequence $(e^{-b(\boldsymbol{\nu})})_{\boldsymbol{\nu} \in \mathcal{F}}$. Optimal values of κ can be shown given explicit forms of the function $b(\boldsymbol{\nu})$, see, e.g., the results of [113, Propositions 4 & 5] where κ is derived in the cases of the Taylor and Legendre systems. This estimate is asymptotically sharp as $s \rightarrow \infty$, providing the optimal rate in the finite-dimensional setting, and carries an explicit dependence on the dimension. Coefficient bounds satisfying Assumption 5.12 have been shown for many parameterized PDE problems, including nonlinear elliptic PDEs, initial value problems, and parabolic and hyperbolic equations [28, 71, 72, 75, 77]. Therefore we expect similar results to hold in the finite parametric dimension versions of those problems.

5.3.2 Error estimates for Hilbert-valued recovery

With Propositions 3.19 & 3.20 and error estimates such as [113, Theorem 2] for quasi-optimal approximations, we are now able to provide convergence rates for approximations to parameterized PDEs obtained through the sparse Hilbert-valued recovery techniques of Section 3.6. The next theorem provides a benchmark for performance of sparse Hilbert-valued reconstruction in the general setting of solutions to problem (2.1) having parametric expansion with coefficients $(\mathbf{c}_{\boldsymbol{\nu}})_{\boldsymbol{\nu} \in \mathcal{F}}$ satisfying $\|\mathbf{c}_{\boldsymbol{\nu}}\|_{\mathcal{V}} \lesssim e^{-b(\boldsymbol{\nu})}$ for every $\boldsymbol{\nu} \in \mathcal{F}$, with $b(\boldsymbol{\nu})$ obeying Assumption 5.12.

Theorem 5.13. Let $\varepsilon > 0$ be arbitrary and assume that the solution u of (2.1) with parametric expansion (3.5) in terms of a bounded orthonormal system $(\Psi_\nu)_{\nu \in \mathcal{F}}$ has coefficients $(\mathbf{c}_\nu)_{\nu \in \mathcal{F}}$ satisfying $\|\mathbf{c}_\nu\|_{\mathcal{V}} \lesssim e^{-b(\nu)}$ for every $\nu \in \mathcal{F}$ with $b : [0, \infty)^d \rightarrow \mathbb{R}$ also satisfying Assumption 5.12. Denote by $\mathcal{J}_s$ the set of indices corresponding to the s largest bounds of the sequence $(e^{-b(\nu)})_{\nu \in \mathcal{F}}$ with s large enough such that (5.38) holds. Choose $\mathcal{J}$ such that $\mathcal{J}_s \subseteq \mathcal{J}$, and assume that the number of samples $m \in \mathbb{N}$ satisfies

$$m \geq C\Theta^2 s \max\{\log^2(\Theta^2 s) \log(N), \log(\Theta^2 s) \log(\log(\Theta^2 s) N^{\log(s)})\}, \quad (5.39)$$

as in (3.50), with Θ as in (3.34), $N = \#(\mathcal{J})$, and $\gamma = N^{-\log(s)}$. Then with probability $1 - N^{-\log(s)}$, the solution $\mathbf{c}^\#$ of

$$\text{minimize}_{\mathbf{z} \in \mathcal{V}^N} \|\mathbf{z}\|_{\mathcal{V},1} \quad \text{subject to} \quad \|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2} \leq \eta/\sqrt{m} \quad (5.40)$$

approximates u with error

$$\left\| u - \sum_{\nu \in \mathcal{J}} \mathbf{c}_\nu^\# \Psi_\nu \right\|_{L_\varrho^2(\mathcal{U}; \mathcal{V})} \leq \tilde{C}_\varepsilon \sqrt{s} \exp \left[- \left(\frac{\kappa s}{(1 + \varepsilon)} \right)^{1/d} \right], \quad (5.41)$$

where $\kappa, \tilde{C}_\varepsilon > 0$ are independent of s .

Proof. The requirement on m in (5.39) ensures the $\ell_{\mathcal{V},2}$ -RIP and hence $\ell_{\mathcal{V},2}$ -robust NSP holds for the normalized sampling matrix $\mathbf{A}$ as a consequence of Proposition 3.20. Hence, in the considered setting, recovery up to the best s term and truncation errors, as in estimate (3.82), occurs with probability $1 - N^{-\log(s)}$ when solving problem (5.40). All that remains is to combine the error estimate (5.38) for the quasi-optimal approximation $u_{\mathcal{J}_s}$ with the estimate for the Hilbert-valued BPDN problem under the $\ell_{\mathcal{V},2}$ -RIP. Let $u_{\mathcal{J}} = \sum_{\nu \in \mathcal{J}} \mathbf{c}_\nu \Psi_\nu$ as in (3.7) and $u_{\mathcal{J}}^\# = \sum_{\nu \in \mathcal{J}} \mathbf{c}_\nu^\# \Psi_\nu$ be the approximation obtained after solving (5.40). Then

$$\|u - u_{\mathcal{J}}^\#\|_{L_\varrho^2(\mathcal{U}; \mathcal{V})} = \|u - u_{\mathcal{J}} + u_{\mathcal{J}} - u_{\mathcal{J}}^\#\|_{L_\varrho^2(\mathcal{U}; \mathcal{V})} \leq \|u - u_{\mathcal{J}}\|_{L_\varrho^2(\mathcal{U}; \mathcal{V})} + \|u_{\mathcal{J}} - u_{\mathcal{J}}^\#\|_{L_\varrho^2(\mathcal{U}; \mathcal{V})}.$$

By Parseval's identity, we have

$$\|u - u_{\mathcal{J}}\|_{L_\varrho^2(\mathcal{U}; \mathcal{V})} = \left\| \sum_{\nu \in \mathcal{J}^c} \mathbf{c}_\nu \Psi_\nu \right\|_{L_\varrho^2(\mathcal{U}; \mathcal{V})} = \left(\sum_{\nu \in \mathcal{J}^c} \|\mathbf{c}_\nu\|_{\mathcal{V}}^2 \right)^{1/2} = \|\mathbf{c}_{\mathcal{J}^c}\|_{\mathcal{V},2} = \eta_{\mathcal{J}}/\sqrt{m},$$

where $\mathbf{c}_{\mathcal{J}^c} = (\mathbf{c}_\nu)_{\nu \in \mathcal{J}^c}$, and similarly

$$\|u_{\mathcal{J}} - u_{\mathcal{J}}^\#\|_{L_\varrho^2(\mathcal{U}; \mathcal{V})} = \left\| \sum_{\nu \in \mathcal{J}} (\mathbf{c}_\nu - \mathbf{c}_\nu^\#) \Psi_\nu \right\|_{L_\varrho^2(\mathcal{U}; \mathcal{V})} = \left(\sum_{\nu \in \mathcal{J}} \|\mathbf{c}_\nu - \mathbf{c}_\nu^\#\|_{\mathcal{V}}^2 \right)^{1/2} = \|\mathbf{c}_{\mathcal{J}} - \mathbf{c}_{\mathcal{J}}^\#\|_{\mathcal{V},2},$$

where $\mathbf{c}_{\mathcal{J}} = (\mathbf{c}_{\nu})_{\nu \in \mathcal{J}}$ and $\mathbf{c}_{\mathcal{J}}^{\#} = (\mathbf{c}_{\nu}^{\#})_{\nu \in \mathcal{J}}$. Under (5.39), we have

$$\|\mathbf{c}_{\mathcal{J}} - \mathbf{c}_{\mathcal{J}}^{\#}\|_{\mathcal{V},2} \leq \frac{C_1}{\sqrt{s}} \sigma_s(\mathbf{c}_{\mathcal{J}})_{\mathcal{V},1} + \frac{C_2}{\sqrt{m}} \eta_{\mathcal{J}},$$

where $C_1, C_2 > 0$ are the constants from Proposition 3.19. Grouping terms, we see that $u_{\mathcal{J}}^{\#}$ satisfies

$$\begin{aligned} \|u - u_{\mathcal{J}}^{\#}\|_{L_{\mathcal{Q}}^2(\mathcal{U};\mathcal{V})} &\leq \frac{C_1}{\sqrt{s}} \sigma_s(\mathbf{c}_{\mathcal{J}})_{\mathcal{V},1} + (1 + C_2) \frac{\eta_{\mathcal{J}}}{\sqrt{m}} \\ &\leq \frac{C_1}{\sqrt{s}} \left(\sum_{\nu \in \mathcal{J} \setminus \mathcal{J}_s} \|\mathbf{c}_{\nu}\|_{\mathcal{V}} \right) + (1 + C_2) \left(\sum_{\nu \in \mathcal{J}^c} \|\mathbf{c}_{\nu}\|_{\mathcal{V}}^2 \right)^{1/2} \\ &\leq \frac{C_1}{\sqrt{s}} \left(\sum_{\nu \in \mathcal{J}_s^c} e^{-b(\nu)} \right) + (1 + C_2) \left(\sum_{\nu \in \mathcal{J}_s^c} e^{-2b(\nu)} \right)^{1/2} \\ &\leq \frac{C_1}{\sqrt{s}} C_{\varepsilon} s \exp \left[- \left(\frac{\kappa_1 s}{(1 + \varepsilon)} \right)^{1/d} \right] + (1 + C_2) \sqrt{C_{\varepsilon} s} \left(\exp \left[- \left(\frac{\kappa_2 s}{(1 + \varepsilon)} \right)^{1/d} \right] \right)^{1/2}, \end{aligned}$$

where κ_1 and κ_2 are the volumes of the limiting polytopes $\mathcal{Q}_1$ and $\mathcal{Q}_2$ associated with the sequences $(e^{-b(\nu)})_{\nu \in \mathcal{F}}$ and $(e^{-2b(\nu)})_{\nu \in \mathcal{F}}$, respectively, satisfying $|\mathcal{Q}_2| = 2^d |\mathcal{Q}_1|$ so that $\kappa_2 = 2^d \kappa_1$, see [113, Theorem 2]. Hence, after substituting κ_2 and simplifying we see that $u_{\mathcal{J}}^{\#}$ satisfies

$$\|u - u_{\mathcal{J}}^{\#}\|_{L_{\mathcal{Q}}^2(\mathcal{U};\mathcal{V})} \leq \left(C_1 C_{\varepsilon} + (1 + C_2) \sqrt{C_{\varepsilon}} \right) \sqrt{s} \exp \left[- \left(\frac{\kappa_1 s}{(1 + \varepsilon)} \right)^{1/d} \right],$$

and therefore (5.41) holds with $\tilde{C}_{\varepsilon} := (C_1 C_{\varepsilon} + (1 + C_2) \sqrt{C_{\varepsilon}}) > 0$ and $\kappa = \kappa_1$ under the condition on the number of samples m from (5.39). ■

Remark 5.14 (Constants in Theorem 5.13 in the case of the orthonormal Legendre system and the parameterized elliptic PDE (2.2)). *From [113, Theorem 2] we observe that the constant κ in the above result depends on the size and shape of the quasi-optimal index set $\mathcal{J}_s$ while C_{ε} has explicit definition $C_{\varepsilon} = (4e + 4\varepsilon e - 2) \frac{e}{e-1}$. For many examples of $b(\nu)$ one can explicitly derive κ and prove the optimality of the above result, see, e.g., [113, Propositions 3-6] for the results in the cases of the Taylor and Legendre series. We repeat this derivation for the orthonormal Legendre system relevant to the sparse polynomial approximation methods of Sections 3.5 & 3.6, and the numerical results in Section 6.3. Given a vector $\gamma = (\gamma_i)_{1 \leq i \leq d}$ with $\gamma_i > 1 \ \forall i$, we define $\lambda = (\lambda_i)_{1 \leq i \leq d}$ such that*

$$\lambda_i = \log(\gamma_i) \quad \forall 1 \leq i \leq d.$$

Under Assumptions (A1) & (A2), we have Theorem 2.4, expressing the regularity of the solution of the parameterized elliptic PDE (2.2) with respect to $\mathbf{y} \in \mathcal{U}$. Moreover, we have the existence of a γ as above such that the polyellipse domain $\mathcal{E}_{\gamma}$ is contained in the domain

of uniform ellipticity $U(a, \delta)$ from (2.12). One can further show that in this setting we have

$$\|\mathbf{c}_\nu\|_\nu \leq C_{\gamma, \delta} \gamma^{-\nu} \prod_{i=1}^d \sqrt{2\nu_i + 1} \quad (5.42)$$

for every $\nu \in \mathcal{F}$ with the same γ . Applying [113, Theorem 2] with the above bounds on the sequence $(\|\mathbf{c}_\nu\|_\nu)_{\nu \in \mathcal{F}}$ we obtain the result of [113, Proposition 5], that the quasi-optimal approximation $u_{\mathcal{J}_s}$ satisfies

$$\left\| u - \sum_{\nu \in \mathcal{J}_s} \mathbf{c}_\nu \Psi_\nu \right\|_{L^2_\varrho(\mathcal{U}; \mathcal{V})}^2 \leq C_{\gamma, \delta}^2 C_\varepsilon s \exp \left[-2 \left(\frac{d! s \prod_{i=1}^d \lambda_i}{(1 + \varepsilon)} \right)^{1/d} \right],$$

for constant $C_{\gamma, \delta} > 0$. Adapting the analysis of Theorem 5.13 to this setting, with $\mathcal{J}_s \subseteq \mathcal{J}$, under (5.39) the solution $u_{\mathcal{J}}^\# = \sum_{\nu \in \mathcal{J}} \mathbf{c}_\nu^\# \Psi_\nu$ obtained by solving (5.40) approximates the true solution u of the parameterized elliptic PDE (2.2) with error

$$\left\| u - \sum_{\nu \in \mathcal{J}} \mathbf{c}_\nu^\# \Psi_\nu \right\|_{L^2_\varrho(\mathcal{U}; \mathcal{V})} \leq \tilde{C}_\varepsilon \sqrt{s} \exp \left[- \left(\frac{d! s \prod_{i=1}^d \lambda_i}{(1 + \varepsilon)} \right)^{1/d} \right], \quad (5.43)$$

where $\tilde{C}_\varepsilon = \left(C_1 + 1 + \frac{C_2}{\sqrt{m}} \right) C_{\gamma, \delta} \sqrt{C_\varepsilon}$ and $C_1, C_2 > 0$ are the constants from Proposition 3.19.

Chapter 6

Numerical experiments

In this chapter, we present illustrative numerical experiments exploring the computational complexity and convergence properties of the methods presented in Chapter 3. Our results are organized as follows. Section 6.1 provides a detailed study of the computational complexity of stochastic Galerkin method for solving the parameterized elliptic PDE (2.2). Under various types of parametric dependence in the elliptic operator, including the affine, polynomial, and transcendental functions presented in Examples 2.1-2.3, we study the overall cost in terms of FLOPs of the SG approximations using the cost metrics from Section 5.1.1. Our results indicate that, while the SG method enjoys optimal complexity under affine parameterizations, that optimality is lost for higher-order parametric dependence. Moreover, our numerical results in Section 6.1.3 indicate that the SG system that results from transcendental coefficients in the elliptic operator, after applying the projection techniques described in Section 3.2.2 to sparsify the system, requires orders of magnitude more work to solve over other methods of approximation.

Section 6.2 presents experiments on smooth high-dimensional function approximation with the compressed sensing polynomial approximation strategy from Section 3.5. More specifically, we provide a thorough comparison of weighting strategies for ℓ_1 -minimization, demonstrating the superior error properties of our novel choice of weights enhancing recovery on the lower sets discussed in Section 3.5.2. Our results demonstrate that our choice of weights $\omega_\nu = \|\Psi_\nu\|_{L^\infty(\mathcal{U})}$ outperforms unweighted ℓ_1 -minimization and mitigates the spurious oscillations often present in high-order polynomial approximations, thereby dramatically reducing the number of samples m required to reach a given tolerance. Moreover, the weighting strategy is readily applicable to the setting of parameterized PDEs and sparse HV recovery through solving (3.69), after a modification of the $\ell_{\nu,1}$ -norm to its weighted counterpart, and we expect application to that setting to yield similar improvements. We leave a detailed study of the weighted approach for sparse HV recovery to a future work.

Finally, in Section 6.3 we present results of applying the sparse HV recovery techniques of Section 3.6 to parameterized elliptic PDE (2.2). In particular, in Section 6.3.1 we compare the point-wise functional recovery strategy from (3.61) to the HV recovery approach. Our results in Section 6.3.1 show that the sparse HV recovery technique yields a global approximation over D which has superior error properties over the approximation obtained by interpolating the sparse point-wise functionals. We conclude this chapter in Section 6.3.2 with a comparison of the sparse HV recovery technique to other popular methods

of approximation, including the stochastic Galerkin, stochastic collocation, and Monte Carlo methods. These results indicate that the minimal sample complexity requirements of compressive sensing-based methods of polynomial approximation carry over to the setting of parameterized PDEs, allowing highly accurate solutions to be constructed with dramatically fewer PDE solves than standard sampling-based approximations. Moreover, when compared to the stochastic Galerkin method, which equates degrees of freedom with the cardinality of the basis used in construction, our numerical experiments reveal that the HV recovery method yields comparable approximations with far fewer degrees of freedom.

6.1 Computational complexity of SG approximations

In this section, we provide illustrative numerical examples exploring the computational complexity of the SG method with FE physical discretization in the three cases of Examples 2.1-2.3. We then compare these results with SC method and the results of the theoretical complexity comparison of Section 5.2.2. In all examples considered, we solve the parameterized elliptic PDE (2.2), on the unit square $D = [0, 1]^2$. For a general coefficient $a(x, \mathbf{y})$ the exact solution to (2.2) is difficult to derive. Therefore, we check the convergence against a “highly enriched” approximation, which we consider close enough to the exact solution. To construct this “exact” solution $u_{\text{ex}}(x, \mathbf{y})$, we make use of the isotropic SC method based on Clenshaw-Curtis abscissas on a sufficiently large level L_{ex} . We approximate the computational error for the SG method with orders $p = 0, 1, 2, \dots, p_{\text{max}}$ and for the SC method with levels $L = 0, 1, 2, \dots, L_{\text{max}}$ as

$$\|\mathbb{E}[\varepsilon_{\text{SG}}]\|_{L^\infty(D)} \approx \|\mathbb{E}[\mathbf{u}_{\text{ex}} - \tilde{\mathbf{u}}_{h,p}]\|_{\ell^\infty} \quad \text{and} \quad \|\mathbb{E}[\varepsilon_{\text{SC}}]\|_{L^\infty(D)} \approx \|\mathbb{E}[\mathbf{u}_{\text{ex}} - \tilde{\mathbf{u}}_{h,L}]\|_{\ell^\infty}, \quad (6.1)$$

where $\tilde{\mathbf{u}}_{h,p}$ and $\tilde{\mathbf{u}}_{h,L}$ are the vectors of nodal values associated with the fully discrete SG and SC approximations $\tilde{u}_{h,p}$ and $\tilde{u}_{h,L}$ from (3.10) and (3.27), respectively, found by the FE method and PCG solver, as described in Sections 3.2 and 3.4. In Section 6.1.3, we measure $\|\mathbb{E}[\varepsilon_{\text{SG}}]\|_{L^\infty(D)} \approx \|\mathbb{E}[\mathbf{u}_{\text{ex}} - \tilde{\mathbf{u}}_{h,p}^r]\|_{\ell^\infty}$ where $\tilde{\mathbf{u}}_{h,p}^r$ is the vector of nodal values of the SG approximation with the FE method $\tilde{u}_{h,p}^r$, obtained by solving (3.21) with PCG after discretization by substituting the projected coefficient $a^r(x, \mathbf{y})$.

As stated in Sections 5.1.1 and 5.2.1, we use PCG with the mean-based preconditioner for the SG method and the level-zero preconditioner for the SC method. We believe this puts both methods at a similar starting point for comparison, if not providing a slight advantage for the SG method due to the fact that the level-zero preconditioner is only a constant approximation to $\mathbb{E}[\mathbf{B}(\mathbf{y})]$. With these choices, the complexity results are presented in terms of the work estimates (5.5) and (5.30), respectively. The amount of work to reach a given error in PCG is also dependent on the tolerance used by the solver. If the tolerance is too small, we may see that the PCG method “over-resolves” the solution. To ensure that we do not over-resolve either solution, we set the tolerance of the solvers to be $\|\mathbb{E}[\varepsilon_{\text{SG}}]\|_{L^\infty(D)}/10$ and $\|\mathbb{E}[\varepsilon_{\text{SC}}]\|_{L^\infty(D)}/10$, respectively, using the approximations (6.1). These quantities are first estimated for each order p and level L using a tolerance of $\varepsilon_{\text{PCG}} = 10^{-12}$. In practice, we find that this does not significantly affect the convergence results.

In all three examples, we use the SG approximation constructed in terms of the orthonormal Legendre polynomials $(\Psi_\nu)_{\nu \in \mathcal{J}_p}$ for given index sets $\mathcal{J}_p$. In the presentation of the results that follow, we use the following abbreviations. For the SG method, we use: “SG-TD” to denote the approximation in the total degree subspace $P_{\mathcal{J}_p^{\text{TD}}}(\mathcal{U})$ with $\mathcal{J}_p^{\text{TD}}$ given in (3.9), and “SG-SM” to denote the approximation in the sparse Smolyak subspace $P_{\mathcal{J}_p^{\text{SM}}}(\mathcal{U})$ with $\mathcal{J}_p^{\text{SM}}$ also given in (3.9). For the SC method, we use: “SC-GL” and “SC-LJ” to denote the Smolyak approximation constructed on Gauss-Legendre abscissas and the Leja approximation constructed on Clenshaw-Curtis abscissas, both defined in terms of g_{TD} and $\mathbf{m}$ given in (3.25) and (3.29), respectively, and “SC-CC” to denote the Smolyak approximation constructed on Clenshaw-Curtis abscissas with g_{SM} and $\mathbf{m}$ given in (3.25) and (3.28).

We follow the convention from [7, 54] in identifying the important *stochastic degrees of freedom* (SDOF) used in constructing the SG and SC approximations. In those works, the SDOF for the SG method is defined to be the cardinality N_p of the set $\mathcal{J}_p^{\text{TD}}$ used in constructing the fully discrete approximation $u_{h,p}$ from (3.10) by solving (3.14). The SDOF for the SC method is defined to be the number of points $M_L = \#(\mathcal{G}_L^{\mathbf{m},g})$ corresponding to the index set $\mathcal{J}_L^{\mathbf{m},g}$ used in constructing the fully discrete approximation $u_{h,L}$ from (3.27). Recalling the discussions of Section 5.1.1 and 5.2.1, and the approximations of W^{pSG} from (5.5) and W^{pSC} from (5.30), we note that this definition of SDOF may not be an accurate representation of the actual computational complexity of solving the SG systems for general parameterizations since $\mathcal{M}(p, r)$ is problem-dependent. Therefore in the convergence plots to follow, we include both error vs. SDOF and error vs. cost plots to highlight the difference in practice of solving these systems.

6.1.1 Piecewise affine coefficients

One common example in uncertainty quantification for engineering and the physical sciences is that of isotropic thermal diffusion problem with a stochastic conductivity coefficient. Consider a partitioning of $D = [0, 1]^2$ into eight circular inclusions of radius $r = 0.13$ arrayed about one square inclusion $F = [0.4, 0.6]^2$ as in Figure 6.1. We present the following example from [7], where the coefficient was given by

$$a(x, \mathbf{y}) = b_0(x) + \sum_{i=1}^8 y_i \chi_i(x), \quad (6.2)$$

where $b_0 = 1$, $y_i \sim U(-0.99, -0.2)$ (uniform on $(-0.99, -0.2)$) for $i = 1, \dots, 8$, and χ_i for $i = 1, \dots, 8$ are indicator functions on the circular inclusions. In this example, we also set the forcing term to be

$$f(x) = 100\chi_F(x). \quad (6.3)$$

Figure 6.1 shows the approximation of the expected value of the solution to this problem point-wise in D .

To solve (2.2) with the coefficient (6.2) and forcing function (6.3), we use a piecewise linear finite element basis over the physical space on a nonuniform mesh $\mathcal{T}_h$. More specifically, the

nodes of $\mathcal{T}_h$ are adapted to the geometry of our problem, i.e., we first select a large set of fixed nodes on the boundaries of the inclusions to guarantee that the generated triangulation approximates the true geometry of the problem. From this fixed boundary data, we then use the MATLAB program `distmesh` [97] to generate a non-degenerate triangulation that adequately resolves the details of our subdomain geometry. We further manually specify the subsets of the total set of nodes belonging to each geometric inclusion and its respective boundary, so that the interface conditions for the coefficient may be correctly applied. The final mesh, also plotted in Figure 6.1, consists of 10,604 elements, 5,377 total nodes, and 5,229 unknowns.

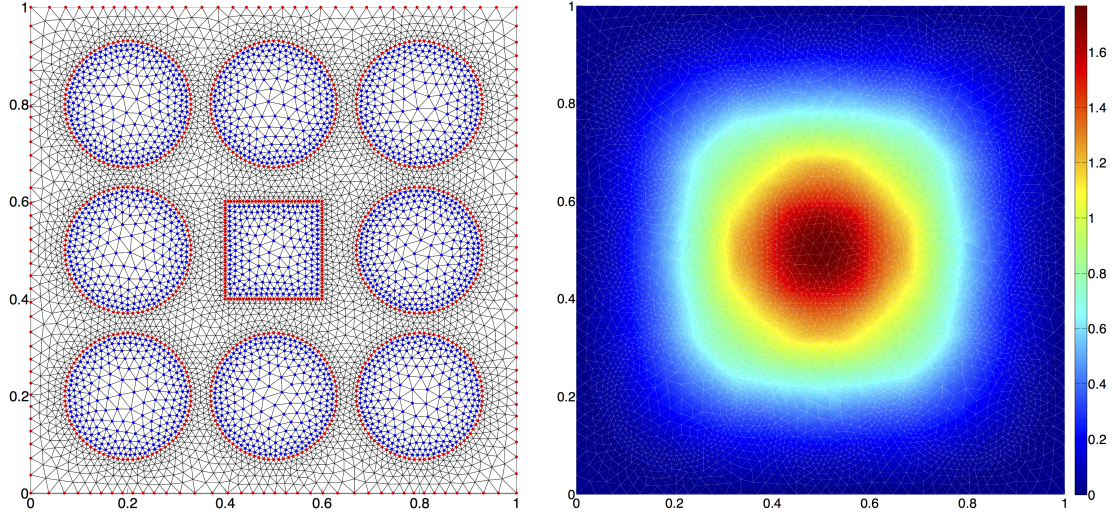


Figure 6.1: We plot (left) a triangulation of the domain D generated with `distmesh`, with circular and square inclusions. Red nodes highlight the boundary of an inclusion or the domain D , blue nodes highlight nodes on the interior of an inclusion. We also plot (right) the expected value of the solution of (2.2) with stochastic conductivity coefficient (6.2).

The coefficient (6.2) fits the description of Example 2.1, i.e., has affine dependence on the random variables $y_i \forall i$. Figure 6.2 displays the convergence of the stochastic Galerkin and collocation methods against the total number of SDOF. From the discussion of Section 5.2.2, we expect to see that the approximation obtained with the SG method requires fewer SDOF than required by the SC method to achieve the same error, and this is indeed the observed result. For example, both the SG-TD and SC-LJ approximations require the same number of SDOF, but the error of the SC-LJ approximation is much higher. This, of course, is a consequence of the estimate (5.31), where the errors of the SC approximations are bounded above by their respective Lebesgue constants $\mathbb{L}_{\mathcal{J}_L^{m,g}}$ times the best approximation error in the space $P_{\mathcal{J}_L^{m,g}}(\mathcal{U})$.

Figure 6.2 also displays the convergence of both methods in terms of error versus the total computational cost in FLOPs (normalized by an $\mathcal{O}(K_h)$ FE stiffness matrix-vector product) of solving the system with the work estimates of (5.5) and (5.30), respectively. Here, we compute the approximations to the error $\|\mathbb{E}[\varepsilon_{SG}]\|_{L^\infty(D)}$ and $\|\mathbb{E}[\varepsilon_{SC}]\|_{L^\infty(D)}$ as given in (6.1),

measuring the cost for the SG method as

$$\left(N_p + \sum_{\alpha \in \mathcal{J}_r} \text{nnz}(\mathbf{G}_\alpha) \right) * k_{\text{iter}}^{\text{pSG}} = (N_p + \mathcal{M}(p, r)) * k_{\text{iter}}^{\text{pSG}}. \quad (6.4)$$

Therefore, we explicitly count the number of FE matrix-vector products per iteration in (6.4), computing the number of nonzero elements in each $\mathbf{G}_\alpha$ contributing to the sum in the code for the SG method in each case of $\mathcal{J}_r = \mathcal{J}_r^{\text{TD}}, \mathcal{J}_r^{\text{SM}}$. We also count the quantity $2 * \sum_{i=1}^{M_L} N_{\text{iter}}^{p(i)}$ in the code for the SC method. Our analysis shows the work corresponding to the SG discretization for SG-TD is asymptotically bounded by $\mathcal{O}([\log(1/\varepsilon)]^d)$ while the analysis from [60] shows that the work corresponding to the SC discretization for SC-CC is asymptotically bounded by $\mathcal{O}([\log(1/\varepsilon)]^d [\log \log(1/\varepsilon)]^{d-1})$. These theoretical results on complexity match the numerical results of Figure 6.2, where it can be seen that for polynomial order $p \geq 2$, the SG-TD approximation yields the best results with the least computational cost for this problem. However, the savings in FLOPs for the SG method appears to be marginal, which can be accounted for by the presence of the extra term $\mathcal{O}([\log \log(1/\varepsilon)]^{d-1})$ in the asymptotic SC cost estimates.

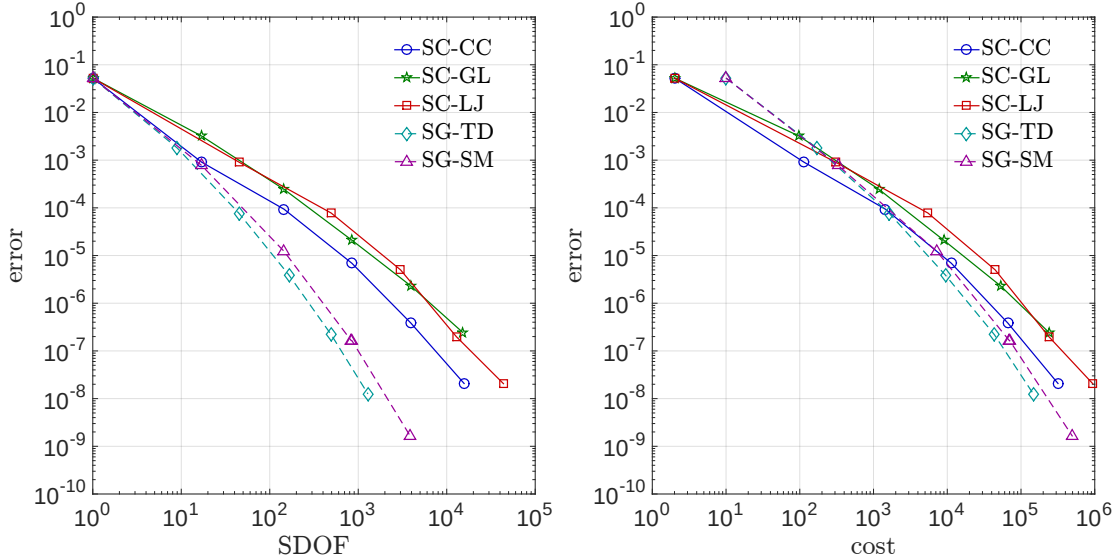


Figure 6.2: We plot **(left)** error versus SDOF and **(right)** error versus cost for solving problem (2.2) with coefficient (6.2) and forcing function (6.3), using the approximations to the error metrics $\|\mathbb{E}[\varepsilon_{\text{SG}}]\|_{L^\infty(D)}$ and $\|\mathbb{E}[\varepsilon_{\text{SC}}]\|_{L^\infty(D)}$ from (6.1) for error and the cost metrics (5.5) and (5.30) (normalized to an $\mathcal{O}(K_h)$ FE stiffness matrix-vector product).

6.1.2 Polynomial coefficients

The next example we present is that of a polynomial function of the random parameters $\mathbf{y}$, e.g., the coefficient from Example 2.2. We consider the function

$$a(x, \mathbf{y}) = 5 + \sum_{|\boldsymbol{\alpha}| \leq \bar{r}} e^{-1.5|\boldsymbol{\alpha}|} \varsigma_{\boldsymbol{\alpha}}(x) \mathbf{y}^{\boldsymbol{\alpha}}, \quad (6.5)$$

$$\varsigma_{\boldsymbol{\alpha}}(x) = \begin{cases} \sin(|\boldsymbol{\alpha}|\pi x_1) \cos(|\boldsymbol{\alpha}|\pi x_2) & \text{if } |\boldsymbol{\alpha}| \text{ is even,} \\ \cos(|\boldsymbol{\alpha}|\pi x_1) \sin(|\boldsymbol{\alpha}|\pi x_2) & \text{if } |\boldsymbol{\alpha}| \text{ is odd,} \end{cases}$$

with $y_i \sim U(-1, 1)$ (uniform on $(-1, 1)$) for all $i = 1, \dots, d$ and forcing term $f(x) = 1 \forall x \in \bar{D}$. For the results that follow we fix $d = 4$ and study the convergence of the SG and SC methods in the cases $\bar{r} = 1, 3, 7$ in (6.5). As in Section 6.1.1, we set the finite element space for the spatial discretization to be the span of piecewise linear polynomials, but here we use a uniform triangulation of D with 4,934 elements and 2,340 spatial unknowns.

Figure 6.3 displays the convergence of the SG and SC methods in terms of error versus SDOF and error versus total computational cost of solving the system with the work estimates of (5.5) and (5.30). There we can see, as we increase the order $\bar{r}$ in (6.5), the work for the SG method to reach a given tolerance dramatically increases by a large multiplicative constant, dependent on $\bar{r}$, while the error versus SDOF increases much more slowly. Our analysis from Section 5.1.2 shows that the work of matrix-vector multiplications with $\mathbf{K}$ when problem (2.2) is discretized with coefficient (6.5) can be bounded as

$$\mathcal{M}(p, \bar{r}) \leq 2 \sum_{j=0}^{\bar{r}} \min \left\{ 2^j, \binom{d + \lceil j/2 \rceil}{d} \right\} \binom{d-1+j}{d-1} \binom{d+p-\lceil j/2 \rceil}{d}, \quad (6.6)$$

for fixed $\bar{r}$, see the green line in Figure 5.2 for an illustration of this fact as p increases and Remark 5.4 for discussion. Table 6.1 shows the growth of the above bound with respect to d and $\bar{r}$ for fixed $p = 5$. The dramatic increase in work to reach a given tolerance corresponds to the decreasing sparsity of the matrix $\mathbf{K}$ from (3.14). As a result, we see that for $\bar{r} = 1$, the SGFEM outperforms the other methods for $p \geq 4$. However, for $\bar{r} = 3, 7$, the extra work of the matrix-vector multiplications of the coupled SG system dominates the overall convergence. We also observe that the convergence rate of the SG approximation with respect to FLOPs does not change in these cases, thus verifying our conclusions of Section 5.2.2 on the asymptotic complexity of the SG method in the case of fixed polynomial order dependence on the parameters.

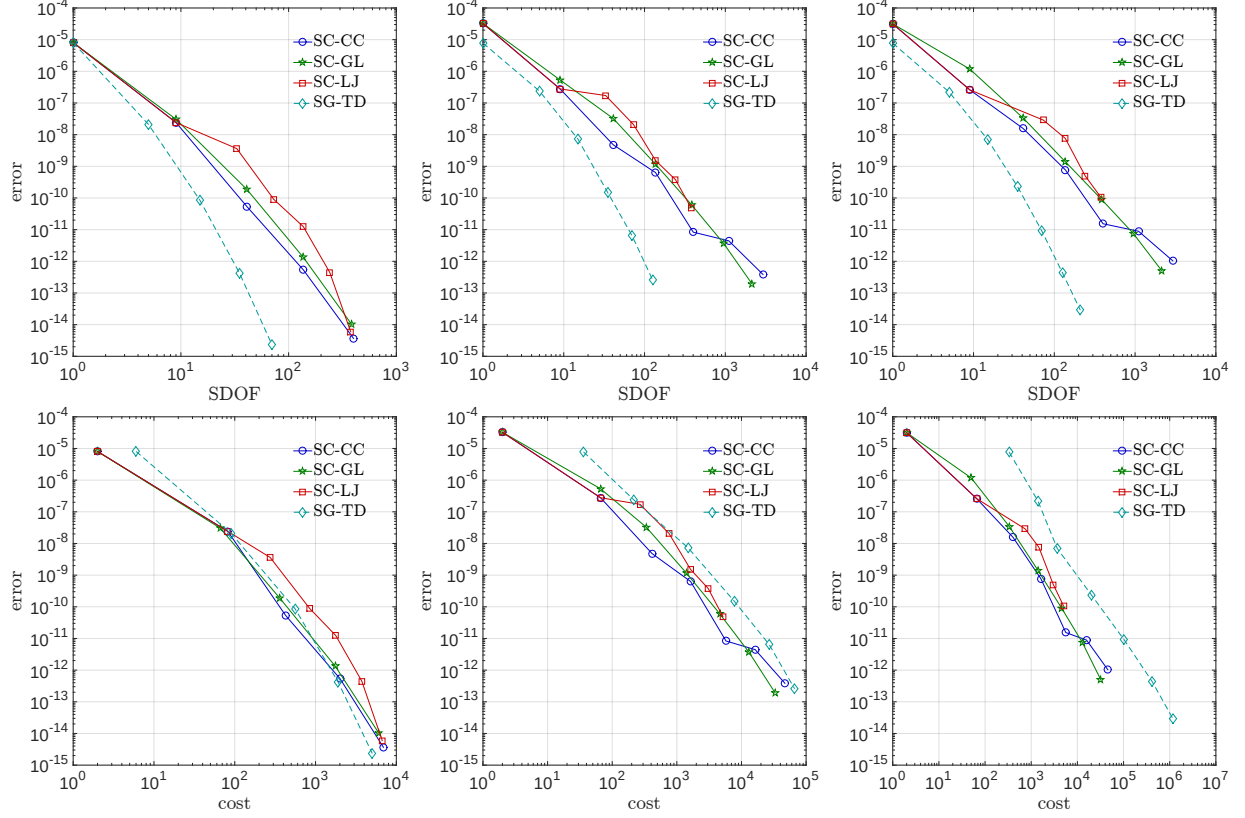


Figure 6.3: We plot (top) error vs SDOF and (bottom) error versus cost for solving problem (2.2) with coefficient (6.5) with $\bar{r} = 1$ (left), $\bar{r} = 3$ (middle), and $\bar{r} = 7$ (right), with forcing $f(x) = 1 \forall x \in \bar{D}$, using the approximations to the error metrics $\|\mathbb{E}[\varepsilon_{\text{SG}}]\|_{L^\infty(D)}$ and $\|\mathbb{E}[\varepsilon_{\text{SC}}]\|_{L^\infty(D)}$ from (6.1) for error and the cost metrics (5.5) and (5.30) (normalized to an $\mathcal{O}(K_h)$ FE stiffness matrix-vector product).

Table 6.1: Growth of the bound (6.6) on the constant $\mathcal{M}(p, \bar{r})$, with respect to dimension d and $\bar{r}$ when $p = 5$ is fixed.

$\bar{r}$	$d = 4$	$d = 8$	$d = 16$
1	1,372	18,414	350,778
2	6,972	160,974	5,622,138
3	18,172	477,774	18,273,402
4	54,922	2,220,174	138,460,410
5	108,682	4,501,134	290,275,578
6	196,882	14,385,294	1,352,981,754
7	280,882	22,292,622	2,095,189,242
8	396,382	51,945,102	6,362,882,298
9	451,822	63,659,662	7,701,766,394
10	523,894	103,489,166	14,396,186,874

6.1.3 Transcendental coefficients

The next example we present is that of a random coefficient defined in terms of the truncated Karhunen-Loève expansion of the function $\log(a(x, \mathbf{y}) - a_{\min})$, for $a_{\min} > 0$. This example, commonly found in groundwater flow problems, is often also presented in the context of enforcing the positivity of $a(x, \mathbf{y})$ as required by Assumption (A1). For such models, the permeability can exhibit large variance within each layer of sediment, and as a result is best represented on a logarithmic scale. Moreover, this problem gives a concrete example of a diffusion coefficient having transcendental dependence on the random parameters as in Example 2.3.

We recall the following problem from [91, 92] of solving (2.2) with deterministic forcing $f(x_1, x_2) = 2 \cos(x_1) \sin(x_2)$ and coefficient having one-dimensional (layered) spatial dependence given by

$$\log(a(x, \mathbf{y}) - 0.5) = 1 + y_1 \left(\frac{\sqrt{\pi}\Lambda}{2} \right)^{1/2} + \sum_{i=2}^d \zeta_i \vartheta_i(x) y_i, \quad (6.7)$$

$$\zeta_i := (\sqrt{\pi}\Lambda)^{1/2} \exp \left(-\frac{\left(\lfloor \frac{i}{2} \rfloor \pi \Lambda\right)^2}{8} \right), \text{ for } i > 1, \quad (6.8)$$

$$\vartheta_i(x) := \begin{cases} \sin \left(\lfloor \frac{i}{2} \rfloor \pi x_1 / L_p \right), & i \text{ even}, \\ \cos \left(\lfloor \frac{i}{2} \rfloor \pi x_1 / L_p \right), & i \text{ odd}. \end{cases} \quad (6.9)$$

Here, $y_i \sim U(-\sqrt{3}, \sqrt{3})$ (uniform on $(-\sqrt{3}, \sqrt{3})$) $\forall i = 1, \dots, d$. For $x_1 \in [0, b]$, let L_c be a desired physical correlation length for the random field $a(x, \mathbf{y})$, chosen so that the random variables $a(x_1, \mathbf{y})$ and $a(x'_1, \mathbf{y})$ become essentially uncorrelated for $|x_1 - x'_1| \gg L_c$. Also, let $L_p = \max\{b, 2L_c\}$ and $\Lambda = L_c/L_p$. Expression (6.7) represents a possible truncation of a

one-dimensional random field with stationary covariance,

$$\text{cov}[\log(a - 0.5)](x_1, x_2) = \exp\left(\frac{-(x_1 - x_2)^2}{L_c^2}\right),$$

and is often referred to as the log-transformed Karhunen-Loève expansion. Direct integration with the coefficient $a(x, \mathbf{y})$ from (6.7) yields a fully-block dense linear system $\mathbf{K}$ from (3.14) that is computationally infeasible to solve, see, e.g., [87, 117]. The purpose of this example is to highlight difficulties of obtaining the fully discrete approximation with the SG and FE methods in the case of transcendental parametric dependence.

As in the previous example in Section 6.1.2, we set the finite element space for the spatial discretization to be the span of piecewise linear polynomials and use a uniform triangulation of D with 4,934 elements and 2,340 spatial unknowns. For the results that follow, we fix the truncation length $d = 9$ and correlation length $L_c = 1/64$ in (6.7). To maintain sparsity of the SG system, we use the strategy of projecting the coefficient $a(x, \mathbf{y})$ from (6.7) onto the space $P_{\mathcal{J}_r}(\mathcal{U})$, as in (3.16), where $\mathcal{J}_r = \mathcal{J}_r^{\text{TD}}$ for the SG-TD approximation, and $\mathcal{J}_r = \mathcal{J}_r^{\text{SM}}$ for the SG-SM approximation, obtaining the matrix $\mathbf{K}_r$ from (3.21). We then increase r while p is fixed until the error in the solution stagnates, in practice finding that, for this problem, $r = p$ is sufficient to guarantee the error of the projection does not exceed that of the solution, while maintaining sparsity of the linear system.

Figure 6.4 compares the error versus SDOF. There we see that for order $p \geq 3$, the SG-TD approximation provides the best approximation with respect to SDOF. As discussed in Section 5.2.2, this is to be expected since the computational complexity of solving the coupled and decoupled systems is not taken into account by the SDOF cost metric. Figure 6.4 also displays the convergence in error versus the total computational cost of solving the system with the work estimates of (5.5) and (5.30). Here however, the results show that the SG method requires orders of magnitude more work to obtain the same error as the SC method. We also observe the change in rate discussed in Section 5.2.2 in this case, as the work required to solve (2.2) with the coefficient $a(x, \mathbf{y})$ from (6.7) now depends on the order r of the projection used in the approximation $a^r(x, \mathbf{y})$ of $a(x, \mathbf{y})$.

Recalling the discussion of Remark 5.5 relevant to the TD-SG approximation, this can be seen as a consequence of the fact that when $r = p$, the cost of solving (3.21) with the PCG method is of the order $\mathcal{O}(K_h N_{[p/2]}^3 k_{\text{iter}}^{\text{SG}})$, growing much more rapidly than the cost in the affine and polynomial coefficient cases, e.g., Examples 2.1-2.2, as we increase the order p . Table 6.2 shows the amount of work required to achieve an error on the order of 10^{-k} for some values of k in terms of the total number of matrix-vector products required by both the SC-CC and SG-TD approximations. The huge discrepancy between the computational cost of the SG and SC methods in these results can be seen as a consequence of the fact that the scaling of the exponential dependence on d in (5.25) changes when the coefficient requires a further projection, degrading as the order of the projection increases towards $2p$.

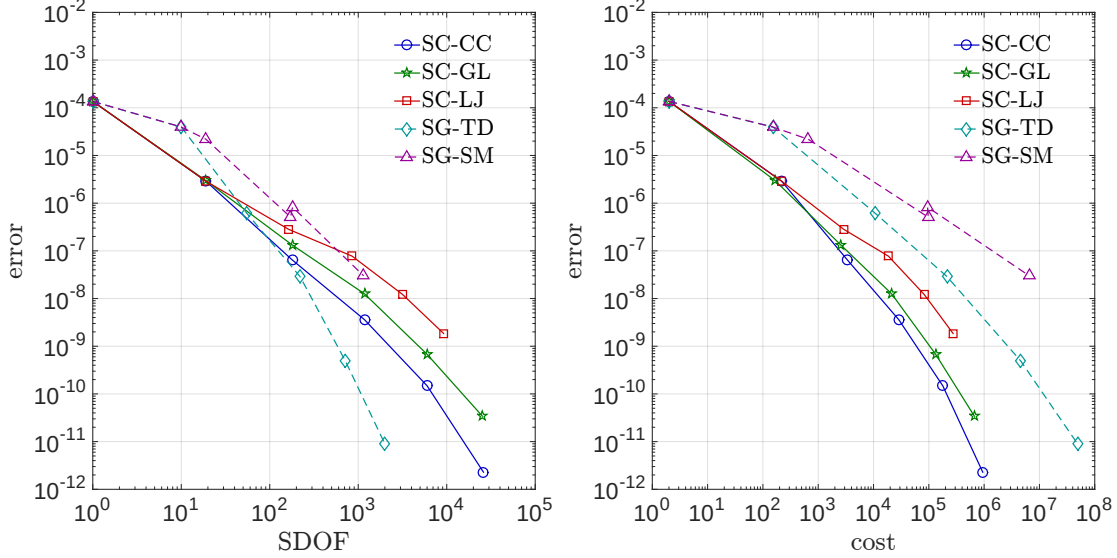


Figure 6.4: We plot (left) error versus SDOF and (right) error versus cost in solving problem (2.2) with coefficient (6.7) and forcing deterministic $f(x_1, x_2) = 2 \cos(x_1) \sin(x_2)$, using the approximations to the error metrics $\|\mathbb{E}[\varepsilon_{SG}]\|_{L^\infty(D)}$ and $\|\mathbb{E}[\varepsilon_{SC}]\|_{L^\infty(D)}$ from (6.1) for error and the cost metrics (5.5) and (5.30) (normalized to an $\mathcal{O}(K_h)$ FE stiffness matrix-vector product).

6.2 Approximating high-dimensional smooth functions with lower sparse recovery

In this section, we provide several numerical examples to demonstrate the efficiency of our weighted ℓ_1 minimization with $\omega_\nu = \|\Psi_\nu\|_{L^\infty(\mathcal{U})}$ for smooth multivariate function recovery. We focus here on the approximation of a function $g : \mathcal{U} \rightarrow \mathbb{R}$ for $\mathcal{U} = [-1, 1]^d$ in terms of orthonormal Legendre expansions by solving

$$\min_{\mathbf{z} \in \mathbb{C}^N} \|\mathbf{z}\|_{\omega,1} \quad \text{subject to} \quad \|\tilde{\mathbf{g}} - \mathbf{A}\mathbf{z}\|_2 \leq \eta/\sqrt{m}, \quad (6.10)$$

i.e., the weighted version of problem (3.60) for various choices of weights, where $\eta = \sqrt{m}\|g_{\mathcal{H}_s^c}\|_2$. This tail bound complies with [29, Theorem 4.5], based on an estimate of $\|\mathbf{e}\|_2$ (or $\|\tilde{\mathbf{g}} - \mathbf{A}\mathbf{c}\|_2$) via $\|g_{\mathcal{H}_s^c}\|_2$, and is less conservative than those analytically studied in [29, Theorem 4.8]. As we test with simple functions, the expansion of g can be computed numerically with the use of a quadrature approximation yielding an estimate of the tail η a priori. The MATLAB software code SPGL1 [118, 119] is employed to solve (6.10).

Table 6.2: Comparison of cost in matrix-vector products for solving problem (2.2) with coefficient (6.7) and forcing $f(x_1, x_2) = 2 \cos(x_1) \sin(x_2)$ using the SC-CC and SG-TD approximations, with the strategy of picking the CG tolerance to be $\|\mathbb{E}[\varepsilon_{\text{SG}}]\|_{L^\infty(D)}/10$ for the SG method and $\|\mathbb{E}[\varepsilon_{\text{SC}}]\|_{L^\infty(D)}/10$ for the SC method using the approximations from (6.1) and the cost metrics (5.5) and (5.30) (normalized to an $\mathcal{O}(K_h)$ FE stiffness matrix-vector product).

SC-CC Level	SC-CC Error	Mat-vec cost of SC-CC	SG-TD Order	SG-TD Error	Mat-vec cost of SG-TD
0	1.3626×10^{-4}	2	0	1.3626×10^{-4}	4
1	2.8884×10^{-6}	218	1	3.9444×10^{-5}	152
2	6.3652×10^{-8}	3,398	2	6.1427×10^{-7}	10,710
3	3.6021×10^{-9}	28,638	3	2.8851×10^{-8}	213,010
4	1.4794×10^{-10}	178,894	4	4.9210×10^{-10}	4,579,575
5	2.2869×10^{-12}	944,220	5	8.9123×10^{-12}	49,089,051

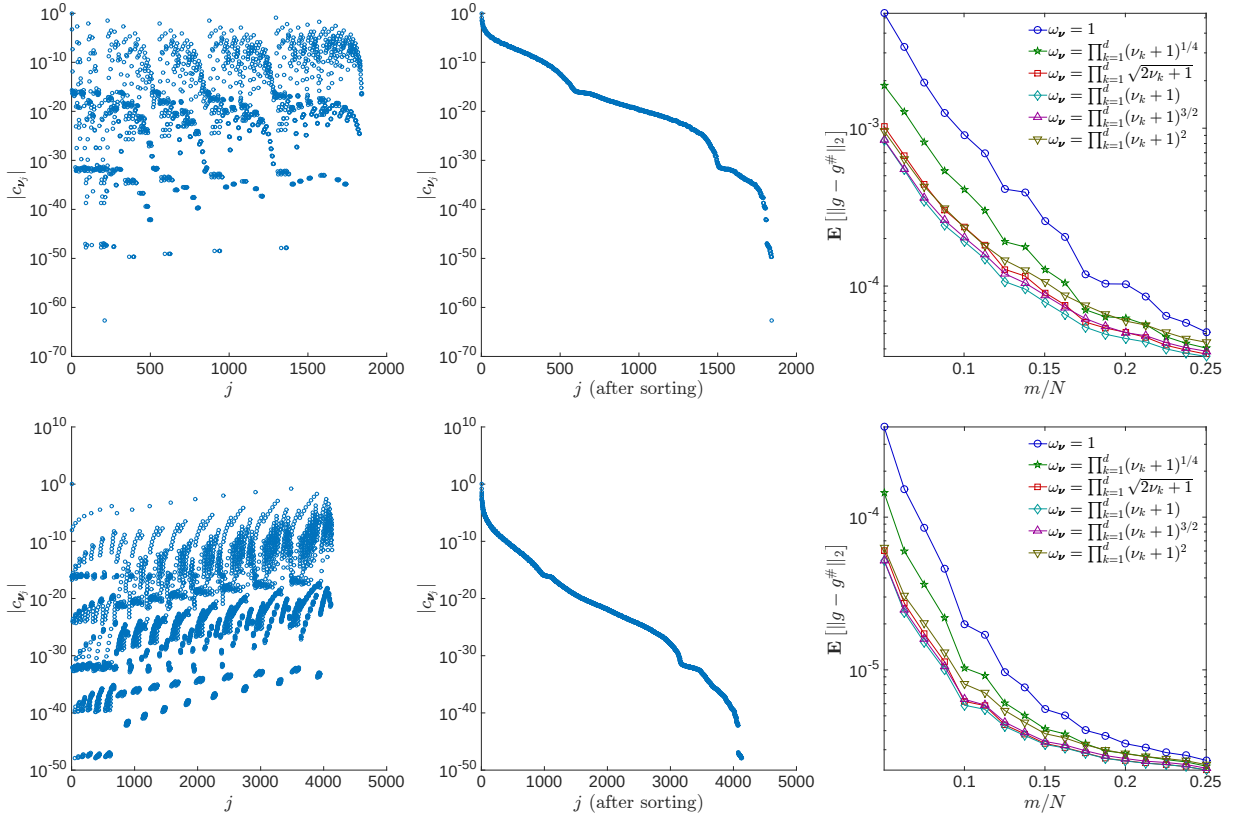


Figure 6.5: Comparison of the averaged $L^2(\mathcal{U})$ error in approximating g from (6.11) using weighted ℓ^1 -minimization with various choices of weights in **(top)** $d = 8$, $N = 1843$, and $\|g_{\mathcal{H}^s}\|_2 = 1.5922e - 05$, and **(bottom)** $d = 16$, $N = 4129$, and $\|g_{\mathcal{H}^s}\|_2 = 1.3388e - 06$.

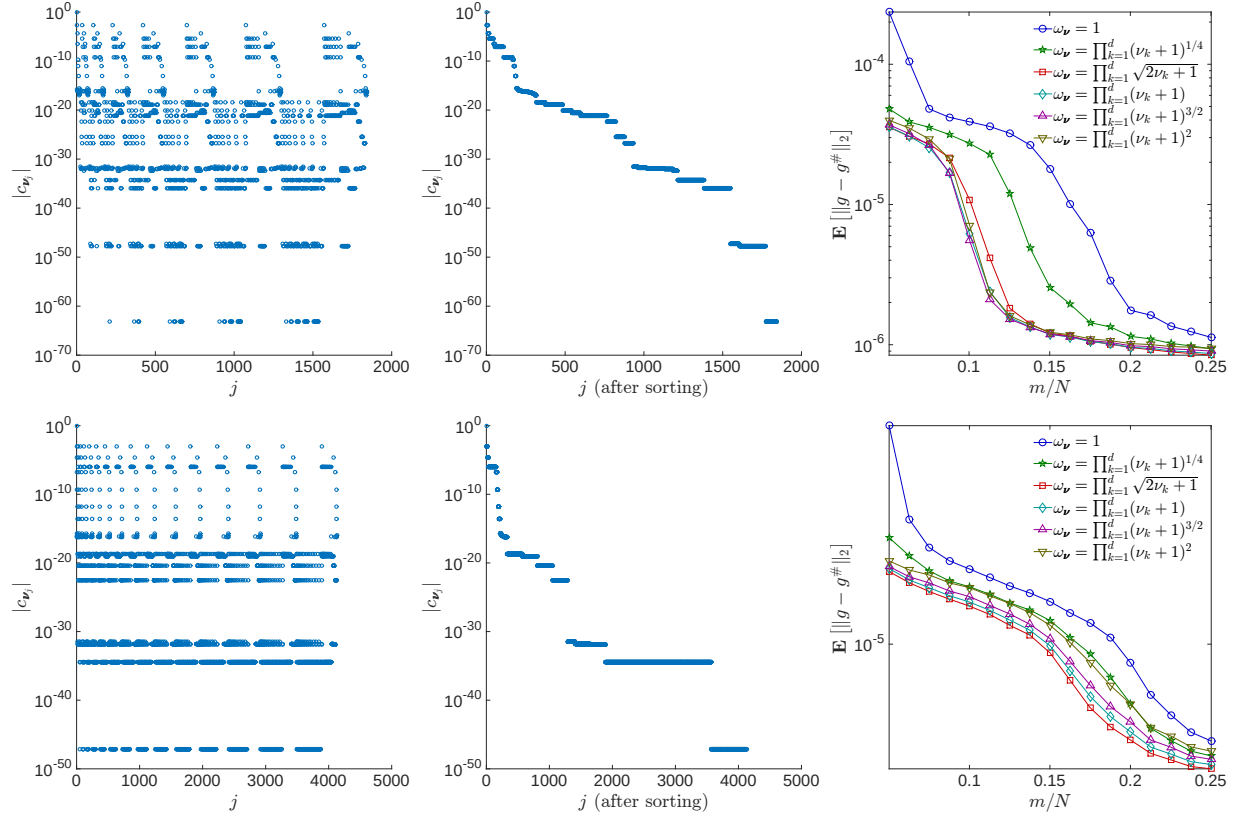


Figure 6.6: Comparison of the averaged $L^2(\mathcal{U})$ error in approximating g from (6.12) using weighted ℓ^1 minimization with various choices of weights in **(top)** $d = 8$, $N = 1843$, and $\|g_{\mathcal{H}_s^c}\|_2 = 4.0232e - 07$, and **(bottom)** $d = 16$, $N = 4129$, and $\|g_{\mathcal{H}_s^c}\|_2 = 2.0155e - 06$.

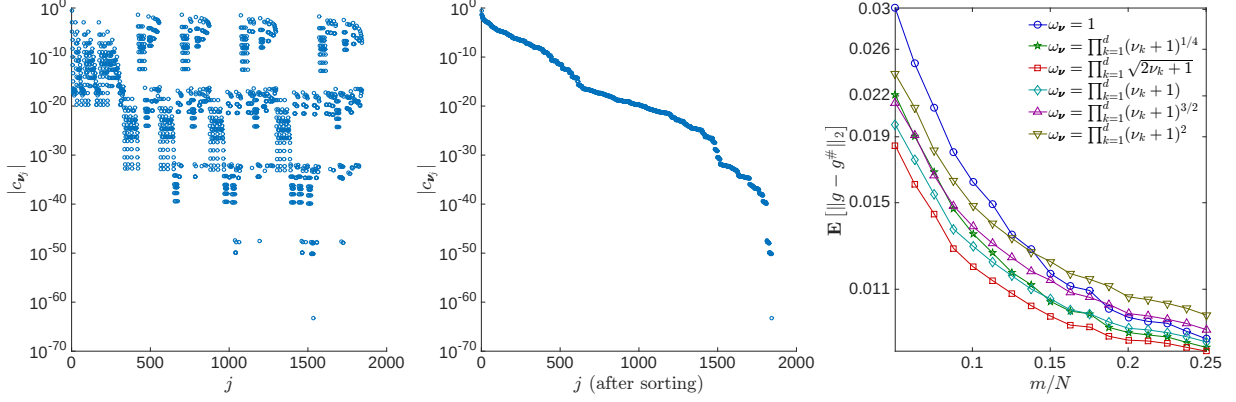


Figure 6.7: Comparison of the averaged $L^2(\mathcal{U})$ error in approximating g from (6.13) using weighted ℓ^1 minimization with various choices of weights with $d = 8$, $N = 1843$ and $\|g_{\mathcal{H}_s^c}\|_2 = 6.1018e - 3$.

In each example, we choose the polynomial subspace $P_{\mathcal{J}}(\mathcal{U}) \equiv P_{\mathcal{H}_s}(\mathcal{U})$, and increase the number of samples m up to some $m_{\max} < N = \#(\mathcal{H}_s)$. For each ratio m/N , the set of random samples is fixed over various choices of weights for a performance comparison. We then run 50 trials for the averaged $L^2_{\mathcal{Q}}$ error, setting the maximum number of iterations in SPGL1 per trial to 10,000. Our results are shown in Figures 6.5-6.8 for several multivariate functions. The left panels represent the magnitudes of polynomial coefficients (computed with MATLAB and C++ the sparse-grid package TASMANNIAN [107, 108]) indexed in $\mathcal{H}_s$ and sorted lexicographically. The center panels depict the decay of the coefficients once sorted by magnitude. The right panels show the corresponding convergence results. In Figure 6.9, we include some results of the approximation errors in the $L^\infty(\mathcal{U})$ norm, showing little change in the performance of the investigated weights compared to the $L^2_{\mathcal{Q}}(\mathcal{U})$ case.

Our experiments indicate that for functions with very fast decaying polynomial expansions, $\omega_{\nu} = \|\Psi_{\nu}\|_{L^\infty(\mathcal{U})}$ is virtually the optimal weight. Indeed, for the function

$$g(\mathbf{y}) = \frac{\prod_{k=\lceil d/2 \rceil+1}^d \cos(16y_k/2^k)}{\prod_{k=1}^{\lceil d/2 \rceil} (1 - y_k/4^k)} \quad (6.11)$$

concerned in Figure 6.5, we see in both $d = 8$ and $d = 16$ that the weight $\omega_{\nu} = \|\Psi_{\nu}\|_{L^\infty(\mathcal{U})}$ significantly outperforms the unweighted ℓ_1 approach. We note that in $d = 16$, higher weights begin to have decreasing benefit, performing worse than our proposed weight.

The results in Figure 6.6 are related to the function

$$g(\mathbf{y}) = \exp\left(\frac{-\sum_{k=1}^d \cos(y_k)}{8d}\right). \quad (6.12)$$

We note that only a small fraction of coefficients exceed 10^{-16} in both $d = 8$ and $d = 16$. In this case, we see that the weighted ℓ^1 with the weight $\omega_{\nu} = \|\Psi_{\nu}\|_{L^\infty(\mathcal{U})}$ performs the best out of all of the weights supplied. We also observe that increasing the weights beyond $\|\Psi_{\nu}\|_{L^\infty(\mathcal{U})}$ leads to a corresponding increase in the averaged $L^2_{\mathcal{Q}}(\mathcal{U})$ error.

In Figure 6.7, we test the function

$$g(\mathbf{y}) = \left[\frac{\prod_{k=1}^{\lceil d/2 \rceil} (1 + 4^k y_k^2)}{\prod_{k=\lceil d/2 \rceil+1}^d (100 + 5y_k)} \right]^{1/d}. \quad (6.13)$$

Here again, the weight $\omega_{\nu} = \|\Psi_{\nu}\|_{L^{\infty}(\mathcal{U})}$ performs the best, and the approximation errors grow as the weights increase beyond this weight. The two highest weights even perform worse than unweighted ℓ^1 -minimization for higher values of m/N . Comparing Figures 6.5 and 6.7, the similar center panels suggest similar decay rates for both functions inside $\mathcal{H}_s$. The performance of the weights between the two examples however drastically differs, possibly due to the different support of the large coefficients. For this function, we were unable to test with $d = 16$ and $N = 4129$ due to the high expansion tail.

On the other hand, if the polynomial expansion is less sparse, or has largest coefficients outside a lower set, our weight may be not optimal. Figure 6.8 shows the results for the function

$$g(\mathbf{y}) = \exp \left(-\frac{\sum_{k=1}^d y_k}{2d} \right). \quad (6.14)$$

For this function, over half of the coefficients in $\mathcal{H}_s$ exceed 10^{-8} , and $\omega_{\nu} = \|\Psi_{\nu}\|_{L^{\infty}(\mathcal{U})}$ performs worse than the larger weights. Still, we observe here, as in all tests for smooth functions in higher dimensions, i.e., $d = 8$ and $d = 16$, that our weight consistently provides improved accuracy compared with unweighted ℓ_1 , thus confirming the theory presented throughout.

All of the results have been presented in terms of the $L^2_{\varrho}(\mathcal{U})$ error thus far. To investigate how the approximations behave pointwise, we also measure all the errors in the $L^{\infty}(\mathcal{U})$ norm. The achieved results show little or no change in the relative performance with respect to different choice of weights, compared to that of the $L^2_{\varrho}(\mathcal{U})$ case, i.e., our weight is always more accurate than unweighted approach and in cases of Figure 6.5-6.7 is the optimal weight. For brevity, we only show here the $L^{\infty}(\mathcal{U})$ error of the approximations considered in Figure 6.6. Figure 6.9 highlights a key advantage of our approach that focusing on best lower s -term recovery mitigates spurious oscillations that occur when using high-degree polynomial approximations.

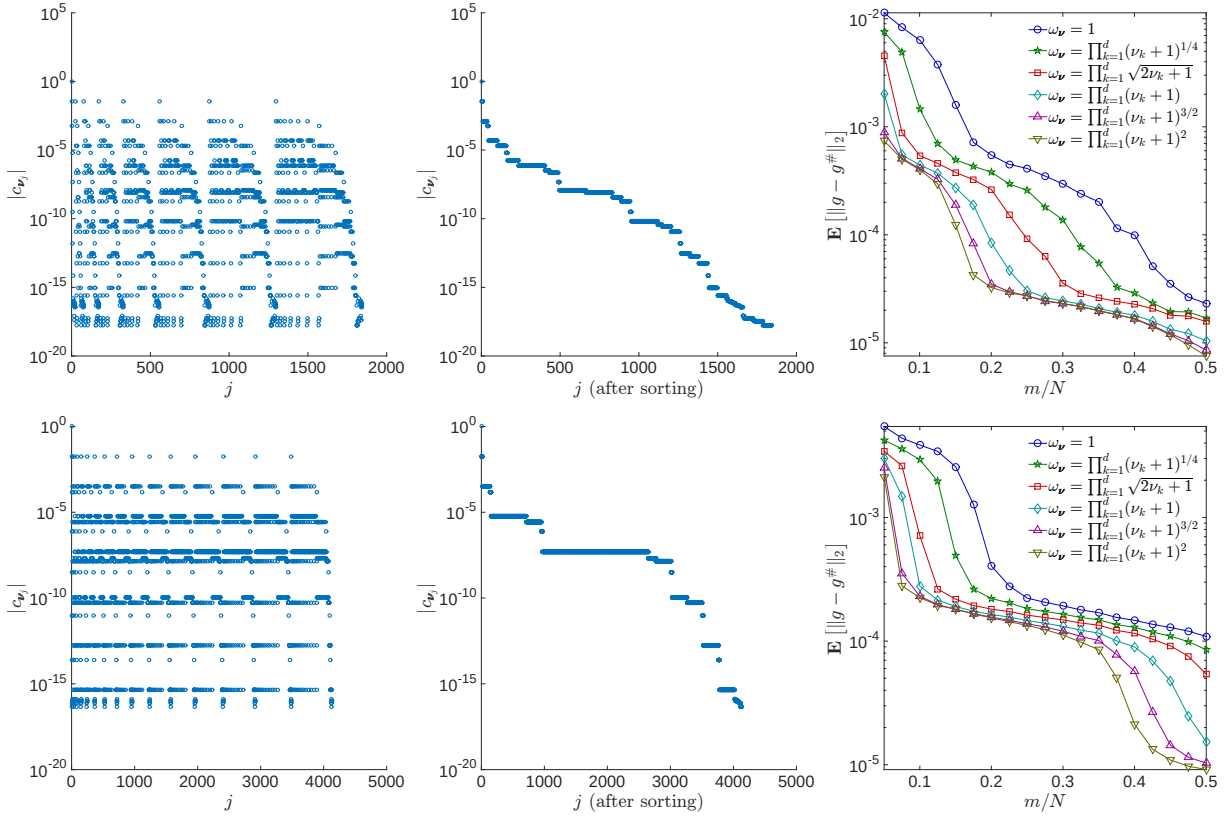


Figure 6.8: Comparison of the averaged L^2_ρ error in approximating g from (6.14) using weighted ℓ^1 minimization with various choices of weights in **(top)** $d = 8$, $N = 1843$, and $\|g_{\mathcal{H}_s^c}\|_2 = 7.2714e - 07$, and **(bottom)** $d = 16$, $N = 4129$, and $\|g_{\mathcal{H}_s^c}\|_2 = 3.7412e - 07$.

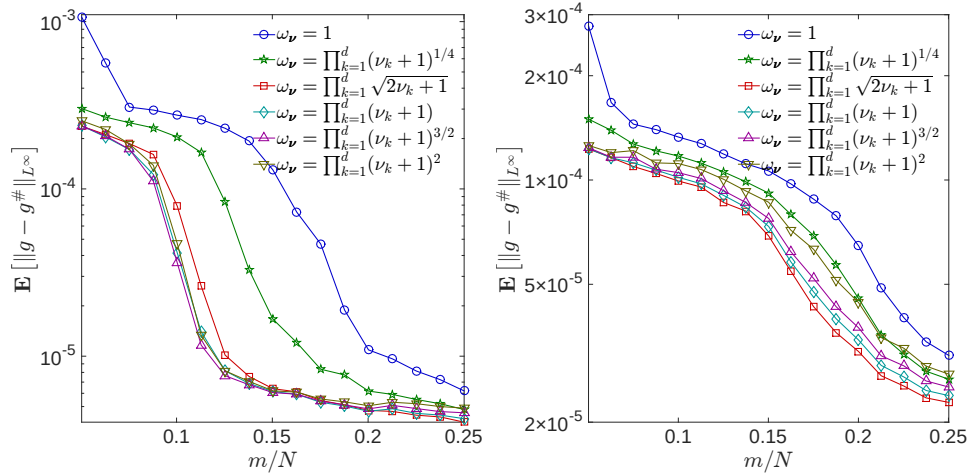


Figure 6.9: Comparison of the averaged $L^\infty(\mathcal{U})$ error of approximations studied in Figure 6.6 in **(left)** $d = 8$ and **(right)** $d = 16$.

6.3 Numerical results for the approximation of solutions to the parameterized elliptic PDE with sparse Hilbert-valued recovery

In this section we provide several numerical experiments solving the parameterized problem (2.2) on the unit square $D = [0, 1]^2 \subset \mathbb{R}^2$, demonstrating the efficiency of sparse HV recovery relative to other approaches. As in Section 6.1, for a general coefficient $a(x, \mathbf{y})$, we do not know the exact solution u to (2.2). Therefore, we check convergence of parametric discretizations against “highly-enriched” reference sparse grid SC approximations, as in Section 6.1, which we denote here $u_{h,\text{ex}}(x, \mathbf{y})$. More specifically, all approximations (including the enriched reference solution) are computed on fixed finite element meshes $\mathcal{T}_h$ (to be described), and our enriched SC approximation is computed using Clenshaw-Curtis abscissas with level L_{ex} , larger than that of the other SC approximations. We then approximate the relative errors of the expectation and standard deviation in the $H_0^1(D)$ -norm for each of the approximations $u_{h,\text{approx}}$ as

$$\varepsilon_{h,\text{approx}}^{\text{rel-}\mathbb{E}} \approx \frac{\|\mathbb{E}_{\mathbf{y}}[u_{h,\text{ex}} - u_{h,\text{approx}}]\|_{H_0^1(D)}}{\|\mathbb{E}_{\mathbf{y}}[u_{h,\text{ex}}]\|_{H_0^1(D)}} \quad \varepsilon_{h,\text{approx}}^{\text{rel-}\sigma} \approx \frac{\|\text{Var}_{\mathbf{y}}[u_{h,\text{ex}} - u_{h,\text{approx}}]^2\|_{H_0^1(D)}}{\|\text{Var}_{\mathbf{y}}[u_{h,\text{ex}}]^2\|_{H_0^1(D)}} \quad (6.15)$$

where $\mathbb{E}_{\mathbf{y}}[\cdot]$ and $\text{Var}_{\mathbf{y}}[\cdot]$ denote expectation and variance with respect to $\mathbf{y} \in \mathcal{U}$ from (2.3) & (2.4), respectively.

As in Section 6.1, we follow the convention from [7, 54] in identifying the important *stochastic degrees of freedom* (SDOF) as the number of random sample points m for the CS, MC, and sparse HV recovery approaches and number of structured sparse grid points M_L for the SC method. For the SG method, SDOF remains N , the cardinality of the particular index set $\mathcal{J}$ used in construction. While this metric is not an accurate representation of the computational complexity of obtaining the approximation, it provides a useful benchmark in comparing the sample complexity of the sampling-based methods. We include the SG method only to compare the $L_q^2(\mathcal{U})$ -optimal (w.r.t. SDOF) error of the Galerkin projection against the error of the sampling-based approximations.

In Section 6.3.1 we give a comparison between our sparse HV recovery approach and global recovery through compressed sensing-based polynomial approximation of the pointwise functionals (3.61). Section 6.3.2 compares the sparse HV recovery approach to several popular alternatives for high-dimensional approximation of solutions to parameterized PDEs, including the Monte Carlo, stochastic Galerkin, and stochastic collocation methods. In our plots and discussion in these sections, we use the following abbreviations. For the CS and sparse HV recovery methods, we use: “CS-TD” and “HV-TD” to denote the approximations obtained using the CS and sparse HV recovery methods in the total degree subspace $P_{\mathcal{J}_p^{\text{TD}}}(\mathcal{U})$ with $\mathcal{J}_p^{\text{TD}}$ given in (3.9) for a fixed p to be provided. Similarly, for the SG method, we use: “SG-TD” to denote the SG approximation obtained in the subspace $\mathcal{J}_p^{\text{TD}}$ for increasing p . For the SC method, we use: “SC-CC” to denote the sparse grid Smolyak approximation constructed on Clenshaw-Curtis abscissas of level L increasing, with g_{SM} and $\mathbf{m}$ given in (3.25) and (3.28). For the Monte Carlo approximation, we use “MC” often also including the convergence rate $\mathcal{O}(m^{-1/2})$ of the MC method in plots for comparison.

In all examples excluding MC and SC, we use the orthonormal Legendre polynomials $(\Psi_\nu)_{\nu \in \mathcal{J}}$ in parametric discretization. For the MC, CS-TD, and HV-TD methods, we use the same set of random sample points $(\mathbf{y}_i)_{i=1}^m$ in obtaining a given approximation. However, since each method relies on random sampling, in order to better understand the average performance of the algorithms, we use the strategy of running all three methods on 24 trials, fixing the initial seed q for the pseudorandom number generator per trial, and then solving each trial's problem with the same set $(\mathbf{y}_i^{(q)})_{i=1}^m$. We then average the results when plotting convergence over the trials.

6.3.1 Comparison of sparse HV recovery and CS point-wise functional recovery for global approximation of solutions to parameterized PDEs

In this section we give a comparison of the sparse HV recovery and CS point-wise functional recovery techniques of Sections 3.6.1 & 3.5.3, respectively. For both approaches we construct a fully discrete approximation of the solution u to the parameterized elliptic PDE (2.2). In this comparison, both methods use the same set of random points $(\mathbf{y}_i)_{i=1}^m$ drawn from the measure ϱ and corresponding normalized samples $(u_h(x, \mathbf{y}_i)/\sqrt{m})_{i=1}^m \in \mathcal{V}_h^m$, with $u_h(x, \mathbf{y})$ from (3.2), obtained with the finite element method from Section 3.1. We describe details of obtaining the approximations as follows.

For the sparse HV recovery method, we solve the unconstrained optimization problem (3.70), with measurement matrix $\mathbf{A}$ given by (3.68) and data $\mathbf{u}$ given by $\mathbf{u}^h = (u_h(x, \mathbf{y}_i)/\sqrt{m})_{i=1}^m \in \mathcal{V}_h^m$. To yield solutions to the HV-BPDN problem (3.69) through problem (3.70), we apply Bregman-FPC iterations of Algorithm 2 with the values from Table 4.1 where $\mathbf{c}^* = \mathbf{c}^{h,*} \in \mathcal{V}_h^N$, used in parameterizing b_{tol} to satisfy the residual constraint of (3.69), is given by

$$\mathbf{c}_\nu^{h,*}(x) = \sum_{k=1}^{K_h} [\mathbf{u}_{h,\nu}]_k \varphi_k(x) \in \mathcal{V}_h(D) \quad \forall \nu \in \mathcal{J}.$$

Here $(\mathbf{u}_{h,\nu})_{\nu \in \mathcal{J}}$ from (3.12) are the coefficients of the solution obtained with the SG and FE methods (sufficiently resolved by the PCG solver), see Sections 3.2 & 6.1.

Letting $\mathbf{c}^{h,\#} \in \mathcal{V}_h^N$ denote the HV approximation obtained by the combined Bregman-FPC solver, our fully discrete sparse HV recovery approximation to u is then given by

$$u_{h,\mathcal{J}}^{\text{HV},\#}(x, \mathbf{y}) = \sum_{\nu \in \mathcal{J}} \mathbf{c}_\nu^{h,\#}(x) \Psi_\nu(\mathbf{y}) = \sum_{\nu \in \mathcal{J}} \sum_{k=1}^{K_h} \hat{\mathbf{c}}_{\nu,k}^{h,\#} \varphi_k(x) \Psi_\nu(\mathbf{y}). \quad (6.16)$$

When $\mathcal{J}$ is the total degree set from (3.9), we refer to $u_{h,\mathcal{J}}^{\text{HV},\#}$ as the HV-TD approximation to u . We can also view our algorithm as simultaneously solving for the matrix of coefficients $\hat{\mathbf{c}}^{h,\#} \in \mathbb{C}^{N \times K_h}$ given by $[\hat{\mathbf{c}}^{h,\#}]_{\nu,k} = \hat{\mathbf{c}}_{\nu,k}^{h,\#}$ for each $k = 1, \dots, K_h$ and $\nu \in \mathcal{J}$, i.e., approximating the matrix $\hat{\mathbf{c}}^{h,*} \in \mathbb{C}^{N \times K_h}$ associated with $\mathbf{c}^{h,*} \in \mathcal{V}_h^N$ given by $[\hat{\mathbf{c}}^{h,*}]_{\nu,k} = [\mathbf{u}_{h,\nu}]_k$ for each $k = 1, \dots, K_h$ and $\nu \in \mathcal{J}$. Therefore $\hat{\mathbf{c}}^{h,\#}$ directly approximates the coefficients of the SG approximation in the finite element and orthonormal bases $(\varphi_k)_{k=1}^{K_h}$ and $(\Psi_\nu)_{\nu \in \mathcal{J}}$,

respectively. Due to the sparse regularization, enforced by the $\ell_{\nu,1}$ -norm, we expect the matrix $\hat{\mathbf{c}}^{h,\#}$ to be row-sparse, or approximately row-sparse.

Recalling the modifications of the forward-backward and Bregman-FPC iterations from Remark 4.3 and Sections 4.2.1 & 4.3, at each iteration involving J_τ from (4.10) we must compute $\|\cdot\|_\nu$ -norms of the components of the Hilbert-valued vectors $\mathbf{c}^h$ in regularization to solve (3.70). With the FE expansions (3.79) we can easily compute the energy norms required in regularization as $\|\mathbf{c}_\nu^h\|_\nu = \|\hat{\mathbf{c}}_\nu^h\|_{\mathbf{B}} = \langle \hat{\mathbf{c}}_\nu^h, \mathbf{B}(\hat{\mathbf{c}}_\nu^h)^\top \rangle_2$ where $\hat{\mathbf{c}}_\nu^h$ denotes the ν -th row of $\hat{\mathbf{c}}^h$ and $\mathbf{B}$ is the FE stiffness matrix resulting from substituting $a \equiv 1$ in $\mathcal{B}[u, v](\mathbf{y})$ from (2.10), i.e., given by $[\mathbf{B}]_{l,k} = \int_D \nabla \phi_k \cdot \nabla \phi_l \, dx$ for $l, k = 1, \dots, K_h$.

For the CS point-wise functional recovery approach, we solve the decoupled problems (3.60) with the same measurement matrix $\mathbf{A}$ from (3.68) and data $\mathbf{g}^{(k)} := (u_h(x_k, \mathbf{y}_i))_{i=1}^m \subset \mathbb{C}^m$ at all points x_k on the finite element mesh $\mathcal{T}_h$. We apply the standard Bregman-FPC algorithm from [129], using the same parameters from Table 4.1 with $\mathbf{c}^\star$ as above. From this procedure, we obtain a set of sparse vectors $(\tilde{\mathbf{c}}^{(k),\#})_{k=1}^{K_h} \subset \mathbb{C}^m$ where the vector $\tilde{\mathbf{c}}^{(k),\#}$ defined by $[\tilde{\mathbf{c}}^{(k),\#}]_\nu = \tilde{c}_{\nu,k}^{h,\#}$ for each $\nu \in \mathcal{J}$ from (3.62) approximates $\mathbf{c}^{(k)} \in \mathbb{C}^N$ given by $[\mathbf{c}^{(k)}]_\nu = c_\nu(x_k)$ for each $\nu \in \mathcal{J}$ with $c_\nu(x_k)$ from the expansion (3.61). As in the case of sparse HV recovery, we can define the matrix $\tilde{\mathbf{c}}^{h,\#} \in \mathbb{C}^{N \times K_h}$ by $\tilde{\mathbf{c}}^{h,\#} = [\mathbf{c}^{(1),\#}, \mathbf{c}^{(2),\#}, \dots, \mathbf{c}^{(K_h),\#}]$. This matrix will also approximate the matrix $\hat{\mathbf{c}}^{h,\star}$ of coefficients of the SG and FE approximations. However, recalling the discussion of Remark 3.14, this matrix $\tilde{\mathbf{c}}^{h,\#}$ may not be (approximately) row-sparse due to the fact that the set of s largest terms of the point-wise GPC expansions may have disjoint supports.

To construct the fully discrete approximation to u from the point-wise approximations, we interpolate the GPC expansions $g_{k,\mathcal{J}}^\#(\mathbf{y}) \approx g_k(\mathbf{y})$ from (3.62) associated with the vectors $(\tilde{\mathbf{c}}^{(k),\#})_{k=1}^{K_h}$ in the basis $(\varphi_k)_{k=1}^{K_h}$ to obtain

$$u_{h,\mathcal{J}}^{\text{pwCS},\#}(x, \mathbf{y}) = \sum_{\nu \in \mathcal{J}} \sum_{k=1}^{K_h} \tilde{c}_{\nu,k}^{h,\#} \varphi_k(x) \Psi_\nu(\mathbf{y}). \quad (6.17)$$

When $\mathcal{J}$ is the total degree set from (3.9), we refer to $u_{h,\mathcal{J}}^{\text{pwCS},\#}$ as the CS-TD approximation to u , which can be used to compute any desired quantity of interest, such as the error metrics of (6.15) in comparing to the HV-TD approximation.

Recalling the example from Section 6.1.3, we study problem (2.2) when parameterized with a deterministic load $f \equiv 1$ and diffusion coefficient $a(x, \mathbf{y})$ given by a modification of the right hand side of (6.7), i.e., the affine function

$$a(x, \mathbf{y}) = 10 + y_1 \left(\frac{\sqrt{\pi}\Lambda}{2} \right)^{1/2} + \sum_{i=2}^d \zeta_i \vartheta_i(x) y_i. \quad (6.18)$$

Here ζ_i and ϑ_i are given in (6.8) and (6.9), respectively, $y_i \sim U(-\sqrt{3}, \sqrt{3})$ (uniform on $(-\sqrt{3}, \sqrt{3})$) $\forall i = 1, \dots, d$, and the constant 10 is chosen to guarantee $a(x, \mathbf{y})$ satisfies Assumption (A1). For $x_1 \in [0, b]$, let L_c be a desired physical correlation length for the random field $a(x, \mathbf{y})$, chosen so that the random variables $a(x_1, \mathbf{y})$ and $a(x'_1, \mathbf{y})$ become essentially uncorrelated for $|x_1 - x'_1| \gg L_c$. Also, let $L_p = \max\{b, 2L_c\}$ and $\Lambda = L_c/L_p$.

In the following example, with $L_c = 1/4$ and $d = 100$ in (6.18), we construct the approximations HV-TD and CS-TD from (6.16) & (6.17), respectively, in a total degree basis of order $p = 2$ having cardinality $N = \binom{d+p}{p} = 5151$. We compute the random samples of our FE approximations $(u_h(x, \mathbf{y}_i)/\sqrt{m})_{i=1}^m \in \mathcal{V}_h^m$ using a quasi-uniform mesh $\mathcal{T}_h$ having 206 degrees of freedom, corresponding to a maximum mesh size of $h \approx 1/16$.

The left panel of Figure 6.10 shows a comparison of the relative error statistic $\varepsilon_{h,\text{approx}}^{\text{rel}-\mathbb{E}}$ from (6.15) for the HV-TD and CS-TD approximations. For each data point, we increase the number of random samples m following the rule $m_k = \lceil kN/8 \rceil$ for $k = 1, 2, \dots, 7$, keeping $N = 5151$ fixed. The top middle and top right panels display the decay of the SG coefficients $\mathbf{c}_{\boldsymbol{\nu}}^{h,*}$ in $\|\cdot\|_{\mathcal{V}}$ -norm, before and after sorting by magnitude, and the bottom middle and bottom right panels show the decay of $|\mathbf{c}_{\boldsymbol{\nu}}^{h,*}(x_i)|$ at a selection of points x_i on the mesh for $\mathcal{T}_h$, before and after sorting by magnitude. We note that the choice of parameterization implies u is highly compressible, in the sense that the coefficients $\mathbf{c}_{\boldsymbol{\nu}} \in \mathcal{V}$ from (3.5) satisfy $\|\mathbf{c}_{\boldsymbol{\nu}}\|_{\mathcal{V}} \rightarrow 0$ rapidly as $|\boldsymbol{\nu}| \rightarrow \infty$, making it ideal for approximation with compressed sensing-based polynomial strategies. This can be seen as a consequence of the choice of correlation length L_c , resulting in the coefficient $a(\mathbf{y})$ and hence $u(\mathbf{y})$ having highly anisotropic dependence on the variables $\mathbf{y} \in \mathcal{U}$, with decreasing dependence on y_i as $i \rightarrow \infty$.

While the difference in errors between the HV-TD and CS-TD approximations is somewhat dramatic, it should be noted that the parameterization for the solvers use the tolerance $b_{\text{tol}} = 1.2 \cdot \|\mathbf{A}\mathbf{c}^{h,*} - \mathbf{u}\|_{\mathcal{V},2}$, a global error statistic. Since the CS-TD approximations involve a sequence of point-wise solves, this value of b_{tol} may lead to poor point-wise approximations as it does not account for the error of truncation at a particular point x_k in the mesh. In practice however, such point-wise truncation estimates are often unavailable for practical parameterized PDE problems. On the other hand, *a priori* estimates in global energy norms have been shown for a wide variety of parameterized PDEs under reasonable assumptions on the input data, see Section 5.3.

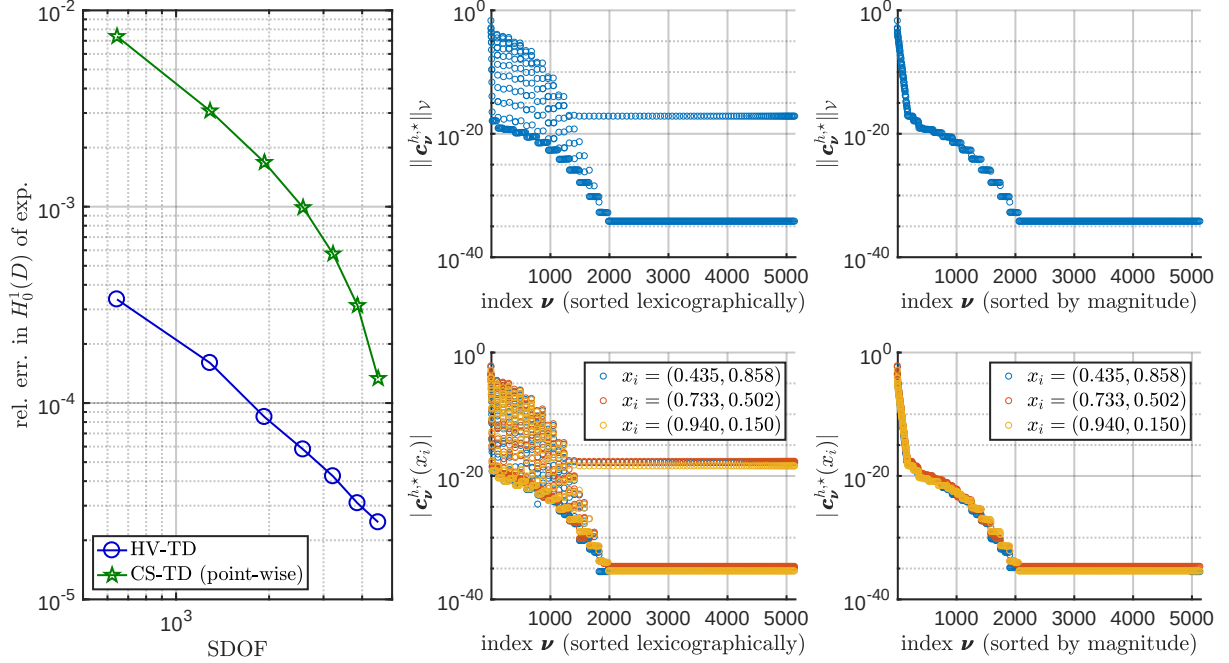


Figure 6.10: (left) Comparison of relative error statistics $\varepsilon_{h,\text{approx}}^{\text{rel}-\mathbb{E}}$ from (6.15) for the sparse HV recovery method (HV-TD) and compressed sensing point-wise functional recovery (CS-TD) methods, both computed with a total degree basis of order $p = 2$ in $d = 100$ dimensions having cardinality $N = 5151$. (center) Magnitudes of the coefficients in (top) energy norm and (bottom) pointwise at three physical locations x_i sorted lexicographically, (right) same as center after sorting by largest in magnitude. Here $\mathbf{c}^{h,*}$ is obtained with the stochastic Galerkin method in the same total degree polynomial subspace.

6.3.2 Comparison of sparse HV recovery with SC, SG, and MC methods for parameterized PDEs

In this section, we give a comparison of the approximation errors that can be obtained solving the stochastic elliptic PDE (2.2) with the sparse HV recovery method and the SG, MC, and SC methods from Sections 3.2, 3.3, and 3.4, respectively. For all of the methods considered, we use a fixed triangulation $\mathcal{T}_h$ of D . For the MC and HV-TD methods, we parameterize both solvers with the same set of points $(\mathbf{y}_i)_{i=1}^m$ drawn from the measure ϱ . We focus on two types of parameterizations, namely the affine coefficient from (6.18) over a range of values of dimension d and correlation length L_c , and the transcendental coefficient from (6.7). Each example highlights an important aspect of the convergence of the sparse HV recovery method.

The first example we study is that of the affine coefficient (6.18) with fixed $L_c = 1/4$ and increasing d . As in the previous section, the physical FE discretization for this problem uses a fixed quasi-uniform mesh $\mathcal{T}_h$ of $D = [0, 1]^2$ having $K_h = 206$ degrees of freedom, i.e., corresponding to a maximum mesh size of $h \approx 1/16$ in $n = 2$ physical dimensions. Each row of Figure 6.11 plots the relative error metrics $\varepsilon_{h,\text{approx}}^{\text{rel-}\mathbb{E}}$ on the left and $\varepsilon_{h,\text{approx}}^{\text{rel-}\sigma}$ on the right from (6.15), while keeping the order p of the total degree polynomial subspace used in computing the HV-TD approximation fixed. Each row increases d , from 20 on the top row to 60 on the middle row, and 100 on the bottom row. We note that $\varepsilon_{h,\text{approx}}^{\text{rel-}\sigma}$ uses all of the coefficients $\mathbf{c}_\nu^{h,\#}$ $\nu \in \mathcal{J}$ from (6.16), and is therefore a much better representation of the error of the HV-TD method than $\varepsilon_{h,\text{approx}}^{\text{rel-}\mathbb{E}}$, which only uses $\mathbf{c}_0^{h,\#}$, i.e., the HV-TD approximation of the first stochastic mode of (3.5), due to the orthonormality of $(\Psi_\nu)_{\nu \in \mathcal{F}}$.

As mentioned in Section 6.3.1, this choice of correlation length L_c results in highly anisotropic dependence of the solution u on the parameter vector $\mathbf{y}$, with most of the important terms in (6.18) corresponding to low index i of y_i . Moreover, the solution u is highly compressible, with the coefficients $\mathbf{c}_\nu$ satisfying $\|\mathbf{c}_\nu\|_\nu \rightarrow 0$ rapidly as $|\nu| \rightarrow \infty$. Hence, as d increases, the relative sparsity s/N decreases, since $N = \#(\mathcal{J})$ depends exponentially on d for total degree $\mathcal{J}$, while the best s terms of (3.5) scale approximately linearly with d under the coefficient (6.18). As a result, in higher dimensions the HV-TD approximation outperforms the SC and MC methods dramatically. This can be seen as a consequence of the fact that the SC-CC method is using an isotropic growth rule so that M_L is growing exponentially in d . Under an anisotropic growth rule, adapted to the coefficient decay, better performance of the SC-CC approximation would be observed. However, such anisotropic rules often require detailed knowledge of the parametric dependence.

Figure 6.12 displays the effect of doubling the correlation length L_c in the coefficient $a(x, \mathbf{y})$ from (6.7) from $1/4$ to $1/2$ in $d = 100$ dimensions, resulting in an expansion of the solution with even fewer large terms $\|\mathbf{c}_\nu\|_\nu$. For the relative standard deviation error statistic $\varepsilon_{h,\text{approx}}^{\text{rel-}\sigma}$, we begin to see the HV-TD method approaches the error of the final point of the SG-TD method, both of which are constructed using the same basis $(\Psi_\nu)_{\nu \in \mathcal{J}}$ of a total degree polynomial space of order $p = 2$. As $m_k = \lceil kN/8 \rceil$, $k = 1, 2, \dots, 7$, approaches $N = \#(\mathcal{J})$ (the cardinality of the set used in both constructions), the error of the HV-TD approximation begins to stagnate, suggesting that the Bregman-FPC algorithm has converged to the tolerance b_{tol} . Since $b_{\text{tol}} = 1.2 \cdot \|\mathbf{A}\mathbf{c}^{h,*} - \mathbf{u}\|_{\nu,2}$, with $\mathbf{c}^{h,*}$ the SG solution

of total order $p = 2$, this suggests the relative errors of both approximations are of the same order.

Both of Figures 6.11 & 6.12 highlight a key advantage of the sparse HV recovery approach that is common to all CS-based methods of approximation. Without *a priori* knowledge of the coefficient decay, or use of anisotropic weighting in the set $\mathcal{J}$, the HV-TD approximation is able to naturally detect the underlying anisotropy of the problem in refinement as the number of random samples m increases. Moreover, our results suggest that simply choosing $\mathcal{J}$ large enough to surely cover the best s terms in the expansion (3.5) is sufficient to yield highly accurate approximations. These approximations would only be further enhanced if *a priori* knowledge of the coefficient decay is given and anisotropic weights are incorporated into the set $\mathcal{J}$. Indeed, we expect the lower set-based ℓ_1 -weighting strategies for smooth function approximation in Section 6.2 to also yield improved accuracy. We leave a detailed study of such weighting techniques to a future work.

Finally, Figure 6.13 shows the result of choosing $a(x, \mathbf{y})$ as the log transformed Karhunen-Loève expansion from problem (6.7) with correlation length $L_c = 1/8$ in $d = 17$ dimensions. As observed in Section 6.1.3, such problems with transcendental parametric dependence in the elliptic operator are inherently more difficult to solve than those with affine or polynomial dependence. Here all approximations are computed on a fixed finite element mesh with $K_h = 713$ degrees of freedom, i.e., corresponding to a maximum mesh size of $h \approx 1/32$.

For this problem, in $d = 17$ dimensions, the stochastic Galerkin system becomes prohibitively difficult to discretize and solve, even when applying the sparse projection technique from Section 3.2.2. As a result, we only compare the Monte Carlo, joint-sparse, and stochastic collocation methods, using the SC error as an approximation to $\|\mathbf{A}\mathbf{c}^{h,*} - \mathbf{u}\|_{V,2}$ in parameterizing b_{tol} . Here we only plot the HV-TD method for $m_k = \lceil kN/8 \rceil$ for $k = 1, \dots, 4$. In both the $\varepsilon_{h,\text{approx}}^{\text{rel-}\sigma}$ and $\varepsilon_{h,\text{approx}}^{\text{rel-}\mathbb{E}}$ error metrics, the HV-TD method outperforms the MC and SC-CC approximations with respect to sample complexity, even in the case of slower anisotropic decay associated with $L_c = 1/8$. Furthermore, the rate of convergence of the HV-TD approximation is $\mathcal{O}(m^{-3/2})$, the fastest out of all methods included, further bolstering our claim of the ability of our approach to accurately detect and refine the most important terms of the expansion (3.5).

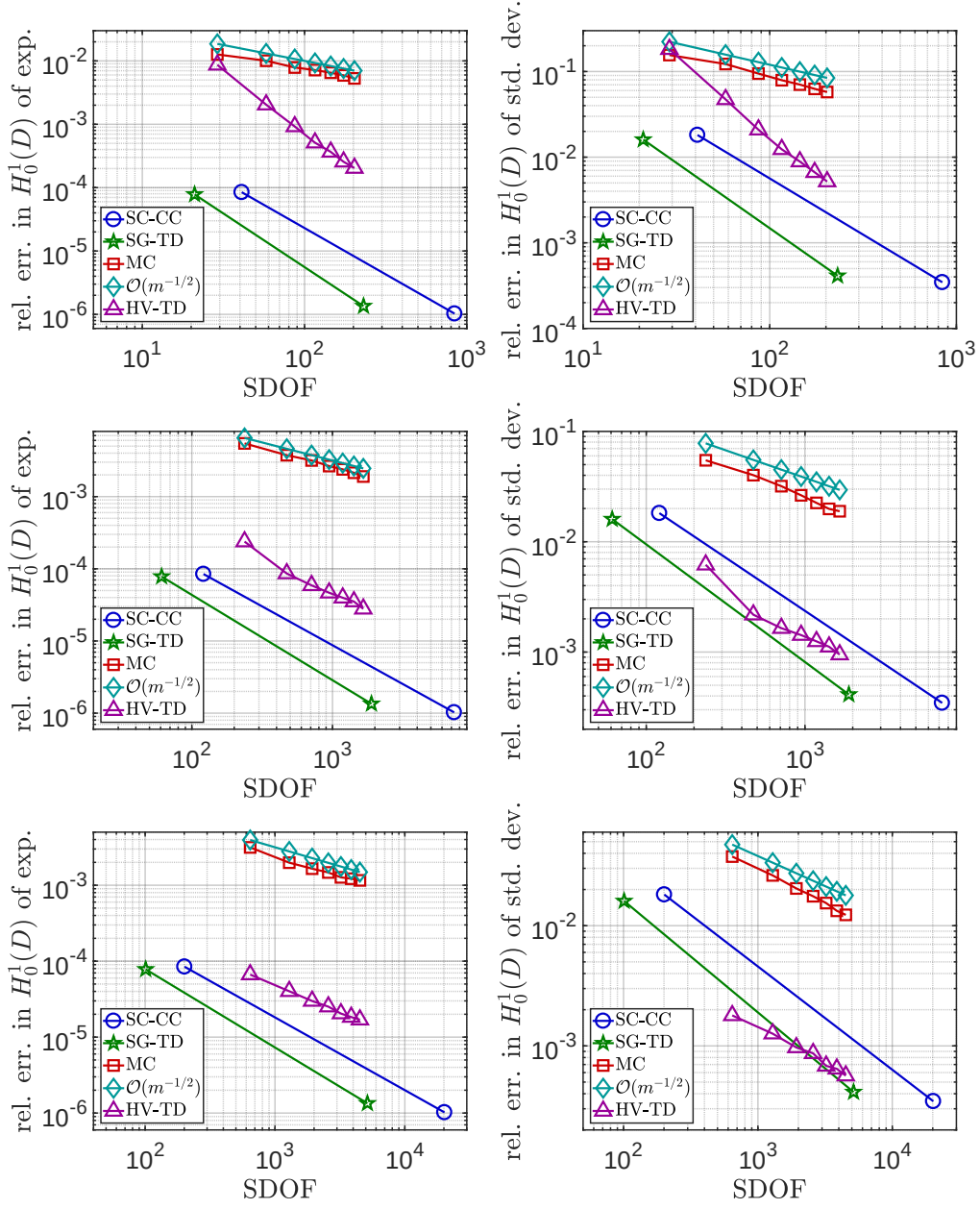


Figure 6.11: Comparison of relative approximation errors $\varepsilon_{h,\text{approx}}^{\text{rel-E}}$ (left) and $\varepsilon_{h,\text{approx}}^{\text{rel-}\sigma}$ (right) from (6.15) for the $L = 2, 3$ stochastic collocation (SC-CC), $p = 1, 2$ stochastic Galerkin (SG-TD), Monte Carlo (MC), and total degree order $p = 2$ sparse HV recovery (HV-TD) methods for solving (2.2) with coefficient (6.18) and correlation length $L_c = 1/4$. (top) $d = 20$, $N = 231$, (middle) $d = 60$, $N = 1891$, (bottom) $d = 100$, $N = 5151$.

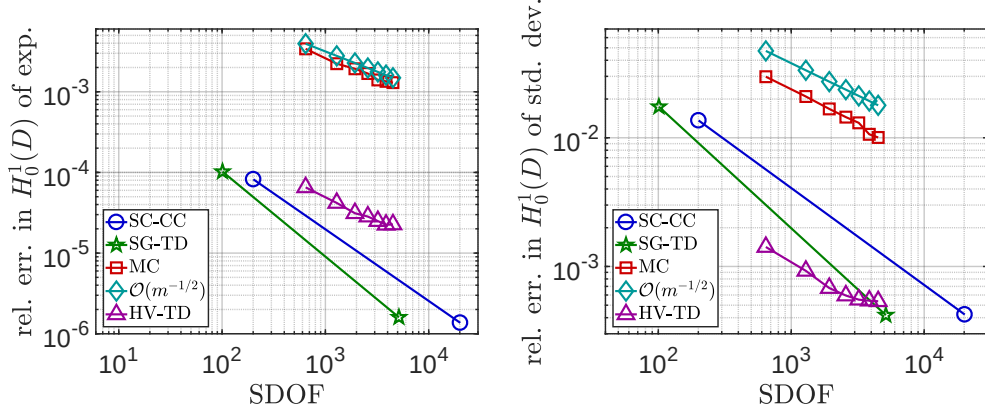


Figure 6.12: Comparison of relative approximation errors $\varepsilon_{h,\text{approx}}^{\text{rel-E}}$ (left) and $\varepsilon_{h,\text{approx}}^{\text{rel-}\sigma}$ (right) from (6.15) for the $L = 2, 3$ stochastic collocation (SC-CC), $p = 1, 2$ stochastic Galerkin (SG-TD), Monte Carlo (MC), and total degree order $p = 2$ with $N = 5151$ sparse HV recovery (HV-TD) methods for solving (2.2) with coefficient (6.18) and correlation length $L_c = 1/2$ in $d = 100$ dimensions.

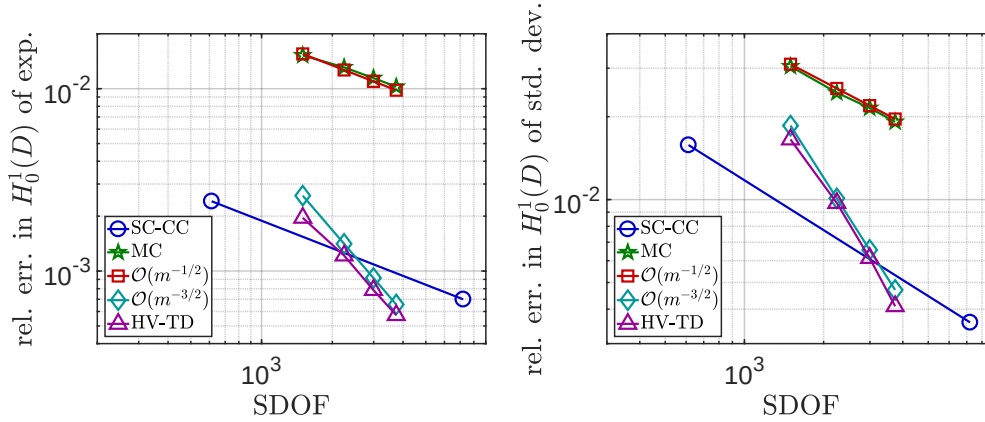


Figure 6.13: Comparison of relative approximation errors $\varepsilon_{h,\text{approx}}^{\text{rel-E}}$ (left) and $\varepsilon_{h,\text{approx}}^{\text{rel-}\sigma}$ (right) from (6.15) for the $L = 2, 3$ stochastic collocation (SC-CC), Monte Carlo (MC), and total degree order $p = 4$ with $N = 5985$ sparse HV recovery (HV-TD) methods for solving (2.2) with coefficient (6.7) and correlation length $L_c = 1/8$ in $d = 17$ dimensions.

Chapter 7

Conclusions

The major contribution presented in this work was the extension of compressed sensing and joint-sparse recovery methods to the recovery of sparse *Hilbert-valued vectors*. Our novel sparse HV recovery approach enables efficient approximation of solutions to parameterized PDEs, under the minimal sample complexity requirements of both methods. Moreover, with quasi-optimal error estimates from [113], we are able to show that our method achieves sub-exponential convergence rates similar to the popular stochastic Galerkin and stochastic collocation methods. Throughout this work, we have addressed a number of key issues related to the computational complexity of methods for approximation and convergence of algorithms. We summarize our achievements as follows.

In Section 3.2, we presented the stochastic Galerkin method, which approximates the solution u to parameterized PDEs by orthogonal projection, obtaining optimal error estimates and fast convergence rates under mild assumptions on the input data. It was known that the computational complexity of the method depends on the order of dependence on the parameters in the model, though a rigorous analysis of this cost in the *general nonlinear* case was lacking in the literature. We addressed this in Section 5.1, proving bounds on the sparsity of the linear systems that result from the stochastic Galerkin discretization in the *general* case, i.e., bounds that hold for *all* parameterizations. Using these bounds to analyze the cost in FLOPs, we show the method enjoys optimal cost in the *affine* case, in agreement with previous results [7, 55]. However, for higher-order nonlinear parameterizations, the optimality of the SG method is lost. Indeed, when discretizing the elliptic PDE (2.2) with the SG method, higher order parameterizations result in denser matrices consisting of fully coupled systems of finite element equations, making matrix-vector products required for solution cumbersome. In Section 6.1, we demonstrate this fact with extensive numerical experiments, comparing performance in FLOPs for a variety of affine and nonlinear parameterizations.

Recently, methods developed in the compressed sensing community for the recovery of real or complex-valued signals have been adapted to the problem of polynomial approximation of high-dimensional functions. Such methods rely on the smooth dependence on parameters and fast decay of coefficients in orthogonal expansions of the target function, enabling highly accurate approximations with minimal sample complexity. Motivated by the polynomial approximation problem and the fact that the most important terms of such expansions of smooth functions often lie in lower sets, in Section 3.5 we focused on improving sample

complexity estimates for sparse polynomial approximation. We achieve this through an extension of chaining arguments of Bourgain [15] from Hadamard to general bounded orthonormal systems, leading to a reduction in the order of dependence on $\log(s)$ in sample complexity estimates by one unit for sparse recovery. To further improve results for smooth function approximation, in Section 3.5.2 we develop a structured sparsity model for polynomial approximation on lower sets, enforced by a new lower restricted isometry property (RIP), a specific case of the weighted RIP from [103]. Finally, in Section 6.2, we presented numerical experiments with a novel weighting strategy, with explicit weight choice enhancing lower set recovery, demonstrating the virtual optimality of our choice of weights on several examples of smooth high-dimensional function approximation.

In Section 3.6, we presented the problem setting and measurement scheme required for sparse HV reconstruction. The key extension from standard compressive sensing is a reformulation of standard basis-pursuit denoising in terms of a mixed $\ell_{V,p}$ -norm which enables energy norm regularized sparse reconstruction of solutions to parameterized PDEs. As described in Section 3.6.3, the problem of Hilbert-valued recovery is intimately related to compressed sensing and joint-sparse recovery, and can be seen as a generalization of those problems. In Section 4.1, we show convergence results for a formulation of forward-backward splitting, relevant to both joint-sparse and sparse HV recovery. Our analysis establishes a novel angular convergence result using the firmly nonexpansive property for soft thresholding operators. With this, and the well-known weak convergence of the method, we are able to prove the strong convergence of the iterates in the infinite dimensional setting. Moreover, in Sections 4.2 and 4.3, we give extensions to the Bregman iterative regularization and fixed point continuation methods, enabling efficient recovery of sparse HV signals.

In Section 5.3, we develop a framework for the approximation of solutions to parameterized PDEs through sparse HV recovery. With only general assumptions on the decay of coefficients of the parametric expansion of the solution, we are able to derive sub-exponential convergence rates for sparse HV recovery with respect to the number of terms s . Such compressibility results have been shown for a variety of parameterized PDE problems [28], including the parameterized parabolic and hyperbolic PDEs. Therefore we expect the sparse HV recovery approach to also perform well in those settings. In Section 6.3 we demonstrate the minimal sample complexity requirements of the method, comparing to popular alternative methods of solution including the SG, MC, and SC methods. Our results reveal the superior approximation capabilities in high-dimension problems, resulting from the ability of compressed sensing-based approximations to naturally detect underlying problem anisotropy.

Bibliography

- [1] Adcock, B. (2017). Infinite-Dimensional ℓ_1 Minimization and Function Approximation from Pointwise Data. *Constructive Approximation*, 45(3):345–390.
- [2] Adcock, B. (2018). Infinite-dimensional compressed sensing and function interpolation. *Foundations of Computational Mathematics*, 18(3):661–701.
- [3] Atkinson, K. E. and Han, W. (2005). *Theoretical Numerical Analysis: A Functional Analysis Framework*, volume 39 of *Texts in Applied Mathematics*. Springer New York, New York, NY, 3 edition.
- [4] Babuska, I., Nobile, F., and Tempone, R. (2007). A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Journal On Numerical Analysis*, 45:1005–1034.
- [5] Babuška, I., Tempone, R., and Zouraris, G. (2004). Galerkin finite element approximations of stochastic elliptic partial differential equations. *SIAM J. Numer. Anal.*, 42:800–825.
- [6] Babuška, I., Tempone, R., and Zouraris, G. E. (2005). Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation. *Comput. Methods Appl. Mech. Engrg.*, 194(12-16):1251–1294.
- [7] Bäck, J., Nobile, F., Tamellini, L., and Tempone, R. (2011). Stochastic Spectral Galerkin and Collocation Methods for PDEs with Random Coefficients: A Numerical Comparison. In Hesthaven, J. S. and Rønquist, E. M., editors, *Spectral and High Order Methods for Partial Differential Equations*, volume 76 of *Lecture Notes in Computational Science and Engineering*, pages 43–62. Springer Berlin Heidelberg.
- [8] Bäck, J., Nobile, F., Tamellini, L., and Tempone, R. (2014). Convergence of quasi-optimal stochastic Galerkin methods for a class of PDEs with random coefficients. *Computers and Mathematics with Applications*, 67(4):732–751.
- [9] Bäck, J., Tempone, R., Nobile, F., and Tamellini, L. (2012). On the Optimal Polynomial Approximation of Stochastic PDEs by Galerkin and Collocation Methods. *Mathematical Models and Methods in Applied Sciences*, 22(09):1250023 (33 pages).
- [10] Bauschke, H. H. and Combettes, P. L. (2010). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing.
- [11] Bellman, R. (1957). *Dynamic Programming*. Rand Corporation research study. Princeton University Press.
- [12] Bellman, R. and Bellman, R. E. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton Legacy Library. Princeton University Press.
- [13] Biegler, L. T., Ghattas, O., Heinkenschloss, M., and van Bloemen Waanders, B. (2003). Large-scale pde-constrained optimization: An introduction. In Biegler, L. T., Heinkenschloss, M., Ghattas, O., and van Bloemen Waanders, B., editors, *Large-Scale PDE-Constrained Optimization*, pages 3–13, Berlin, Heidelberg. Springer Berlin Heidelberg.

- [14] Bieri, M., Andreev, R., and Schwab, C. (2009). Sparse tensor discretization of elliptic spdes. *SIAM J. Sci. Comput.*, 31:4281–4304.
- [15] Bourgain, J. (2014). An improved estimate in the restricted isometry problem. In *Geometric Aspects of Functional Analysis*, Lecture Notes in Mathematics, pages 65–70.
- [16] Bregman, L. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming.
- [17] Brenner, S. C. and Scott, L. R. (1994). *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag, New York.
- [18] Bruck Jr., R. E. (1975). An iterative solution of a variational inequality for certain monotone operators in Hilbert space. *Bulletin of the American Mathematical Society*, 81(5):890–893.
- [19] Burger, M. and Osher, S. (2013). *A Guide to the TV Zoo*, pages 1–70. Springer International Publishing, Cham.
- [20] Candès, E., Romberg, J., , and Tao, T. (2006). Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(1):489–509.
- [21] Candès, E. and Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. on Information Theory*, 52(12):5406–5425.
- [22] Chen, G. H.-G. and Rockafellar, R. T. (1997). Convergence Rates in Forward–Backward Splitting. *SIAM Journal on Optimization*, 7(2):421–444.
- [23] Cheraghchi, M., Guruswami, V., and A., V. (2013). Restricted isometry of Fourier matrices and list decodability of random linear codes. In *SODA*, pages 432–442.
- [24] Chernov, A. and Dũng, D. (2016). New explicit-in-dimension estimates for the cardinality of high-dimensional hyperbolic crosses and approximation of functions having mixed smoothness. *Journal of Complexity*, 32:92–121.
- [25] Chkifa, A., Cohen, A., DeVore, R., and Schwab, C. (2013). Sparse adaptive taylor approximation algorithms for parametric and stochastic elliptic pdes. *Modél. Math. Anal. Numér.*, 47(1):253–280.
- [26] Chkifa, A., Cohen, A., Migliorati, G., Nobile, F., and Tempone, R. (2015a). Discrete least squares polynomial approximation with random evaluations - application to parametric and stochastic elliptic PDEs. *ESAIM: M2AN*, 49(3):815–837.
- [27] Chkifa, A., Cohen, A., and Schwab, C. (2014). High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs. *Foundations of Computational Mathematics*, 14(4):601–633.

- [28] Chkifa, A., Cohen, A., and Schwab, C. (2015b). Breaking the curse of dimensionality in sparse polynomial approximation of parametric pdes. *J. Math. Pures Appl.*, 103(2):400–428.
- [29] Chkifa, A., Dexter, N., Tran, H., and Webster, C. G. (2018). Polynomial approximation via compressed sensing of high-dimensional functions on lower sets. *Mathematics of Computation*, 87(311):1415–1450.
- [30] Clenshaw, C. W. and Curtis, A. R. (1960). A method for numerical integration on an automatic computer. *Numerische Mathematik*, 2(1):197–205.
- [31] Cohen, A., Davenport, M. A., and Leviatan, D. (2013). On the stability and accuracy of least squares approximations. *Found Comput Math*, 13:819–834.
- [32] Cohen, A. and DeVore, R. (2015). Approximation of high-dimensional parametric PDEs. *Acta Numer.*, 24:1–159.
- [33] Cohen, A., DeVore, R., and Schwab, C. (2010). Convergence rates of best n-term galerkin approximations for a class of elliptic spdes. *Found Comput Math*, 10:615–646.
- [34] Cohen, A., DeVore, R., and Schwab, C. (2011). Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDEs. *Analysis and Applications*, 9(1):11–47.
- [35] Cohen, A., Migliorati, G., and Nobile, F. (2016). Discrete least-squares approximations over optimized downward closed polynomial spaces in arbitrary dimension. *Constructive Approximation*, page to appear.
- [36] Combettes, P. L. (2004). Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, 53(5-6):475–504.
- [37] Cotter, S., Rao, B. D., and Kreutz-Delgado, K. (2005). Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Transactions on Signal Processing*, 53(7):2477–2488.
- [38] Daubechies, I., Defrise, M., and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457.
- [39] Davies, M. and Eldar, Y. (2012). Rank awareness in joint sparse recovery. *IEEE Trans. Inform. Theory*, 58(2):1135–1146.
- [40] Davis, P. J. (1975). *Interpolation and Approximation*. Dover.
- [41] De Marchi, S. (2004). On Leja sequences: Some results and applications. *Applied Mathematics and Computation*, 152(3):621–647.
- [42] Deb, M., Babuška, I., and Oden, J. (2001). Solution of stochastic partial differential equations using Galerkin finite element techniques. *Comput. Methods Appl. Mech. Engrg.*, 190(48):6359–6372.

- [43] DeVore, R. (1998). Nonlinear approximation. *Acta. Numer*, 7:51–150.
- [44] DeVore, R. A., Howard, R., and Micchelli, C. (1989). Optimal nonlinear approximation. *manuscripta mathematica*, 63(4):469–478.
- [45] Dexter, N., Tran, H., and Webster, C. G. (2017). On the strong convergence of forward-backward splitting in reconstructing jointly sparse signals. *Submitted*.
- [46] Dexter, N., Webster, C. G., and Zhang, G. (2016). Explicit cost bounds of stochastic Galerkin approximations for parameterized PDEs with random coefficients. *Computers and Mathematics with Applications*, 71(11):2231–2256.
- [47] Donoho, D. L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306.
- [48] Doostan, A. and Owhadi, H. (2011). A non-adapted sparse approximation of PDEs with stochastic inputs. *Journal of Computational Physics*, 230:3015–3034.
- [49] Dunford, N. and Schwartz, J. (1958). *Linear Operators: General theory*. Pure and applied mathematics. Interscience Publishers.
- [50] Durrett, R. (1996). *Probability: theory and examples*. Duxbury Press, Belmont, CA, second edition.
- [51] Dzjadyk, V. K. and Ivanov, V. V. (1983). On asymptotics and estimates for the uniform norms of the Lagrange interpolation polynomials corresponding to the Chebyshev nodal points. *Analysis Mathematica*, 9(2):85–97.
- [52] Eiermann, M., Ernst, O. G., and Ullmann, E. (2007). Computational aspects of the stochastic finite element method. In *Computing and Visualization in Science*, volume 10, pages 3–15.
- [53] Eldar, Y. and Rauhut, H. (2010). Average case analysis of multichannel sparse recovery using convex relaxation. *IEEE Trans. Inform. Theory*, 56(1):505–519.
- [54] Elman, H. C., Miller, C. W., Phipps, E. T., and Tuminaro, R. S. (2011). Assessment of collocation and Galerkin approaches to linear diffusion equations with random data. *International Journal for Uncertainty Quantification*, 1(1):19–33.
- [55] Ernst, O. G. and Ullmann, E. (2010). Stochastic Galerkin Matrices. *SIAM Journal on Matrix Analysis and Applications*, 31(4):1848–1872.
- [56] Evans, L. (2012). *An Introduction to Stochastic Differential Equations*. American Mathematical Society.
- [57] Evans, L. C. (2010). *Partial differential equations*. American Mathematical Society, Providence, R.I.
- [58] Fornasier, M. and Rauhut, H. (2008). Recovery algorithms for vector-valued data with joint sparsity constraints. *SIAM J. Numer. Anal.*, 46(2):577–613.

- [59] Foucart, S. and Rauhut, H. (2013). *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser.
- [60] Galindo, D., Jantsch, P., Webster, C. G., and Zhang, G. (2016). Accelerating stochastic collocation methods for partial differential equations with random input data. *SIAM/ASA Journal on Uncertainty Quantification*, 4:1111–1137.
- [61] Ghanem, R. and Spanos, P. (2002). *Stochastic Finite Elements: A Spectral Approach*. Dover, New York, 2nd edition.
- [62] Goldstein, A. A. (1964). Convex Programming in Hilbert Space. *Bulletin of the American Mathematical Society*, 70(5):709–710.
- [63] Goldstein, T. and Osher, S. (2009). The Split Bregman Method for L1-Regularized Problems. *SIAM Journal on Imaging Sciences*, 2(2):323–343.
- [64] Gribonval, R., Rauhut, H., Schnass, K., and Vandergheynst, P. (2008). Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms. *Journal of Fourier Analysis and Applications*, 14(5-6):655–687.
- [65] Gunzburger, M., Webster, C. G., and Zhang, G. (2014). Stochastic finite element methods for partial differential equations with random input data. *Acta Numerica*, 23:521–650.
- [66] Guo, Z., Wittman, T., and Osher, S. (2009). L1 unmixing and its application to hyperspectral image enhancement. In *Proc.SPIE*, number April 2009 in 73341M.
- [67] Hale, E., Yin, W., and Zhang, Y. (2008). Fixed-point continuation for ℓ_1 -minimization: methodology and convergence. *SIAM J. Optim.*, 19(3):1107–1130.
- [68] Hale, E., Yin, W., and Zhang, Y. (2010). Fixed-point continuation applied to compressed sensing: implementation and numerical experiments. *J. Comput. Math.*, 28(2):170–194.
- [69] Hale, E. T., Yin, W., and Zhang, Y. (2007). A Fixed-Point Continuation Method for ℓ_1 -Regularized Minimization with Applications to Compressed Sensing. Technical report, Rice University, Houston, Texas.
- [70] Hampton, J. and Doostan, A. (2015). Compressive sampling of polynomial chaos expansions: Convergence analysis and sampling strategies. *Journal of Computational Physics*, 280:363–386.
- [71] Hansen, M. and Schwab, C. (2013a). Analytic regularity and nonlinear approximation of a class of parametric semilinear elliptic PDEs. *Math. Nachr.*, 286(8–9):832–860.
- [72] Hansen, M. and Schwab, C. (2013b). Sparse adaptive approximation of high dimensional parametric initial value problems. *Vietnam Journal of Mathematics*, 41(2):181–215.

- [73] Haviv, I. and Regev, O. (2016). The Restricted Isometry Property of Subsampled Fourier Matrices. *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 288–297.
- [74] Helton, J. C. and Davis, F. J. (2003). Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering and System Safety*, 81(1):23–69.
- [75] Hoang, V. and Schwab, C. (2012). Regularity and generalized polynomial chaos approximation of parametric and random 2nd order hyperbolic partial differential equations. *Anal. Appl.*, 10(3):295–326.
- [76] Hoang, V. and Schwab, C. (2014). N -term Wiener Chaos Approximation Rates for elliptic PDEs with lognormal Gaussian random inputs. *Math. Models Methods Appl. Sci.* 24, 797, 24(4).
- [77] Hoang, V. H. and Schwab, C. (2013). Sparse Tensor Galerkin Discretizations for Parametric and Random Parabolic PDEs - Analytic Regularity and Generalized Polynomial Chaos Approximation. *SIAM J. Mathematical Analysis*, 45(5):3050–3083.
- [78] Holden, H., Øksendal, B., Ubøe, J., and Zhang, T. (1996). *Stochastic partial differential equations*, pages 141–191. Birkhäuser Boston, Boston, MA.
- [79] Jantsch, P., Webster, C. G., and Zhang, G. (2018). On the Lebesgue constant of weighted Leja points for Lagrange interpolation on unbounded domains. *IMA Journal of Numerical Analysis*, page dry002.
- [80] Jantsch, P. A. (2017). *Efficient Methods for Multidimensional Global Polynomial Approximation with Applications to Random PDEs*. PhD diss., University of Tennessee.
- [81] Johnson, C. (2012). *Numerical solution of partial differential equations by the finite element method*. Courier Dover Publications.
- [82] Kühn, T., Sickel, W., and Ullrich, T. (2015). Approximation of Mixed Order Sobolev Functions on the d -Torus: Asymptotics, Preasymptotics, and d -Dependence. *Constr Approx*, 42:353–398.
- [83] Lai, M. J. and Liu, Y. (2011). The null space property for sparse recovery from multiple measurement vectors. *Applied and Computational Harmonic Analysis*, 30(3):402–406.
- [84] Le Maître, O. P. and Knio, O. M. (2010). *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*. Scientific Computation. Springer.
- [85] Loève, M. (1978). *Probability Theory*. Number v. 2 in Graduate Texts in Mathematics. Springer.
- [86] Mathelin, L. and Gallivan, K. (2012). A compressed sensing approach for partial differential equations with random input data. *Commun. Comput. Phys.*, 12:919–954.

- [87] Matthies, H. G. and Keese, A. (2005). Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 194:1295–1331.
- [88] Mishali, M. and Eldar, Y. (2008). Reduce and boost: Recovering arbitrary sets of jointly sparse vectors. *IEEE Transactions in Signal Processing*, 56(10):4692–4702.
- [89] Narayan, A. and Zhou, T. (2015). Stochastic collocation on unstructured multivariate meshes. *Commun. Comput. Phys.*, 18(1):1–36.
- [90] Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pennsylvania, 1st edition.
- [91] Nobile, F., Tempone, R., and Webster, C. G. (2008a). A Sparse Grid Stochastic Collocation Method for Partial Differential Equations with Random Input Data. *SIAM Journal on Numerical Analysis*, 46(5):2309–2345.
- [92] Nobile, F., Tempone, R., and Webster, C. G. (2008b). An Anisotropic Sparse Grid Stochastic Collocation Method for Partial Differential Equations with Random Input Data. *SIAM Journal on Numerical Analysis*, 46:2411–2442.
- [93] Osher, S., Burger, M., Goldfarb, D., Xu, J., and Yin, W. (2005). An iterated regularization method for total variation-based image restoration. *Multiscale Model. Simul.*, 4:460–489.
- [94] Pellissetti, M. F. and Ghanem, R. G. (2000). Iterative Solution of Systems of Linear Equations Arising in the Context of Stochastic Finite Elements. *Advances in Engineering Software*, 31:607–616.
- [95] Peng, J., Hampton, J., and Doostan, A. (2014). A weighted ℓ_1 -minimization approach for sparse polynomial chaos expansions. *Journal of Computational Physics*, 267:92–111.
- [96] Peng, J., Hampton, J., and Doostan, A. (2016). On polynomial chaos expansion via gradient-enhanced ℓ_1 -minimization. *Journal of Computational Physics*, 310:440–458.
- [97] Persson, P.-O. and Strang, G. (2004). A Simple Mesh Generator in MATLAB. *SIAM Review*, 46(2):329–345.
- [98] Powell, C., Lord, G., and Shardlow, T. (2014). *An Introduction to Computational Stochastic PDEs*. Texts in Applied Mathematics. Cambridge University Press, United Kingdom, 1 edition.
- [99] Powell, C. E. and Elman, H. C. (2009). Block-diagonal preconditioning for spectral stochastic finite-element systems. *IMA Journal of Numerical Analysis*, 29(2):350–375.
- [100] Rauhut, H. (2010). Compressive sensing and structured random matrices. In Fornasier, M., editor, *Theoretical Foundations and Numerical Methods for Sparse Recovery*, volume 9 of *Radon Ser. Comput. Appl. Math.*, pages 1–92. de Gruyter.

- [101] Rauhut, H. and Schwab, C. (2017). Compressive sensing Petrov-Galerkin approximation of high-dimensional parametric operator equations. *Mathematics of Computation*, 86(304):661–700.
- [102] Rauhut, H. and Ward, R. (2012). Sparse legendre expansions via ℓ_1 -minimization. *Journal of Approximation Theory*, 164:517–533.
- [103] Rauhut, H. and Ward, R. (2016). Interpolation via weighted ℓ_1 -minimization. *Applied and Computational Harmonic Analysis*, 40(2):321–351.
- [104] Rudelson, M. and Vershynin, R. (2008). On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61:1025–1045.
- [105] Smolyak, S. A. (1963). Quadrature and interpolation formulas for tensor products of certain classes of functions. *Dokl. Akad. Nauk SSSR*, 148(5):1042–1045.
- [106] Stojnic, M., Parvaresh, F., and Hassibi, B. (2009). On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Transactions on Signal Processing*, 57(8):3075–3085.
- [107] Stoyanov, M. (2015). User manual: Tasmanian sparse grids. Technical Report ORNL/TM-2015/596, Oak Ridge National Laboratory, One Bethel Valley Road, Oak Ridge, TN.
- [108] Stoyanov, M. K. and Webster, C. G. (2016). A dynamically adaptive sparse grids method for quasi-optimal interpolation of multidimensional functions. *Computers & Mathematics with Applications*, 71(11):2449–2465.
- [109] Szegő, G. (1975). *Orthogonal polynomials*, volume XXIII. Amer. Math. Soc., 4 edition.
- [110] Szlam, A., Guo, Z., and Osher, S. (2010). A split Bregman method for non-negative sparsity penalized least squares with applications to hyperspectral demixing. *Proceedings - International Conference on Image Processing, ICIP*, pages 1917–1920.
- [111] Temlyakov, V. N. (1998). Nonlinear Kolmogorov widths. *Mathematical Notes*, 63(6):785–795.
- [112] Todor, R. A. and Schwab, C. (2007). Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients. *IMA Journal of Numerical Analysis*, 27(2):232–261.
- [113] Tran, H., Webster, C. G., and Zhang, G. (2017). Analysis of quasi-optimal polynomial approximations for parameterized PDEs with deterministic and stochastic coefficients. *Numerische Mathematik*, 137(2):451–493.
- [114] Tropp, J. (2006). Algorithms for simultaneous sparse approximation. part ii: Convex relaxation. *Signal Processing*, 86:589–602.
- [115] Tropp, J., Gilbert, A., and Strauss, M. (2006). Algorithms for simultaneous sparse approximation. part i: Greedy pursuit. *Signal Processing*, 86:572–588.

- [116] Ullmann, E. (2010). A Kronecker Product Preconditioner for Stochastic Galerkin Finite Element Discretizations. *SIAM Journal on Scientific Computing*, 32(2):923–946.
- [117] Ullmann, E., Elman, H. C., and Ernst, O. G. (2012). Efficient Iterative Solvers for Stochastic Galerkin Discretizations of Log-Transformed Random Diffusion Problems. *SIAM Journal on Scientific Computing*, 34(2):A659–A682.
- [118] van den Berg, E. and Friedlander, M. P. (2007). SPGL1: A solver for large-scale sparse reconstruction. <http://www.cs.ubc.ca/labs/scl/spgl1>.
- [119] van den Berg, E. and Friedlander, M. P. (2008). Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912.
- [120] van der Berg, E. and Friedlander, M. (2010). Theoretical and empirical results for recovery from multiple measurements. *IEEE Trans. Inform. Theory*, 56(5).
- [121] Wasilkowski, G. W. and Woźniakowski, H. (1995). Explicit Cost Bounds of Algorithms for Multivariate Tensor Product Problems. *Journal of Complexity*.
- [122] Xiu, D. (2010). *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, Princeton, NJ, USA.
- [123] Xiu, D. and Karniadakis, G. E. (2002a). Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos. *Computer Methods in Applied Mechanics and Engineering*, 191:4927–4948.
- [124] Xiu, D. and Karniadakis, G. E. (2002b). The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2):619–644.
- [125] Yan, L., Guo, L., and Xiu, D. (2012). Stochastic collocation algorithms using ℓ_1 -minimization. *International Journal for Uncertainty Quantification*, 2(3):279–293.
- [126] Yang, X. and Karniadakis, G. E. (2013). Reweighted ℓ_1 -minimization method for stochastic elliptic differential equations. *Journal of Computational Physics*, 248:87–108.
- [127] Yin, W. (2010). Analysis and Generalizations of the Linearized Bregman Method. *SIAM Journal on Imaging Sciences*, 3(4):856–877.
- [128] Yin, W. and Osher, S. (2013). Error forgetting of Bregman iteration. *Journal of Scientific Computing*, 54(2-3):684–695.
- [129] Yin, W., Osher, S., Goldfarb, D., and Darbon, J. (2008). Bregman iterative algorithms for ℓ_1 -minimization with applications to compressed sensing. *SIAM J. Imaging Sci.*, 1(1):143–168.
- [130] Zhu, W. (2017). *Nonlocal Variational Methods in Image and Data Processing*. Phd diss., UCLA.

Appendices

Appendix A

Proofs from the text

A Proofs of results from Chapter 4

In this section, we include proofs from Chapter 4 that are either too long to include in the regular text, or do not add significant insight to the discussion. Several of the proofs from this section can be found in other works, with minor differences.

Proof of Proposition 4.16. The following is the proof of [129, Proposition 3.2], restated with the $\|\cdot\|_{V,q}$ norms for $q = 1, 2$ and corresponding inner product $\langle \cdot, \cdot \rangle_{V,2}$. From the definition of the subgradient (see (4.2)) and because $\mathbf{z}^k$ is a minimizer of the objective function of step 3 of Algorithm 1 we have

$$\begin{aligned} H(\mathbf{z}^k) &\leq H(\mathbf{z}^k) + J(\mathbf{z}^k) - J(\mathbf{z}^{k-1}) - \langle \mathbf{p}^{k-1}, \mathbf{z}^k - \mathbf{z}^{k-1} \rangle_{V,2} \\ &= D_J^{\mathbf{p}^{k-1}}(\mathbf{z}^k, \mathbf{z}^{k-1}) + H(\mathbf{z}^k) \\ &\leq D_J^{\mathbf{p}^{k-1}}(\mathbf{z}^{k-1}, \mathbf{z}^{k-1}) + H(\mathbf{z}^{k-1}) \\ &= H(\mathbf{z}^{k-1}), \end{aligned}$$

which implies part 1. Now if $J(\tilde{\mathbf{z}}) < \infty$, then since

$$\begin{aligned} D_J^{\mathbf{p}^k}(\tilde{\mathbf{z}}, \mathbf{z}^k) - D_J^{\mathbf{p}^{k-1}}(\tilde{\mathbf{z}}, \mathbf{z}^{k-1}) + D_J^{\mathbf{p}^{k-1}}(\mathbf{z}^k, \mathbf{z}^{k-1}) \\ &= J(\tilde{\mathbf{z}}) - J(\mathbf{z}^k) + \langle \mathbf{z}^k - \tilde{\mathbf{z}}, \mathbf{p}^k \rangle_{V,2} \\ &\quad - J(\tilde{\mathbf{z}}) + J(\mathbf{z}^{k-1}) - \langle \mathbf{z}^{k-1} - \tilde{\mathbf{z}}, \mathbf{p}^{k-1} \rangle_{V,2} \\ &\quad + J(\mathbf{z}^k) - J(\mathbf{z}^{k-1}) + \langle \mathbf{z}^{k-1} - \mathbf{z}^k, \mathbf{p}^{k-1} \rangle_{V,2} \\ &= \langle \mathbf{z}^k - \tilde{\mathbf{z}}, \mathbf{p}^k - \mathbf{p}^{k-1} \rangle_{V,2} \end{aligned}$$

Letting $-\mathbf{q}^k = \mathbf{p}^k - \mathbf{p}^{k-1} \in \partial H(\mathbf{z}^k)$, and using the properties of subgradients, we obtain

$$\begin{aligned} D_J^{\mathbf{p}^k}(\tilde{\mathbf{z}}, \mathbf{z}^k) - D_J^{\mathbf{p}^{k-1}}(\tilde{\mathbf{z}}, \mathbf{z}^{k-1}) + D_J^{\mathbf{p}^{k-1}}(\mathbf{z}^k, \mathbf{z}^{k-1}) &= \langle \mathbf{q}^k, \tilde{\mathbf{z}} - \mathbf{z}^k \rangle_{V,2} \\ &\leq H(\tilde{\mathbf{z}}) - H(\mathbf{z}^k), \end{aligned}$$

or, after rearranging,

$$D_J^{\mathbf{p}^k}(\tilde{\mathbf{z}}, \mathbf{z}^k) + D_J^{\mathbf{p}^{k-1}}(\mathbf{z}^k, \mathbf{z}^{k-1}) + H(\mathbf{z}^k) \leq H(\tilde{\mathbf{z}}) + D_J^{\mathbf{p}^{k-1}}(\tilde{\mathbf{z}}, \mathbf{z}^{k-1}). \quad (\text{A.1})$$

If $\tilde{\mathbf{z}}$ is also a minimizer of H , then

$$\begin{aligned} D_J^{\mathbf{p}^k}(\tilde{\mathbf{z}}, \mathbf{z}^k) &\leq D_J^{\mathbf{p}^k}(\tilde{\mathbf{z}}, \mathbf{z}^k) + D_J^{\mathbf{p}^{k-1}}(\mathbf{z}^k, \mathbf{z}^{k-1}) \\ &\leq D_J^{\mathbf{p}^k}(\tilde{\mathbf{z}}, \mathbf{z}^k) + D_J^{\mathbf{p}^{k-1}}(\mathbf{z}^k, \mathbf{z}^{k-1}) - H(\mathbf{z}^k) + H(\tilde{\mathbf{z}}) \\ &\leq D_J^{\mathbf{p}^{k-1}}(\tilde{\mathbf{z}}, \mathbf{z}^{k-1}). \end{aligned}$$

so that the Bregman distances to $\tilde{\mathbf{z}}$ are monotonically nonincreasing. Subtracting $H(\tilde{\mathbf{z}})$ from both sides of (A.1) and summing, we have

$$D_J^{\mathbf{p}^k}(\tilde{\mathbf{z}}, \mathbf{z}^k) + \sum_{j=1}^k \left[D_J^{\mathbf{p}^{j-1}}(\mathbf{z}^j, \mathbf{z}^{j-1}) + H(\mathbf{z}^j) - H(\tilde{\mathbf{z}}) \right] \leq D^0(\tilde{\mathbf{z}}, \mathbf{z}^0) = J(\tilde{\mathbf{z}}).$$

From the nonnegativity of the Bregman distance and the monotonicity of $H(\mathbf{z}^j)$ due to part 1, it follows that

$$D^{\mathbf{p}^k}(\tilde{\mathbf{z}}, \mathbf{z}^k) + k[H(\mathbf{z}^k) - H(\tilde{\mathbf{z}})] \leq J(\tilde{\mathbf{z}}),$$

implying part 2.

Now if $\mathbf{u} = \mathbf{A}\tilde{\mathbf{z}}$, and $J(\mathbf{z}) := \|\mathbf{z}\|_{\mathcal{V},1}$ and $H(\mathbf{z}) := \frac{\mu}{2}\|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2}^2$ for $\mathbf{z} \in \mathcal{H}$, then from parts 1 and 2 we have

$$kH(\mathbf{z}^k) \leq \sum_{j=1}^k H(\mathbf{z}^j) \leq J(\tilde{\mathbf{z}}).$$

Hence

$$\begin{aligned} J(\tilde{\mathbf{z}}) &\geq \sum_{j=1}^k D_J^{\mathbf{p}^{j-1}}(\mathbf{z}^j, \mathbf{z}^{j-1}) = J(\mathbf{z}^k) - \sum_{j=1}^k \langle \mathbf{p}^{j-1}, \mathbf{z}^j - \mathbf{z}^{j-1} \rangle_{\mathcal{V},2} \\ &= J(\mathbf{z}^k) - J(\mathbf{z}^0) - \langle \mathbf{p}^{k-1}, \mathbf{z}^k - \tilde{\mathbf{z}} \rangle_{\mathcal{V},2} + \sum_{j=1}^{k-1} \langle \mathbf{p}^j - \mathbf{p}^{j-1}, \mathbf{z}^j - \tilde{\mathbf{z}} \rangle_{\mathcal{V},2} \\ &= J(\mathbf{z}^k) - \sum_{j=1}^{k-1} \langle \mathbf{q}^j, \mathbf{z}^k - \tilde{\mathbf{z}} \rangle_{\mathcal{V},2} - \sum_{j=1}^{k-1} \langle \mathbf{q}^j, \mathbf{z}^j - \tilde{\mathbf{z}} \rangle_{\mathcal{V},2}. \end{aligned}$$

Since $\mathbf{q}^j = \mathbf{A}^*(\mathbf{A}\mathbf{z}^j - \mathbf{u})$, we have

$$\begin{aligned}
J(\tilde{\mathbf{z}}) &\geq J(\mathbf{z}^k) - \sum_{j=1}^{k-1} \langle \mathbf{A}\mathbf{z}^j - \mathbf{u}, \mathbf{A}\mathbf{z}^k - \mathbf{u} \rangle_{\mathcal{V},2} - \sum_{j=1}^{k-1} \langle \mathbf{A}\mathbf{z}^j - \mathbf{u}, \mathbf{A}\mathbf{z}^j - \mathbf{u} \rangle_{\mathcal{V}} \\
&= J(\mathbf{z}^k) - \sum_{j=1}^{k-1} \langle \mathbf{A}\mathbf{z}^j - \mathbf{u}, \mathbf{A}\mathbf{z}^k - \mathbf{u} \rangle_{\mathcal{V},2} - \sum_{j=1}^{k-1} \|\mathbf{A}\mathbf{z}^j - \mathbf{u}\|_{\mathcal{V},2}^2 \\
&\geq J(\mathbf{z}^k) - \sum_{j=1}^{k-1} \|\mathbf{A}\mathbf{z}^j - \mathbf{u}\|_{\mathcal{V},2} \|\mathbf{A}\mathbf{z}^k - \mathbf{u}\|_{\mathcal{V},2} - \sum_{j=1}^{k-1} \|\mathbf{A}\mathbf{z}^j - \mathbf{u}\|_{\mathcal{V},2}^2 \\
&\geq J(\mathbf{z}^k) - \sum_{j=1}^{k-1} \|\mathbf{A}\mathbf{z}^k - \mathbf{u}\|_{\mathcal{V},2}^2 - \sum_{j=1}^{k-1} \|\mathbf{A}\mathbf{z}^j - \mathbf{u}\|_{\mathcal{V},2}^2 \\
&\geq J(\mathbf{z}^k) - k \|\mathbf{A}\mathbf{z}^k - \mathbf{u}\|_{\mathcal{V},2}^2 - \sum_{j=1}^{k-1} \|\mathbf{A}\mathbf{z}^j - \mathbf{u}\|_{\mathcal{V},2}^2 \\
&\geq J(\mathbf{z}^k) - 2J(\tilde{\mathbf{z}})
\end{aligned}$$

showing that $J(\mathbf{z}^k) \leq 3J(\tilde{\mathbf{z}})$. Hence, when $J(\mathbf{z}^k) = \|\mathbf{z}^k\|_{\mathcal{V},1}$, we see that $\{\mathbf{z}^k\}$ is uniformly bounded in the reflexive Banach space $(\mathcal{V}^N, \|\cdot\|_{\mathcal{V},1})$, and the existence of weak-* convergent subsequences follows from Theorem 3.16 and the Banach-Alaoglu Theorem. That the limit of such subsequences satisfy $\mathbf{A}\mathbf{z} = \mathbf{u}$ follows from part 2. ■

Proof of Theorem 4.17. Let $\mathbf{z}^k$ and $\bar{\mathbf{z}}^k$ denote the solutions to Version 1 and 2, respectively, at the k -th iteration. From (4.37), $D_J^{\mathbf{p}^0}(\mathbf{z}, \mathbf{z}^0) = J(\mathbf{z})$, and (4.40) gives $\mathbf{u}^1 = \mathbf{u}$. Hence, at $k = 0$, (4.38) and (4.42) solve the same problem:

$$\min_{\mathbf{z} \in \mathcal{V}^N} J(\mathbf{z}) + \frac{1}{2} \|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2}^2.$$

Since this problem may have more than one solution, we do not assume $\mathbf{z}^1 = \bar{\mathbf{z}}^1$, and instead rely on Lemma 4.8, noting that $\mathbf{A}^*(\mathbf{u} - \mathbf{A}\mathbf{z}^1) = \mathbf{A}^*(\mathbf{u} - \mathbf{A}\bar{\mathbf{z}}^1)$. Therefore, at step (4.39) we have

$$\mathbf{p}^1 = \mathbf{p}^0 - \mathbf{A}^*(\mathbf{A}\mathbf{z}^1 - \mathbf{u}) = \mathbf{A}^*(\mathbf{u} - \mathbf{A}\mathbf{z}^1) = \mathbf{A}^*(\mathbf{u} - \mathbf{A}\bar{\mathbf{z}}^1) = \mathbf{A}^*(\mathbf{u}^1 - \mathbf{A}\bar{\mathbf{z}}^1).$$

Assume that $\mathbf{p}^k = \mathbf{A}^*(\mathbf{u}^k - \mathbf{A}\bar{\mathbf{z}}^k)$, we show the following: (i) the optimization problems in (4.38) and (4.42) at iteration k are equivalent, (ii) $\mathbf{A}^*(\mathbf{A}\mathbf{z}^{k+1} - \mathbf{u}) = \mathbf{A}^*(\mathbf{A}\bar{\mathbf{z}}^{k+1} - \mathbf{u})$, and (iii) $\mathbf{p}^{k+1} = \mathbf{A}^*(\mathbf{u}^{k+1} - \mathbf{A}\bar{\mathbf{z}}^{k+1})$.

Part (i): we have

$$\begin{aligned}
D_J^{p^k}(\mathbf{z}, \mathbf{z}^k) + \frac{1}{2} \|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2}^2 &= J(\mathbf{z}) - \langle \mathbf{p}^k, \mathbf{z} \rangle_{\mathcal{V},2} + \frac{1}{2} \|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2}^2 + C_1 \\
&= J(\mathbf{z}) - \langle \mathbf{u}^k - \mathbf{A}\bar{\mathbf{z}}^k, \mathbf{A}\mathbf{z} \rangle_{\mathcal{V},2} + \frac{1}{2} \|\mathbf{A}\mathbf{z} - \mathbf{u}\|_{\mathcal{V},2}^2 + C_2 \\
&= J(\mathbf{z}) + \frac{1}{2} \|\mathbf{A}\mathbf{z} - (\mathbf{u} + (\mathbf{u}^k - \mathbf{A}\bar{\mathbf{z}}^k))\|_{\mathcal{V},2}^2 + C_3 \\
&= J(\mathbf{z}) + \frac{1}{2} \|\mathbf{A}\mathbf{z} - \mathbf{u}^{k+1}\|_{\mathcal{V},2}^2 + C_3,
\end{aligned}$$

where C_1 , C_2 , and C_3 are constant in $\mathbf{z}$. Hence steps (4.38) & (4.42) of Versions 1 & 2, respectively, have the same objective function up to a constant in $\mathbf{z}$.

Part (ii) The equivalence of $\mathbf{A}^*(\mathbf{A}\mathbf{z}^{k+1} - \mathbf{u})$ and $\mathbf{A}^*(\mathbf{A}\bar{\mathbf{z}}^{k+1} - \mathbf{u})$ follows from part (i) and Lemma 4.8.

Part (iii) From the induction assumption, as well as (4.39), (4.41), and part (i), we have

$$\begin{aligned}
\mathbf{p}^{k+1} &= \mathbf{p}^k - \mathbf{A}^*(\mathbf{A}\mathbf{z}^{k+1} - \mathbf{u}) = \mathbf{p}^k - \mathbf{A}^*(\mathbf{A}\bar{\mathbf{z}}^{k+1} - \mathbf{u}) \\
&= \mathbf{A}^*(\mathbf{u}^k - \mathbf{A}\bar{\mathbf{z}}^k) - \mathbf{A}^*(\mathbf{A}\bar{\mathbf{z}}^{k+1} - \mathbf{u}) \\
&= \mathbf{A}^*(\mathbf{u} + (\mathbf{u}^k - \mathbf{A}\bar{\mathbf{z}}^k) - \mathbf{A}\bar{\mathbf{z}}^{k+1}) \\
&= \mathbf{A}^*(\mathbf{u}^{k+1} - \mathbf{A}\bar{\mathbf{z}}^{k+1}).
\end{aligned}$$

This completes the proof. ■

Proof of Theorem 4.18. For any $\mathbf{z}$, by the nonnegativity of the Bregman distance, we have

$$\begin{aligned}
J(\mathbf{z}^k) &\leq J(\mathbf{z}) - \langle \mathbf{z} - \mathbf{z}^k, \mathbf{p}^k \rangle_{\mathcal{V},2} \\
&= J(\mathbf{z}) - \langle \mathbf{z} - \mathbf{z}^k, \mathbf{A}^*(\mathbf{u}^k - \mathbf{A}\mathbf{z}^k) \rangle_{\mathcal{V},2} \\
&= J(\mathbf{z}) - \langle \mathbf{A}\mathbf{z} - \mathbf{A}\mathbf{z}^k, \mathbf{u}^k - \mathbf{A}\mathbf{z}^k \rangle_{\mathcal{V},2} \\
&= J(\mathbf{z}) - \langle \mathbf{A}\mathbf{z} - \mathbf{u}, \mathbf{u}^k - \mathbf{u} \rangle_{\mathcal{V},2},
\end{aligned}$$

where the first inequality follows from Theorem 4.17. Therefore $\mathbf{z}^k$ satisfies $J(\mathbf{z}^k) \leq J(\mathbf{z})$ for any $\mathbf{z}$ satisfying $\mathbf{A}\mathbf{z} = \mathbf{u}$, and the result follows. ■

B Proofs of results from Chapter 5

In this section, we include proofs from Chapter 5 that are either too long to include in the regular text, or do not add significant insight to the discussion.

Proof of Corollary 5.3. When $d = 1$ we denote $\alpha = r \in \mathbb{N}_0$ and note that from Theorem 4.1,

$$c(r, \ell) = \begin{cases} \#\mathbf{S}(r, \ell) & r \text{ even, } \ell = r/2 \\ 2\#\mathbf{S}(r, \ell) & \text{otherwise,} \end{cases}$$

with $\mathbf{S}(r, \ell) = \{s \in \mathbb{N}_0 : s = \ell, s \leq r\} = \{\ell\}$ for $\ell \leq r$ and $\emptyset$ otherwise. We distinguish in cases:

- Case $r = 2k$, $k \in \mathbb{N}_0$,

1. when $0 \leq r \leq p$,

$$\begin{aligned} \text{nnz}(\mathbf{G}_r) &= \sum_{\ell=\lceil r/2 \rceil}^r c(r, \ell) \binom{1+p-\ell}{p-\ell} = \binom{1+p-k}{p-k} + \sum_{\ell=k+1}^{2k} 2 \binom{1+p-\ell}{p-\ell} \\ &= (1+p-k) - k(3k-2p-1) \\ &= 1+p-4k^2+2kp+k^2 \\ &= (p-2k+1)(2k+1)+k^2 \\ &= (p-r+1)(r+1)+k^2. \end{aligned}$$

2. when $p+1 \leq r \leq 2p$, we have $\frac{p+1}{2} \leq k \leq p$, so

$$\begin{aligned} \text{nnz}(\mathbf{G}_r) &= \sum_{\ell=\lceil r/2 \rceil}^r c(r, \ell) \binom{1+p-\ell}{p-\ell} = \binom{1+p-k}{p-k} + \sum_{\ell=k+1}^{2k} 2 \binom{1+p-\ell}{p-\ell} \\ &= (1+p-k) + \sum_{\ell=k+1}^p 2 \binom{1+p-\ell}{p-\ell} \\ &= (1+p-k) + (p-k)(p-k+1) \\ &= (1+p-k)^2. \end{aligned}$$

3. when $r > 2p$, then $k > p$, so

$$\text{nnz}(\mathbf{G}_r) = \sum_{\ell=\lceil r/2 \rceil}^r c(r, \ell) \binom{1+p-\ell}{p-\ell} = \binom{1+p-k}{p-k} + \sum_{\ell=k+1}^{2k} 2 \binom{1+p-\ell}{p-\ell} = 0,$$

since $p-k < 0$ and $l > k \Rightarrow p-\ell < p-k < 0$.

- Case $r = 2k+1$, $k \in \mathbb{N}_0$,

1. when $0 \leq r \leq p$, then $\lceil r/2 \rceil = \lceil (2k+1)/2 \rceil = \lceil k+1/2 \rceil = k+1$, so

$$\begin{aligned}
\text{nnz}(\mathbf{G}_r) &= \sum_{\ell=\lceil r/2 \rceil}^r c(r, \ell) \binom{1+p-\ell}{p-\ell} = 2 \sum_{\ell=k+1}^{2k+1} \binom{1+p-\ell}{p-\ell} \\
&= -(1+k)(3k-2p) \\
&= -4k + 2p - 4k^2 + 2kp + k^2 + k \\
&= -2k(2k+2) + p(2k+2) + k^2 + k \\
&= (p-2k)(2k+2) + k^2 + k \\
&= (p-r+1)(r+1) + k^2 + k.
\end{aligned}$$

2. when $p+1 \leq r \leq 2p$, then $p/2 \leq k \leq p-1/2$, so

$$\begin{aligned}
\text{nnz}(\mathbf{G}_r) &= \sum_{\ell=\lceil r/2 \rceil}^r c(r, \ell) \binom{1+p-\ell}{p-\ell} = 2 \sum_{\ell=k+1}^{2k+1} \binom{1+p-\ell}{p-\ell} = 2 \sum_{\ell=k+1}^p \binom{1+p-\ell}{p-\ell} \\
&= (p-k)(p-k+1).
\end{aligned}$$

3. when $r > 2p$, then $k > p-1/2$, so

$$\text{nnz}(\mathbf{G}_r) = \sum_{\ell=\lceil r/2 \rceil}^r c(r, \ell) \binom{1+p-\ell}{p-\ell} = \sum_{\ell=k+1}^{2k+1} c(r, \ell) \binom{1+p-\ell}{p-\ell} = 0,$$

since $k > p-1/2 \Rightarrow p-\ell \leq p-(k+1) = p-k-1 < p-(p-1/2)-1 = -1/2$.

This completes the proof. ■

Vita

Nick Dexter was born in Largo, Florida to parents Areti Tsangaris and Timothy Dexter. He graduated from Countryside High School in 2005, accepting a research scholarship to study computational mathematics at the Rochester Institute of Technology. In 2009, Nick was awarded a competitive research internship at Oak Ridge National Laboratory where he worked on problems of cell migration in response to biochemical gradients and Bayesian inference for neutron scattering problems. From 2009-2011 Nick worked as an Undergraduate Research Assistant at RIT with Professor George Thurston, modeling the formation of cataracts in the lens of the eye. After graduating with a Bachelor of Science degree in 2011, Nick worked as a Research Associate for the University of Tennessee. Nick was awarded a graduate fellowship in 2012 from The Bredesen Center for Interdisciplinary Research and Graduate Education at the University of Tennessee. Nick earned a Doctor of Philosophy in Mathematics from the University of Tennessee while conducting research in the Computational and Applied Mathematics group at Oak Ridge National Laboratory under the advisement of Dr. Clayton Webster. In 2018, Nick was awarded the Pacific Institute for the Mathematical Sciences Postdoctoral Fellowship to continue his research at Simon Fraser University with Dr. Ben Adcock.