Politecnico di Milano

MOX - Modeling and Scientific Computing
Dipartimento di Matematica

Doctoral Program in
Mathematical Models and Methods for Engineering - XXVIII Cycle

---

COmpRessed SolvING:
Sparse Approximation of PDEs based on Compressed Sensing

Candidate:
Simone Brugiapaglia

Advisor and Tutor:
Prof. Simona Perotto

Co-advisor:
Prof. Stefano Micheletti

The Chair of the Doctoral Program:
Prof. Roberto Lucchetti

Academic Year 2015-16

Abstract

In this thesis, we deal with a new framework for the numerical approximation of partial differential equations which employs main ideas and tools from compressed sensing in a Petrov-Galerkin setting. The goal is to compute an s-sparse approximation with respect to a trial basis of dimension N (with $s \ll N$) by picking $m \ll N$ randomly chosen test functions, and to employ sparse optimization techniques to solve the resulting $m \times N$ underdetermined linear system. This approach has been named COmpRessed SolvING (in short, CORSING).

First, we carry out an extensive numerical assessment of CORSING on advection-diffusion-reaction equations, both in a one- and a two-dimensional setting, showing that the proposed strategy is able to reduce the computational burden associated with a standard Petrov-Galerkin formulation.

Successively, we focus on the theoretical analysis of the method. In particular, we prove recovery error estimates both in expectation and in probability, comparing the error associated with the CORSING solution with the best s-term approximation error. With this aim, we propose a new theoretical framework based on a variant of the classical inf-sup property for sparse vectors, that is named Restricted Inf-Sup Property, and on the concept of local a-coherence, that generalizes the notion of local coherence to bilinear forms in Hilbert spaces. The recovery results and the corresponding hypotheses are then theoretically assessed on one-dimensional advection-diffusion-reaction problems, while in the two-dimensional setting the verification is carried out through numerical tests.

Finally, a preliminary application of CORSING to three-dimensional advection-diffusion-reaction equations and to the two-dimensional Stokes problem is also provided.

Keywords: partial differential equations, compressed sensing, Petrov-Galerkin formulation, inf-sup property, local coherence, estimates in expectation and probability.

Sommario

In questa tesi viene proposto un nuovo metodo per l'approssimazione numerica di equazioni differenziali alle derivate parziali, basato sull'applicazione di tecniche e idee del compressed sensing a discretizzazioni di tipo Petrov-Galerkin. L'obiettivo è quello di calcolare una approssimazione s-sparsa rispetto ad una base trial di dimensione N (con $s \ll N$), selezionando $m \ll N$ funzioni test in maniera randomizzata e, successivamente, risolvere il sistema sottodeterminato ottenuto, di dimensione $m \times N$, tramite tecniche di ottimizzazione sparsa. Questo approccio è stato denominato COmpRessed SolvING (in breve, CORSING).

In primis, viene condotta una vasta indagine numerica del CORSING su equazioni di tipo diffusione-trasporto-reazione monodimensionali e bidimensionali, mostrando come la strategia proposta sia capace di ridurre il costo computazionale associato a discretizzazioni di Petrov-Galerkin standard.

Successivamente, il metodo viene studiato dal punto di vista teorico. In particolare, si dimostrano delle stime di errore in valore atteso e in probabilità, mettendo a confronto l'errore della soluzione CORSING e l'errore di miglior approssimazione s-sparsa. L'analisi teorica è basata su una variante della classica proprietà di inf-sup per vettori sparsi, denominata proprietà di inf-sup ristretta, e sul concetto di a-coerenza locale, che generalizza la nozione di coerenza locale al caso di forme bilineari su spazi di Hilbert. I risultati teorici e le corrispettive ipotesi vengono poi specializzati al caso di equazioni di diffusione-traporto-reazione monodimensionali, mentre nel caso bidimensionale le ipotesi vengono verificate numericamente.

Infine, risultati preliminari mostrano come il CORSING possa essere applicato al caso di equazioni di diffusione-trasporto-reazione tridimensionali e al problema di Stokes bidimensionale.

Parole chiave: equazioni differenziali alle derivate parziali, compressed sensing, formulazione di Petrov-Galerkin, proprietà di inf-sup, coerenza locale, stime in valore atteso e in probabilità.

Contents

Introduction

We present a novel technique for approximating Partial Differential Equations (PDEs), taking advantage of concepts from signal processing, nonlinear approximation, sparse representations and, above all, from *Compressed Sensing* (CS). The principal motivation is to reduce the computational cost associated with the *Petrov-Galerkin* (PG) discretization of a PDE.

The COmpRessed SolvING approach

The CS technique was developed by D.L. Donoho [Don06] and E.J. Candès, J.K. Romberg and T. Tao [CRT06], and allows one to sample a signal using far fewer measurements than those required by the Nyquist-Shannon sampling theorem, where the sampling rate must be at least twice the maximum frequency of the signal (the so-called *Nyquist rate*).

This discovery potentially has a substantial impact on real world applications: for example, in *Magnetic Resonance Imaging*, this leads to significant scan time reductions, with benefits for patients and health care economics [LDSP08]. Other remarkable applications of CS are *Radar Imaging* [HS09] or the *Single-Pixel Camera* [DDT+08].

The main hypothesis underlying CS is *sparsity* or, more generally, *compressibility*. Many natural signals fit into this framework, having a concise representation when expressed in the proper basis. In particular, expanding the signal with respect to a basis of N vectors, it is possible to approximate the best s-term approximation to the signal, with $s \ll N$, by means of m random linear nonadaptive measurements, with $s < m \ll N$, modeled as inner products of the signal against suitable test vectors, and employing computationally feasible sparse recovery techniques.

In this thesis, the main idea we use to link the field of signal processing and the one of numerical methods for PDEs is really simple:

why not identifying the solution of a PDE with a signal?

In particular, we focus on discretization methods based on the PG formulation, popularized during the 1970's by A.K. Aziz and I. Babuška [AB72]. This is

Petrov-Galerkin method: solution of a PDE **tests** (bilinear form)	$\longleftrightarrow$	*Sampling:* signal **measurements** (inner product)

Figure 1: Analogy between the PG method and the CS sampling procedure.

a general framework, including *finite elements* (FE), *finite volumes*, *spectral approximations*, where the PDE in a variational form is evaluated against several test functions. This process is analogous to the measurement of a signal in the sampling phase of the CS, with the only difference that the tests are performed using a bilinear form in place of an inner product. We can think of the bilinear form as our *measuring device*, employed to virtually acquire the unknown solution of the PDE.

Inspired by this parallelism (Figure 1) we coin the name COmpRessed Solving, in short CORSING, to refer to the methodology proposed in this thesis.[1]

In practice, CORSING aims at reducing the computational cost associated with the standard PG approximation, by reducing the size, N, of the associated square linear system, selecting just some rows, $m \ll N$, of the stiffness matrix and load vector. This selection amounts to picking a subset of the whole test functions. We propose either a deterministic or a random selection strategy, i.e., we pick the first m tests with respect to a predefined ordering, or we extract them after assigning a probability distribution on the test space. Either way, we are led to an underdetermined system, which we solve by means of sparse recovery techniques, such as ℓ^1-minimization [DL92] or the greedy algorithm *Orthogonal Matching Pursuit* (OMP) [MZ93, PRK93], thus computing an s-sparse approximation to the solution, with $s \ll N$.

An important issue related to CORSING is choosing the trial and test functions. In CS, the main idea consists of picking a trial space with good sparsity properties in the time domain, and a test space sparse in the frequency domain, or vice versa. We will follow the same heuristics, adopting, on the one hand, the hierarchical multiscale basis of hat functions [Yse86, Dah97] and, on the other hand, the basis of sine functions.

Comparison with existing techniques

In order to emphasize the potentialities of the proposed approach, we compare the CORSING approach with other techniques, consolidated in the literature.

[1] We have chosen the word CORSING also due to its assonance with the verb coarsen, in reference to the roughening of the standard PG method.

Adaptive approximation of PDEs The CORSING method aims at computing the best s-term approximation to the solution to a PDE. Therefore, it can be classified among nonlinear approximation methods for PDEs [DeV98, Tem03]. Although the framework for CORSING is very general and can accommodate many different choices of trial and test spaces, when considering hierarchical piecewise polynomials over an initial coarse triangulation as trial basis functions, a possible competitor approach is the Adaptive Finite Element Method (AFEM) (see, e.g., [NSV09] and the references therein). AFEM and CORSING are, however, thoroughly different: in AFEM, the solution is iteratively computed according to the loop

$$\text{SOLVE} \to \text{ESTIMATE} \to \text{MARK} \to \text{REFINE},$$

and exploiting suitable *a posteriori* error estimators. On the contrary, with CORSING, we employ a reduced PG discretization, using a fixed trial space of dimension N (which corresponds ideally to a very fine uniform refinement, expressed in a hierarchical basis) and performing a fixed number of random measurements in the test space. In particular:

(1) the trial space is not iteratively enlarged, but fixed initially;

(2) the measurements in the test space are performed non-adaptively;

(3) no *a posteriori* error estimator/indicator is required.

The CORSING procedure then recovers an s-sparse solution, which can be compared with the AFEM solution on the same ground. We consider (1) as a possible drawback of CORSING, whereas (2) and (3) are upsides. In principle, (1) requires a higher computational cost in the recovery phase, whereas (2) allows for full parallelization and (3) significantly reduces the implementation complexity.

In particular, an earlier attempt to apply CS to the adaptive approximation of a PDE can be found in [JMPY10]. The authors focus on a Galerkin formulation of the Poisson problem, where the trial and test spaces coincide with piecewise linear finite elements. The proposed technique is fully deterministic and relies on the successive refinement of the solution on different hierarchical levels and on a suitable error estimator. Only the ℓ^1-minimization is applied and it appears that m is very close to N.

Infinite-dimensional CS From a different perspective, CORSING can be considered as a variant of the infinite-dimensional CS, where CS is applied to infinite-dimensional Hilbert spaces [AH15, AHPR13]. This is achieved by subsampling a given isometry of the Hilbert space, usually associated with an inner product and a change of basis (e.g., from a wavelet basis to the Fourier basis). The main

idea behind CORSING is different, since it deals with the bilinear form arising from the weak formulation, that can be even nonsymmetric. Nevertheless, we think that the theory developed in [AH15, AHPR13] could play a significant role for a deeper understanding of the CORSING technique and this will be a subject of future investigation.

Reduction strategies based on the SVD The CORSING approach provides a compression of a standard PG discretization, thus it is natural to compare it with compression strategies based on the SVD factorization [GL13]. A remarkable application of the SVD for the model order reduction of PDEs is the Proper Orthogonal Decomposition (see, e.g., [KV02]). This issue is discussed in Section 2.3.1.

ℓ^1-minimization techniques for PDEs This work can also be related to numerical methods for PDEs based on ℓ^1-minimization. To our knowledge, the earliest contributions are the pioneering studies by J.E. Lavery on the inviscid Burgers' equation and on steady scalar conservation laws [Lav88, Lav89]. More recently, similar techniques have been analyzed in [Gue04, GP09], where transport and Hamilton-Jacobi equations are considered. However, the CS principles are applied in none of these works.

CS and high-dimensional stochastic parametric PDEs Finally, it is worth mentioning the recent application of CS to the numerical approximation of high-dimensional stochastic parametric PDEs, of particular interest in *Uncertainty Quantification* [DO11, YK13, PHD14, SSN+14, RS14, BBRS15]. Even though we deal with the issue of reducing the computational cost associated with the numerical approximation of one single deterministic PDE, we think that a combination of CORSING with the aforementioned techniques could be of interest for a future investigation.

Outline of the thesis

The thesis is divided in four chapters.

In Chapter 1, we present the CS technique, introducing its main underlying concepts, namely, sparsity, sensing and recovery. Then, we review some theoretical results about CS. First, we discuss the *Restricted Isometry Property* (RIP), the concept of *coherence* and provide recovery results for the greedy algorithm *Orthogonal Matching Pursuit* (OMP). Moreover, we introduce the theory of *Bounded Orthonormal Systems* (BOS) and some recent sampling strategies based on the notion of *local coherence*, discussing a particular example (Haar *vs*

Fourier), of inspiration for the development of CORSING. Finally, we provide an original generalization of the RIP for generic matrices, that will be applied in the theoretical study of CORSING.

In Chapter 2, after presenting the PG method and the main elements of the Babuška-Nečas theory, we introduce the CORSING approach in its deterministic (D-CORSING) and randomized (R-CORSING) version, following the same heuristic spirit that led us through its discovery. Then, we check, with an extensive numerical assessment, the CORSING accuracy, computational burden, and robustness, on one-dimensional and two-dimensional advection-diffusion-reaction (ADR) problems. In particular, we provide comparisons with the best s-term approximation error, with the full-PG method ($m = N$), with FE and with an SVD-based approach.

The goal of Chapter 3 is to set up a theoretical analysis of R-CORSING, formalizing the empirical recipes given in Chapter 2 and introducing some technical assumptions based on the concept of *local a-coherence*, generalization of the local coherence to bilinear forms in Hilbert spaces. Then, we introduce the *Restricted Inf-Sup Property* (RISP), a combination of the classical inf-sup condition [BF91] and the RIP of CS and provide an analysis of R-CORSING for generic weak problems in Hilbert spaces. Then, we discuss the application of the CORSING theory to the case of one- and two-dimensional ADR problems. Moreover, we provide further numerical assessments in order to corroborate the theoretical results.

Finally, with a view to more practical applications, we validate the CORSING method on the two-dimensional Stokes problem and, through a tensorization strategy, on the three-dimensional ADR problem in Chapter 4.

Chapter 1

Compressed sensing

Compressed Sensing (CS) is a novel research area in the signal processing field, which provides an effective way to acquire a signal by means of a small number of measurements, less than required by the Nyquist-Shannon sampling theorem [Nyq28, Sha49]. CS was proposed in 2006 in the pioneering works by D.L. Donoho [Don06] and by E.J. Candés, J.K. Romberg, and T. Tao [CRT06].

In this chapter, we outline the concepts and results about CS useful for fixing the COmpRessed SolvING approach.

For the sake of generality, all the results are presented in $\mathbb{C}^N$, but they hold in $\mathbb{R}^N$ as well.

Outline of the chapter This chapter is organized in two sections. In Section 1.1 we review three basic concepts underpinning the CS method: sparsity, sensing and recovery. Afterwards, in Section 1.2 we present some elements of the CS theory, such as the Restricted Isometry Property, the concepts of coherence and local coherence, the theory of Bounded Orthonormal Systems and some recovery results for the Orthogonal Matching Pursuit algorithm.

1.1 Three main concepts

The aim of CS is to acquire an unknown signal, assumed to have a sparse representation, using the minimal number of linear nonadaptive measurements and then recovering it by means of efficient optimization procedures.

In order to present the CS method, we need to familiarize with three main concepts. (1) Sparsity: the only hypothesis needed on the unknown signal to be measured, (2) Sensing: how to acquire a signal in a compressed way, using the minimum number of measurements, (3) Recovery: the reconstruction of the signal from its measurements. They will be presented separately in the next sections.

1.1.1 Sparsity: what does it mean, exactly?

The only hypothesis underlying the CS approach is *sparsity*. But what does it mean for a vector to be sparse, exactly? Informally, a vector is sparse if it has only few non-zero entries with respect to the total number of entries. However, due to the importance that this notion has in the CS framework, we need to define sparsity rigorously.

s-sparsity The first way to measure sparsity relies on the so-called *ℓ^0-norm*, that simply counts the non-zero entries of a vector. Given $\mathbf{u} \in \mathbb{C}^N$, we define its *support* as the set of indices corresponding of its non-zero entries, namely

$$\operatorname{supp}(\mathbf{u}) := \{j \in [N] : u_j \neq 0\},$$

with $[N] := \{1, \dots, N\}$, and its ℓ^0-norm as the cardinality of the support, i.e.,

$$\|\mathbf{u}\|_0 := |\operatorname{supp}(\mathbf{u})|. \tag{1.1}$$

Actually, the ℓ^0-norm is not a norm, since $\|\lambda \mathbf{u}\|_0 = \|\mathbf{u}\|_0$, $\forall \lambda \neq 0$; however, we will adopt this terminology due to its ubiquitous presence in the literature.

Using the ℓ^0-norm, we define the concept of s-sparsity.

Definition 1.1. A vector $\mathbf{u} \in \mathbb{C}^N$ is *s-sparse* if $\|\mathbf{u}\|_0 \leq s$.

Moreover, the set containing all the s-sparse vectors of $\mathbb{C}^N$ is denoted

$$\Sigma_s^N := \{\mathbf{u} \in \mathbb{C}^N : \|\mathbf{u}\|_0 \leq s\}. \tag{1.2}$$

This definition of sparsity is employed in [CRT06], where the signal is assumed to be "the superposition of s spikes". We underline that Σ_s^N is a finite union of linear subspaces; a visualization of Σ_2^3 in the real-valued case is given in Figure 1.1.

In CS, one usually deals with situations where the sparsity is much smallest than the dimension of the space, namely,

$$s \ll N.$$

Compressibility A less restrictive and often adopted hypothesis is to assume $\mathbf{u}$ to be *compressible*, instead of sparse. This concept relies on the definition of *best s-term approximation error*.

Definition 1.2. Given $p > 0$, $s \in \mathbb{N}$ and $\mathbf{u} \in \mathbb{C}^N$, the *best s-term approximation error of $\mathbf{u}$ with respect to the ℓ^p-norm* is the quantity

$$\sigma_s(\mathbf{u})_p := \inf_{\mathbf{z} \in \Sigma_s^N} \|\mathbf{u} - \mathbf{z}\|_p.$$

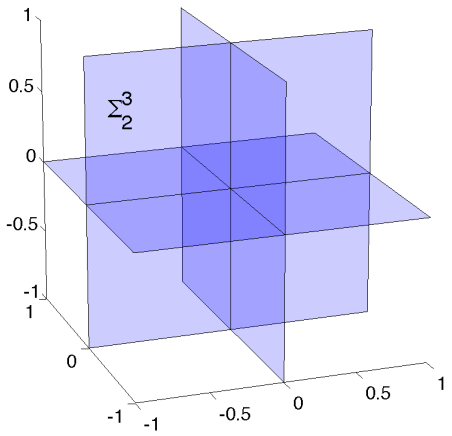


Figure 1.1: The set Σ_2^3 of 2-sparse vectors of $\mathbb{R}^3$.

Notice that the infimum is (not necessarily uniquely) realized by keeping the s entries of $\mathbf{u}$ having the largest magnitude and placing zeros elsewhere.

A vector $\mathbf{u} \in \mathbb{C}^N$ is called *compressible* when its best s-term approximation error decays quickly in s, i.e., if there exist $p, r > 0$ such that

$$\sigma_s(\mathbf{u})_p \lesssim s^{-r}.$$

In order to understand what vectors are good candidates to be compressible, we consider the *Stechkin inequality* (see [FR13, Theorem 2.5])

$$\sigma_s(\mathbf{u})_q \lesssim \frac{\|\mathbf{u}\|_p}{s^{1/p-1/q}}, \quad \forall \mathbf{u} \in \mathbb{C}^N, \forall s \in \mathbb{N}, \forall q > p > 0.$$

As a consequence, all the vectors belonging to a given ℓ^p-ball, with p small enough, have a good level of compressibility. In particular, in his pioneering work [Don06], D.L. Donoho defines $\mathbf{u} \in \mathbb{C}^N$ to be sparse whenever $\|\mathbf{u}\|_p \leq R$ for some $0 < p < 2$ and $R > 0$.

Sparse representations Assuming a vector $\mathbf{u} \in \mathbb{C}^N$ to be sparse, or compressible, is, in general, too restrictive. What is more often done is to assume a given signal $\mathbf{s} \in \mathbb{C}^N$ to have a *sparse representation* with respect to a given *sparsity basis* of $\mathbb{C}^N$. Namely, we assume the existence of an orthonormal basis

$$\mathbf{\Psi} = [\boldsymbol{\psi}_1 | \cdots | \boldsymbol{\psi}_N] \in \mathbb{C}^{N \times N}$$

such that

$$\mathbf{s} = \mathbf{\Psi}\mathbf{u} = \sum_{j \in [N]} u_j \boldsymbol{\psi}_j, \quad \text{with } \mathbf{u} \text{ sparse (or compressible).}$$

Remark 1.1.1. In this chapter, we make no distinction between orthonormal bases of $\mathbb{C}^N$ and unitary matrices, implicitly identifying a unitary matrix $\mathbf{U} = [\mathbf{u}_1|\cdots|\mathbf{u}_N] \in \mathbb{C}^{N\times N}$ with the orthonormal basis formed by its columns $\{\mathbf{u}_1,\dots,\mathbf{u}_N\}$. □

The choice of the sparsity basis actually depends on the particular signal that one wants to measure. If the signal $\mathbf{s}$ itself is sparse, then $\mathbf{\Psi} = \mathbf{I}$, i.e., the identity matrix or, equivalently, the canonical basis, is the most natural choice. Otherwise, one should employ different bases.

Perhaps the oldest example of sparsity basis is the *discrete Fourier basis*, whose origin dates back to the 19^{th} century [Fou22]. Assuming N even, it is defined as

$$\mathbf{F} = \left[\mathbf{f}_{-\frac{N}{2}+1}\middle|\cdots\middle|\mathbf{f}_{\frac{N}{2}}\right] \in \mathbb{C}^{N\times N}, \tag{1.3}$$

whose columns are

$$[\mathbf{f}_r]_j = \frac{1}{\sqrt{N}}\exp\left(\frac{\mathrm{i}2\pi jr}{N}\right), \quad \forall j \in [N],$$

with $r \in \mathbb{Z}$ and $-N/2 < r \leq N/2$, and $\mathrm{i} = \sqrt{-1}$. Using $\mathbf{F}$, we are able to decompose a signal into the sum of harmonics, corresponding to pure frequencies. A remarkable and well-known property of the discrete Fourier system is that the matrix-vector multiplication can be computed with complexity $\mathcal{O}(N\log N)$, thanks to the *Fast Fourier Transform* [CT65].

The Fourier basis can be generalized through tensorization to an arbitrarily high dimension and admits many variants, such as the discrete sine and cosine bases. An outstanding modern application of this basis is the JPEG standard for image compression, exploiting the sparsifying capability of the two-dimensional discrete cosine basis [PM93].

There are many other prominent examples of sparsity bases. For example, it is worth recallig the *Haar system* [Haa10] and the *wavelets* [Dau92], with their recent variants such as the *noiselets* [CGM01], the *curvelets* [SCD02] and the *shearlets* [GKL06].

In Section 1.2.6 we will define the discrete Haar basis, and study its relation with the Fourier basis.

1.1.2 Sensing: the "big soup"

Let $\mathbf{s} \in \mathbb{C}^N$ be a signal having a sparse representation $\mathbf{s} = \mathbf{\Psi}\mathbf{u}$.

Given a set of test vectors $\{\boldsymbol{\varphi}_1,\dots,\boldsymbol{\varphi}_m\} \subseteq \mathbb{C}^N$, with $m \ll N$, the measurement process is performed in a linear and nonadaptive way by computing the inner products

$$\langle \boldsymbol{\varphi}_i, \mathbf{s}\rangle = f_i, \quad \text{for } i = 1,\dots,m,$$

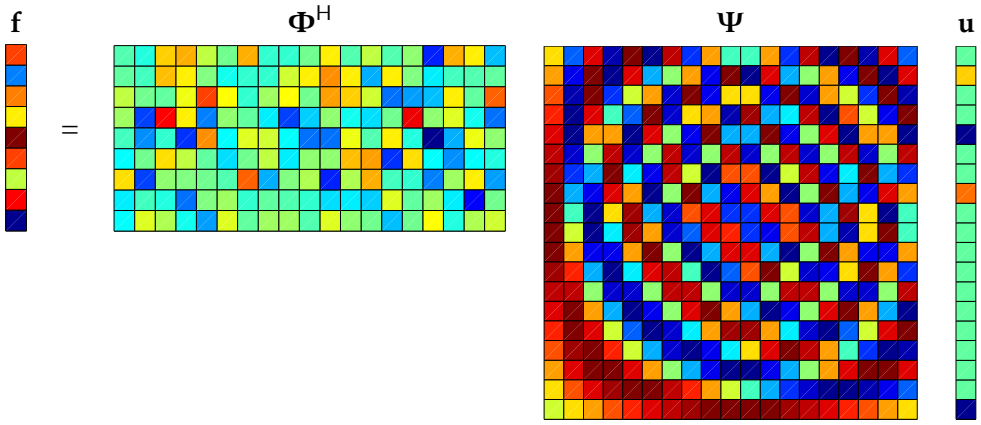


Figure 1.2: Schematization of the sensing process leading to $\mathbf{f} = \mathbf{\Phi}^{\mathsf{H}}\mathbf{\Psi}\mathbf{u}$.

with $\langle \cdot, \cdot \rangle$ the standard Hermitian inner product. If we consider the matrix that collects the vectors $\boldsymbol{\varphi}_i$ as columns, $\mathbf{\Phi} = (\boldsymbol{\varphi}_i) \in \mathbb{C}^{N \times m}$, the whole measurement process can be recast in the linear system

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \tag{1.4}$$

where $\mathbf{A} = \mathbf{\Phi}^{\mathsf{H}}\mathbf{\Psi} \in \mathbb{C}^{m \times N}$ is the *measurement matrix* and $\mathbf{f} \in \mathbb{C}^m$ collects the measurements f_i. The measurement process is often called *sensing* or *encoding* and is represented in Figure 1.2. We notice that the case of a signal $\mathbf{s}$ that is trivially sparse, i.e., sparse in the canonical basis, is a simple subcase of this general framework, with $\mathbf{\Psi} = \mathbf{I}$ and $\mathbf{A} = \mathbf{\Phi}^{\mathsf{H}}$.

At this stage, a fundamental question arises: given a sparsity basis $\mathbf{\Psi}$, how should we choose $\mathbf{\Phi}$ in order to minimize the number of measurements? We will investigate this issue in the next developments, but the following words by D.L. Donoho [Don06] give us a first insight.

> "Surely then, one imagines, the sampling kernels ξ_i underlying the optimal information operator must be simply measuring individual transform coefficients? Actually, no: the information operator is measuring very complex holographic functionals which in some sense mix together all the coefficients in a *big soup*."[1]

Moreover, before answering the question, we need to understand how to recover $\mathbf{u}$ after the sensing process.

[1] Traslating the passage to our notation, Donoho's "sampling kernels ξ_i" are the measurement vectors $\boldsymbol{\varphi}_i$, the "transform coefficients" are the components of $\mathbf{u}$, whereas the "information operator" is the map $\mathbf{u} \mapsto \mathbf{A}^{\mathsf{H}}\mathbf{u}$, corresponding to the sensing process.

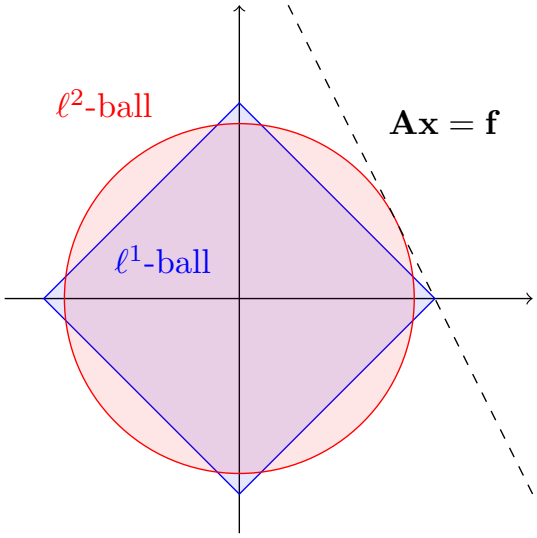


Figure 1.3: The shape of the ℓ^1-ball promotes sparsity.

1.1.3 Recovery: looking for a needle in a haystack

The linear system (1.4) is highly underdetermined ($m \ll N$) and, in general, it may have no solution or infinite solutions. To overcome this limit, a *recovery* (or *decoding*) algorithm is employed. The basic idea is to find the sparsest solution to (1.4), i.e., to solve the nonlinear optimization problem

$$(\mathrm{P}_0) \quad \min_{\mathbf{u}\in\mathbb{C}^N} \|\mathbf{u}\|_0 \quad \text{s.t. } \mathbf{Au} = \mathbf{f}. \tag{1.5}$$

Even though problem (P_0) has been proved to be NP-hard [Nat95], several algorithms have been devised in order to approximate it. We will focus, in particular, on the greedy algorithm Orthogonal Matching Pursuit (OMP), presented in Section 1.2.3.

An alternative and very popular idea, is to relax the ℓ^0-norm in (P_0) with the ℓ^1-norm, yielding the convex optimization problem

$$(\mathrm{P}_1) \quad \min_{\mathbf{u}\in\mathbb{C}^N} \|\mathbf{u}\|_1 \quad \text{s.t. } \mathbf{Au} = \mathbf{f}. \tag{1.6}$$

(P_1) was originally introduced in signal processing in [Log65] and then studied in [DL92] (before CS was born). It can be solved using classical tools of convex optimization. In fact, it turns out to be a linear programming problem when $\mathbf{A}$ is a real matrix, while it can be reduced to a second-order conic programming problem when the entries of $\mathbf{A}$ are complex. A very fast and stable MATLAB® package for the ℓ^1-minimization is SPGL1 (see [vdBF08, vdBF07]).

From an intuitive viewpoint, the ℓ^1-minimization is able to recover the sparsest vector because the intersection of the smaller ℓ^1-ball with the linear subspace $\{\mathbf{u} : \mathbf{Au} = \mathbf{f}\}$ is unique and has minimal sparsity, except for pathological

situations where the kernel of $\mathbf{A}$ is parallel to an edge of the ℓ^1-ball. This situation does not occur employing other convex ℓ^p-norms, such as the ℓ^2-norm (see Figure 1.3).

More practically, due to the numerical impossibility to realize the exact linear constraint in (P_0) and (P_1), we will consider the minimization problems

$$(P_q^\varepsilon) \quad \min_{\mathbf{u}\in\mathbb{C}^N} \|\mathbf{u}\|_q \quad \text{s.t. } \|\mathbf{A}\mathbf{u}-\mathbf{f}\|_2 \le \varepsilon, \quad q = 0, 1, \tag{1.7}$$

where $\varepsilon > 0$ is a given tolerance on the ℓ^2-norm of the residual.

Remark 1.1.2 (Least-squares minimization). A classical strategy to solve the underdetermined linear system (1.4) is least-squares minimization, based on the rank-deficient QR decomposition of $\mathbf{A}$ (see [GL13, Section 5.5]). This method is implemented in the MATLAB® \ (backslash) command. Even though, in some circumstances, this could be a valuable option, we will focus only on ℓ^0- and ℓ^1-minimization, due to a better capability to recover sparse vectors. □

1.2 Theoretical tastes

The theory of CS experienced a huge growth in the last decade. For this reason, summarizing it in a few pages is an impossible task. Therefore, we will just review the theoretical aspects and the main concepts needed for the further developments of this thesis, thus providing to the reader some "theoretical tastes".

For a thorough review about CS, we highly recommend the textbook [FR13] that collects an exhaustive set of results and pointers to the literature; moreover it has the great quality of being self-contained. We also found very helpful the book [Ela10] and the review papers [CW08, FR11, JV11, Kut12]. Finally, the online repository by the Rice University [Ric] gathers a vast catalogue of references.

1.2.1 The Restricted Isometry Property

A deeper understanding of problems (P_0) and (P_1) requires more conceptual results. These optimization problems raise some fundamental questions: when do (P_0) and (P_1) admit a unique solution? When the solutions of (P_0) and (P_1) do coincide? For a certain vector $\mathbf{u}$, with k non-zero entries, what is the minimum number of measurements needed to successfully recover $\mathbf{u}$?

A fundamental tool to answer these questions is the *Restricted Isometry Property* (RIP), first introduced in [CRT06].

Definition 1.3. A matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ satisfies the RIP of order $s < m$ and constant $\delta \in [0, 1)$ if

$$(1-\delta)\|\mathbf{u}\|_2^2 \le \|\mathbf{A}\mathbf{u}\|_2^2 \le (1+\delta)\|\mathbf{u}\|_2^2, \quad \forall \mathbf{u} \in \Sigma_s^N,$$

with Σ_s^N defined as in (1.2). If $\mathbf{A}$ satisfies this definition, we say that $\mathbf{A} \in \mathrm{RIP}(s, \delta)$. Essentially, we are requiring that $\mathbf{A}$ behaves nearly like an isometry on all the s-sparse vectors.

We briefly browse some well-known "RIP-based" results, referring the reader to [FR13, Chapter 6] for an exhaustive review. We start with a fundamental recovery result about (P_0).

Proposition 1.4. *Let $\mathbf{A} \in \mathbb{C}^{m\times N}$, $\mathbf{u} \in \mathbb{C}^N$, with $\|\mathbf{u}\|_0 \leq s$, and $\mathbf{Au} = \mathbf{f}$. If there exists $\delta \in [0, 1)$ such that $\mathbf{A} \in RIP(2s, \delta)$, then (P_0) recovers $\mathbf{u}$ exactly.*

Proof. Let $\mathbf{u}^*$ be a solution to (P_0). Then, thanks to its optimality, we have $\|\mathbf{u}^*\|_0 \leq \|\mathbf{u}\|_0$. This condition, together with the hypothesis $\|\mathbf{u}\|_0 \leq s$, implies $\|\mathbf{u} - \mathbf{u}^*\|_0 \leq 2s$. Finally, using that $\mathbf{A} \in \mathrm{RIP}(2s, \delta)$, we get

$$(1 - \delta)\|\mathbf{u} - \mathbf{u}^*\|_2^2 \leq \|\mathbf{A}(\mathbf{u} - \mathbf{u}^*)\|_2^2 = \|\mathbf{Au} - \mathbf{Au}^*\|_2^2 = \|\mathbf{f} - \mathbf{f}\|_2^2 = 0,$$

hence $\mathbf{u} = \mathbf{u}^*$. □

In practice, this proposition states that, for any fixed s-sparse signal $\mathbf{u}$, with associated measurement vector $\mathbf{f}$, problem (P_0) yields a unique solution, coinciding with $\mathbf{u}$.

An original and slight improvement over Proposition 1.4 is given in terms of a suitable *inf-sup* property [BMP15] by the following

Proposition 1.5. *Let $\mathbf{A} \in \mathbb{C}^{m\times N}$, $\mathbf{u} \in \mathbb{C}^N$, with $\|\mathbf{u}\|_0 \leq s$, and $\mathbf{Au} = \mathbf{f}$. If $\mathbf{A}$ satisfies*

$$\inf_{\mathbf{x}\in\Sigma_{2s}^N;\ \mathbf{x}\neq\mathbf{0}} \sup_{\mathbf{z}\in\mathbb{C}^m;\ \mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}^{\mathsf{H}}\mathbf{Ax}}{\|\mathbf{z}\|_2\|\mathbf{x}\|_2} = \alpha > 0,$$

then (P_0) recovers $\mathbf{u}$ exactly.

Proof. We mimic the proof of Proposition 1.4. The only variant is that, for each solution $\mathbf{u}^*$ of (P_0), there exists a $\mathbf{z} \neq \mathbf{0}$ such that

$$\alpha \|\mathbf{u} - \mathbf{u}^*\|_2 \leq \frac{\mathbf{z}^{\mathsf{H}}\mathbf{A}(\mathbf{u} - \mathbf{u}^*)}{\|\mathbf{z}\|_2} = 0,$$

hence $\mathbf{u} = \mathbf{u}^*$. □

Remark 1.2.1. The inf-sup condition in Proposition 1.5 is equivalent to the property

$$\|\mathbf{Ax}\|_* \geq \alpha\|\mathbf{x}\|_2 \qquad \forall \mathbf{x} \in \Sigma_{2s}^N,$$

where

$$\|\mathbf{y}\|_* = \sup_{\mathbf{z}\in\mathbb{C}^m;\ \mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}^{\mathsf{H}}\mathbf{y}}{\|\mathbf{z}\|_2}$$

is the dual norm of $\mathbf{y} \in \mathbb{C}^m$. The proof of this statement is a consequence of [BBF13, Proposition 3.4.4]. □

Remark 1.2.2. The inf-sup condition in Proposition 1.5 can be replaced by the hypothesis

$$\mathbf{Aw} = \mathbf{0} \Longrightarrow \mathbf{w} = \mathbf{0}, \quad \forall \mathbf{w} \in \Sigma_{2s}^N,$$

or, equivalently, by requiring that any set of $2s$ columns of $\mathbf{A}$ be linearly independent. □

The following recovery result holds for the (P_1) problem, and the proof is provided in [FR11, Theorem 3.4].

Proposition 1.6. *Let* $\mathbf{A} \in \mathbb{C}^{m\times N}$, $\mathbf{u} \in \mathbb{C}^N$ *and* $\mathbf{Au} = \mathbf{f}$. *If* $\mathbf{A} \in RIP(3s, \delta)$, *with* $\delta \in [0, \frac{1}{3})$, *and* $\mathbf{u}^*$ *is a solution to* (P_1), *then there exists a constant* $C = C(\delta) > 0$ *such that the* ℓ_2*-norm error estimate holds*

$$\|\mathbf{u} - \mathbf{u}^*\|_2 \leq C \frac{\sigma_s(\mathbf{u})_1}{\sqrt{s}},$$

where $\sigma_s(\mathbf{u})_1$ *is the best s-term approximation error of* $\mathbf{u}$ *with respect to the* ℓ^1*-norm.*

An immediate consequence of Proposition 1.6 is the following exact recovery result.

Corollary 1.7. *Let* $\mathbf{A} \in \mathbb{C}^{m\times N}$, $\mathbf{u} \in \mathbb{C}^N$, *with* $\|\mathbf{u}\|_0 \leq s$, *and* $\mathbf{Au} = \mathbf{f}$. *If there exists* $\delta \in [0, \frac{1}{3})$ *such that* $\mathbf{A} \in RIP(3s, \delta)$, *then* (P_1) *recovers* $\mathbf{u}$ *exactly.*

Despite its simple definition, the RIP condition is not easy to verify, because of its intrinsic combinatorial nature. In order to give sufficient conditions for the RIP to be fulfilled, we introduce a concept of fundamental importance: the *coherence*. Based on this tool, we provide sufficient conditions on the matrices $\mathbf{\Psi}$ and $\mathbf{\Phi}$ that guarantee $\mathbf{A}$ to fulfill the RIP, with high probability.

1.2.2 The importance of being incoherent

A useful tool to understand how to build matrices $\mathbf{A}$ suitable for CS is the coherence.

Definition 1.8. Given $\mathbf{A} = [\mathbf{a}_1|\cdots|\mathbf{a}_N] \in \mathbb{C}^{m\times N}$ with ℓ^2-normalized columns, we define its *coherence* as

$$\mu(\mathbf{A}) := \max_{i\neq j\in[N]} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|.$$

The coherence measures to what extent the columns of $\mathbf{A}$ are far from being an orthonormal basis. Indeed, the two extreme cases are: (1) when $m = N$ and the columns of $\mathbf{A}$ form an orthonormal basis of $\mathbb{C}^N$, then $\mu(\mathbf{A}) = 0$; (2) if $\mathbf{A}$ has two identical columns, then $\mu(\mathbf{A}) = 1$.

Thanks to the Cauchy-Schwarz inequality, $\mu(\mathbf{A}) \leq 1$. Moreover, a sharp bound for the coherence from below is the following

$$\mu(\mathbf{A}) \geq \sqrt{\frac{N-m}{m(N-1)}}, \quad \forall \mathbf{A} \in \mathbb{C}^{m\times N}. \tag{1.8}$$

This bound was proved for the first time in [Wel74] and it is known as *Welch bound*. There are special matrices guaranteeing the equality in (1.8), called *equiangular tight frames* (see [FR13, Theorem 5.7]).

We can extend the definition of coherence to sets of two bases, introducing the concept of *mutual coherence*.

Definition 1.9. Given $\mathbf{U} = [\mathbf{u}_1|\cdots|\mathbf{u}_N], \mathbf{V} = [\mathbf{v}_1|\cdots|\mathbf{v}_N] \in \mathbb{C}^{N\times N}$ unitary matrices, the *mutual coherence between* $\mathbf{U}$ *and* $\mathbf{V}$ is defined as

$$\mu(\mathbf{U},\mathbf{V}) := \max_{q,j\in[N]} |\langle \mathbf{u}_j, \mathbf{v}_q\rangle|,$$

i.e., the coherence $\mu([\mathbf{U}|\mathbf{V}])$ of the matrix built concatenating $\mathbf{U}$ and $\mathbf{V}$ horizontally.

An immediate consequence of (1.8) is that

$$\mu(\mathbf{U},\mathbf{V}) \geq \sqrt{\frac{1}{2N-1}}.$$

In particular, when this minimum is reached, up to a constant factor, we say that $\mathbf{U}$ and $\mathbf{V}$ are *mutually incoherent*.

An uncertainty principle There is an important discrete uncertainty principle regarding two mutually incoherent bases, proved in [EB02], generalizing the uncertainty principle proved in [DS89] regarding the Fourier basis and the canonical basis.

Theorem 1.10. *Let* $\mathbf{U},\mathbf{V} \in \mathbb{C}^{N\times N}$ *two unitary matrices and* $K > 0$ *a constant such that*

$$\mu(\mathbf{U},\mathbf{V}) \leq \frac{K}{\sqrt{N}}.$$

Let $\mathbf{x},\mathbf{y} \in \mathbb{C}^N$ *be such that* $\mathbf{U}\mathbf{x} = \mathbf{V}\mathbf{y}$. *Then,*

$$\|\mathbf{x}\|_0 + \|\mathbf{y}\|_0 \geq \frac{2\sqrt{N}}{K}.$$

In other words, if a vector is sparse with respect to $\mathbf{U}$, it is forced to be full with respect to $\mathbf{V}$, and vice versa. This provides the main intuition underlying CS: if a signal has a sparse representation with respect to $\mathbf{U}$, then its representation with respect to $\mathbf{V}$ is almost full. Therefore, if one performs scalar products against all the columns of $\mathbf{V}$, there is too much redundancy of information. Thus, a randomized selection of just $m \ll N$ columns of $\mathbf{V}$ as test vectors suffices to store the required amount of information, with high probability.

1.2.3 Orthogonal Matching Pursuit: "greed is good"

The Orthogonal Matching Pursuit (OMP) algorithm has a long genealogy. We can trace its origin back to the late 60s, in the context of statistical regression [HL67, LH70], whereas its first ancestor *projection pursuit regression* appeared a decade later [FS81]. It was published in the form presented here in [CBL89]. Nevertheless, OMP started to attract more and more interest after its application to signal processing, [MZ93, PRK93]. For a wider historical overview see [Tem03].

OMP is presented in Algorithm 1.1, where the matrix $\mathbf{A}$ is assumed to have ℓ^2-normalized columns.

Algorithm 1.1 Orthogonal Matching Pursuit (OMP)

Input:
Matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$, with ℓ^2-normalized columns
Vector $\mathbf{f} \in \mathbb{C}^m$
Tolerance on the residual $\varepsilon > 0$ (or else, sparsity $s \in [N]$)

Output:
Sparse solution $\mathbf{u}$ to $(\mathrm{P}_0^\varepsilon)$ (or else, (P_0^s))

Procedure:
$\mathcal{S} \leftarrow \emptyset$ ▷ Initialization
$\mathbf{u} \leftarrow \mathbf{0}$
while $\|\mathbf{Au} - \mathbf{f}\|_2 > \varepsilon$ (or else, $\|\mathbf{u}\|_0 < s$) **do**
  $\bar{j} \leftarrow \arg\max_{j \in [N]} |[\mathbf{A}^{\mathsf{H}}(\mathbf{Au} - \mathbf{f})]_j|$ ▷ Select new index
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{\bar{j}\}$ ▷ Enlarge support
  $\mathbf{u} \leftarrow \arg\min_{\mathbf{z} \in \mathbb{C}^N} \|\mathbf{Az} - \mathbf{f}\|_2$ s.t. $\mathrm{supp}(\mathbf{z}) \subseteq \mathcal{S}$ ▷ Minimize residual
end while
return u

The OMP algorithm iteratively enlarges the support $\mathcal{S}$ of the sparse solution $\mathbf{u}$, by adding, at each step, the component $\bar{j} \in [N]$ corresponding to the column of $\mathbf{A}$ that maximizes the angle with respect to the residual $\mathbf{Au} - \mathbf{f}$. Then, the residual is orthogonally projected on the span of the columns of $\mathbf{A}$ corresponding to the indices in $\mathcal{S}$. This method is called "greedy", since it aims at reducing

the ℓ^2-norm of the residual as much as possible at each step. The stopping criterion can be related to the ℓ^2-norm of the residual, if we aim at solving $(\mathrm{P}_0^\varepsilon)$, defined in (1.7), or to the sparsity of the solution, if we consider the problem

$$(\mathrm{P}_0^s) \quad \min_{\mathbf{u}\in\mathbb{C}^N} \|\mathbf{A}\mathbf{u}-\mathbf{f}\|_2 \quad \text{s.t. } \|\mathbf{u}\|_0 \le s.$$

The OMP algorithm has been implemented in a very efficient way in [RZE08] through the MATLAB® package OMP-BOX [Rub09].

Recovery results A complete recovery theory for the OMP algorithm is still an open issue. So far, the existing theorems regarding the OMP performances rely on the concepts of coherence (Definition 1.8) and RIP (Definition 1.3). The common goal of the theorems that we are going to discuss is the same: provide sufficient conditions such that the OMP algorithm computes a solution $\mathbf{u}$ that fulfills the inequality

$$\|\mathbf{A}\mathbf{u}-\mathbf{f}\|_2 \lesssim \inf_{\mathbf{w}\in\Sigma_s^N} \|\mathbf{A}\mathbf{w}-\mathbf{f}\|_2,$$

using $\mathcal{O}(s)$ iterations.

The study of OMP based on the coherence was pioneered by A.C. Gilbert et al. in [GMS03] and then analyzed by J.A. Tropp in [Tro04], where sufficient conditions are provided in terms of $\mu(\mathbf{A})$. The first result presented here corresponds to [Tro04, Corollary 4.4].

Theorem 1.11. *For every $s \in \mathbb{N}$, with $s \le \frac{1}{3\mu(\mathbf{A})}$, the OMP algorithm computes a solution $\mathbf{u}$ such that*

$$\|\mathbf{A}\mathbf{u}-\mathbf{f}\|_2 \le \sqrt{1+6s} \inf_{\mathbf{w}\in\Sigma_s^N} \|\mathbf{A}\mathbf{w}-\mathbf{f}\|_2,$$

in s iterations.

The analysis in [Tro04] essentially relies on the *Exact Recovery Condition* (1.9). In particular, he shows that, if $\mathbf{f}$ admits a sparse representation with respect to the columns of $\mathbf{A}$, namely if there exists $\mathbf{u} \in \Sigma_s^N$ such that $\mathbf{f} = \mathbf{A}\mathbf{u}$, then OMP algorithm exactly recovers $\mathbf{u}$ if

$$\max\{\|\mathbf{A}_{\mathcal{S}}^+\mathbf{w}\|_1 : \operatorname{supp}(\mathbf{w}) \subseteq [N] \setminus \mathcal{S}\} < 1, \tag{1.9}$$

where $\mathcal{S} = \operatorname{supp}(\mathbf{u})$, $\mathbf{A}_{\mathcal{S}}$ is the submatrix of $\mathbf{A}$ built by keeping the columns with indices in $\mathcal{S}$, and $\mathbf{X}^+ := (\mathbf{X}^{\mathsf{H}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{H}}$ denotes the *Moore-Penrose* pseudoinverse of a matrix $\mathbf{X}$.

Of course, this condition is, from a concrete viewpoint, unpractical to verify, since it requires the knowledge of the support of $\mathbf{u}$, the exact solution. Nevertheless, condition (1.9) is the main ingredient needed to formulate a series of interesting results, like Theorem 1.11.

Another result based on the coherence is proved in [Liv12, Theorem 2] and relates the residual norm achieved after $2s$ steps of the OMP algorithm with the best s-term approximation error of $\mathbf{f}$ in the space generated by the columns of $\mathbf{A}$.

Theorem 1.12. *For every $s \in \mathbb{N}$, with $s \leq \frac{1}{20\mu(\mathbf{A})}$, the OMP algorithm computes a solution $\mathbf{u}$ such that*

$$\|\mathbf{A}\mathbf{u} - \mathbf{f}\|_2 \leq 2.7 \inf_{\mathbf{w} \in \Sigma_s^N} \|\mathbf{A}\mathbf{w} - \mathbf{f}\|_2,$$

in $2s$ iterations.

If compared with Theorem 1.11, the assumption on $\mu(\mathbf{A})$ is stronger, but the asymptotic constant on the right-hand side is now universally bounded.

The main problem regarding Theorems 1.11 and 1.12 is that they essentially rely on the coherence $\mu(\mathbf{A})$, that can be shown to be small enough only is some particular, not always realistic, situations. An alternative way for characterizing the performance of OMP, is based on the RIP (Definition 1.3). The next result was first proved by T. Zhang in [Zha11]. Then, its proof has been simplified and also generalized to the context of Hilbert spaces by A. Cohen et al. in [CDD15]. It corresponds to [CDD15, Theorem 1.1].

Theorem 1.13. *There exist fixed constants $K \in \mathbb{N}$, $C > 0$ and $\delta \in (0,1)$ such that for every $s \in \mathbb{N}$, the following holds: if $\mathbf{A} \in RIP((K+1)s, \delta)$, then, for any $\mathbf{f} \in \mathbb{C}^m$, the OMP algorithm computes a solution $\mathbf{u}$ that fulfills*

$$\|\mathbf{A}\mathbf{u} - \mathbf{f}\|_2 \leq C \inf_{\mathbf{w} \in \Sigma_s^N} \|\mathbf{A}\mathbf{w} - \mathbf{f}\|_2,$$

in Ks iterations.

Accelerations and extensions In Algorithm 1.1, the projection Step 6 is the most costly one, and it can be practically implemented as

$$\mathbf{u} \leftarrow \mathbf{A}_{\mathcal{S}}^{+}\mathbf{f} = (\mathbf{A}_{\mathcal{S}}^{\mathsf{H}}\mathbf{A}_{\mathcal{S}})^{-1}\mathbf{A}_{\mathcal{S}}^{\mathsf{H}}\mathbf{f}. \tag{1.10}$$

A common strategy to accelerate this step is to make a clever usage of the QR or Cholesky factorization. Indeed, the most heavy operation performed in (1.10) is the inversion of $\mathbf{A}_{\mathcal{S}}^{\mathsf{H}}\mathbf{A}_{\mathcal{S}}$. This can be done efficiently by noticing that, at each step, only the last row and the last column of $\mathbf{A}_{\mathcal{S}}^{\mathsf{H}}\mathbf{A}_{\mathcal{S}}$ are added. For example, if one has the Cholesky factorization of a matrix $\mathbf{M} = \mathbf{L}\mathbf{L}^{\mathsf{H}} \in \mathbb{C}^{k \times k}$, then the Cholesky factorization of

$$\mathbf{M}_{new} = \begin{bmatrix} \mathbf{M} & \mathbf{v} \\ \mathbf{v}^{\mathsf{H}} & x \end{bmatrix} \in \mathbb{C}^{(k+1)\times(k+1)},$$

with $\mathbf{v} \in \mathbb{C}^k$ and $x \in \mathbb{C}$, is easily updated as $\mathbf{M}_{new} = \mathbf{L}_{new}\mathbf{L}_{new}^{\mathsf{H}}$, with

$$\mathbf{L}_{new} = \begin{bmatrix} \mathbf{L} & \mathbf{0} \\ \mathbf{w}^{\mathsf{H}} & \sqrt{x - \mathbf{w}^{\mathsf{H}}\mathbf{w}} \end{bmatrix}, \quad \text{with } \mathbf{w} = \mathbf{L}^{-1}\mathbf{v}.$$

This simple algebraic consideration is the main idea of the accelerated OMP-Cholesky algorithm ([RZE08, Algorithm 2]). Analogously, one can employ the updated QR decomposition.

An interesting variant of OMP is the Weak Orthogonal Matching Pursuit (WOMP), where a relaxation parameter $\omega \in (0,1]$ is introduced and Step 4 is replaced with

$$\text{find } \bar{j} \in [N]: \quad |[\mathbf{A}^{\mathsf{H}}(\mathbf{A}\mathbf{u} - \mathbf{f})]_{\bar{j}}| \geq \omega \max_{j \in [N]} |[\mathbf{A}^{\mathsf{H}}(\mathbf{A}\mathbf{u} - \mathbf{f})]_j|.$$

When $\omega = 1$, WOMP coincides with OMP. Theorems analogous to those reported in this section hold for WOMP, see [Tro04, CDD15].

The OMP-BOX package employed in this thesis contains a very efficient implementation of the OMP algorithm called *Batch-OMP*, where $\mathbf{A}^{\mathsf{H}}\mathbf{A}$ and $\mathbf{A}^{\mathsf{H}}\mathbf{f}$ are precomputed and the Cholesky factorization is employed (see [RZE08, Algorithm 3]).

1.2.4 Bounded Orthonormal Systems

We present some recent results about CS, regarding a class of structured random matrices, arising from random sampling in a suitable finite-dimensional function space, spanned by a so-called *Bounded Orthonormal System* (BOS), i.e., an orthonormal basis whose elements are uniformly bounded with respect to the supremum norm. This presentation is mainly based on [Rau10] and [FR13, Chapter 12].

Definition 1.14 (Bounded Orthonormal System)**.** Let $\mathcal{D} \subseteq \mathbb{C}^d$ be endowed with a probability measure $\mathbb{P}$. A set of functions $\mathcal{B} = \{\beta_1, \dots, \beta_N\}$, with $\beta_j : \mathcal{D} \to \mathbb{C}$, is called a *Bounded Orthonormal System with constant* $K > 0$ if it fulfills

$$\int_{\mathcal{D}} \beta_j \overline{\beta}_k \, \mathrm{d}\,\mathbb{P} = \delta_{jk}, \quad \forall j,k \in [N].$$

and

$$\|\beta_j\|_\infty = \sup_{\mathbf{t} \in \mathcal{D}} |\beta_j(\mathbf{t})| \leq K, \quad \forall j \in [N].$$

Given a BOS, we can construct a family of structured random sampling matrices.

Definition 1.15. Given a BOS $\mathcal{B}$ and the vectors $\mathbf{t}_1, \dots, \mathbf{t}_m \in \mathcal{D}$ selected independently according to the probability measure $\mathbb{P}$, every matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ defined as

$$A_{ij} := \beta_j(\mathbf{t}_i), \quad \forall i \in [m], \ \forall j \in [N],$$

is called *sampling matrix associated with* $\mathcal{B}$.

The most remarkable property concerning the BOSs is that any sampling matrix fulfills the RIP with high probability, after a suitable rescaling. The proof of this result is quite complex. We report it here by omitting some technical details that can be found in [FR13, Theorem 12.31]. A variation of this argument will be employed in Chapter 3, to prove Theorem 1.21.

Theorem 1.16. *Let* $\mathbf{A} \in \mathbb{C}^{m \times N}$ *be a sampling matrix associated with a BOS with constant* $K \geq 1$. *Then, for every* $\delta \in (0,1)$, *provided*

$$m \gtrsim K^2 \delta^{-2} s \log^3(s) \log(N),$$

for some $s \geq K^2 \delta^{-2} \log(N)$, *it holds*

$$\mathbb{P}\{m^{-\frac{1}{2}} \mathbf{A} \in RIP(s, \delta)\} \geq 1 - N^{-\log^3(s)}.$$

Proof. The proof is divided into four parts. In Part I, we introduce a characterization of the RIP constant. Afterwards, we estimate its expectation (Part II) and we quantify its deviation from the expected value in probability (Part III). Finally, in Part IV we conclude by combining the conditions emerging from Part II and III.

Part I) Characterization of the RIP constant First, we define the matrix seminorm $|||\cdot|||_s$ for every matrix $\mathbf{B} \in \mathbb{C}^{N \times N}$ as

$$|||\mathbf{B}|||_s := \sup_{\mathbf{z} \in D_{s,N}} |\langle \mathbf{B}\mathbf{z}, \mathbf{z} \rangle|,$$

where $D_{s,N} := \Sigma_s^N \cap \{\mathbf{z} \in \mathbb{C}^N : \|\mathbf{z}\|_2 \leq 1\}$ is the set of s-sparse vectors belonging to the unit ball. Then, we use a characterization of the RIP constant depending on $|||\cdot|||_s$: namely, if we fix $s \in \mathbb{N}$, then

$$\delta_s := |||\widetilde{\mathbf{A}}^{\mathsf{H}} \widetilde{\mathbf{A}} - \mathbf{I}|||_s,$$

is the minimum positive constant such that $\widetilde{\mathbf{A}} := m^{-\frac{1}{2}} \mathbf{A} \in \mathrm{RIP}(s, \delta_s)$ (see [FR13, Definition 6.1]). Therefore, the goal is to prove that $\mathbb{P}\{\delta_s > \delta\} \leq N^{-\log^3(s)}$. Now, define the random column vectors

$$\mathbf{x}_i = (\overline{\beta_j(\mathbf{t}_i)})_{j=1}^N, \quad \forall i \in [m].$$

In such a way, $\mathbf{x}_i^{\mathsf{H}}$ is a row of $\mathbf{A}$ and the following decomposition holds

$$\widetilde{\mathbf{A}}^{\mathsf{H}}\widetilde{\mathbf{A}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}.$$

The orthogonality of the BOS implies $\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}] = \mathbf{I}$ and, consequently,

$$\delta_s = \frac{1}{m}|||\sum_{i=1}^{m}(\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}} - \mathbb{E}[\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}])|||_s.$$

Now, in order to estimate the expected value of δ_s, a symmetrization argument is employed, that allows us to pass from a sum of arbitrary independent random variables to a Rademacher sum, i.e.,

$$\mathbb{E}[|||\sum_{i=1}^{m}(\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}} - \mathbb{E}[\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}])|||_s] \leq 2\mathbb{E}[|||\sum_{i=1}^{m}\epsilon_i\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}|||_s],$$

where $\boldsymbol{\epsilon} = (\epsilon_i)_{i\in[m]}$ is a Rademacher sequence, i.e., a sequence of variables ϵ_i taking values ± 1 with equal probability, independently of the sampling points $\mathbf{t}_i$. Then, the following technical lemma is employed, concerning the expectation of a Rademacher sum.

Lemma 1.17. *Let $\mathbf{z}_1,\dots,\mathbf{z}_m$ be vectors in $\mathbb{C}^N$, with $\|\mathbf{z}_i\|_\infty \leq K$ for all $i \in [m]$. Then, for $s \leq m$,*

$$\mathbb{E}[|||\sum_{i=1}^{m}\epsilon_i\mathbf{z}_i\mathbf{z}_i^{\mathsf{H}}|||_s] \leq C_1 G(K,s,m,N)\sqrt{|||\sum_{i=1}^{m}\mathbf{z}_i\mathbf{z}_i^{\mathsf{H}}|||_s},$$

with $G(K,s,m,N) := K\sqrt{s}\log(4s)\sqrt{\log(8N)\log(9m)}$ and $C_1 \leq 27$.

Proving this lemma is maybe the hardest part of the whole proof, since one needs to consider a Rademacher process, define a pseudometric on $D_{s,N}$ and estimate the covering number of $D_{s,N}$ with respect to a suitable auxiliary seminorm. In particular, a lot of effort is made in order to control the universal constants. For the proof, see [FR13, Lemma 12.36].

We are now able to prove, first, an estimate in expectation of δ_s, and then to convert it in probability.

Part II) Estimate in expectation Define

$$E := \mathbb{E}[\delta_s] = \frac{1}{m}\mathbb{E}[|||\sum_{i=1}^{m}(\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}} - \mathbf{I})|||_s].$$

Then, in order to employ Lemma 1.17, we apply Fubini-Tonelli's Theorem and integrate separately with respect to the independent variables $\mathbf{X} := (\mathbf{x}_i)_{i\in[m]}$ and $\boldsymbol{\epsilon}$. We obtain

$$E \leq \frac{2}{m}\mathbb{E}_{\mathbf{X}}[\mathbb{E}_{\boldsymbol{\epsilon}}[|||\sum_{i=1}^{m}\epsilon_i\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}|||_s]] \leq \frac{2C_1G(K,s,m,N)}{\sqrt{m}}\mathbb{E}_{\mathbf{X}}[\sqrt{|||m^{-1}\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}|||_s}].$$

Then, adding and subtracting the identity matrix, employing the triangle inequality and Jensen's inequality, yields

$$E \leq \frac{2C_1G(K,s,m,N)}{\sqrt{m}}\mathbb{E}_{\mathbf{X}}[\sqrt{m^{-1}|||\sum_{i=1}^{m}(\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}-\mathbf{I})|||_s+1}] \leq D\sqrt{E+1},$$

where $D := 2C_1G(K,s,m,N)/\sqrt{m}$. Elementary algebraic manipulations show that

$$\begin{aligned} E \leq D\sqrt{E+1} &\Longrightarrow (E-D^2/2)^2 \leq D^2+D^4/4 \\ &\Longrightarrow E \leq \sqrt{D^2+D^4/4}+D^2/2 \\ &\Longrightarrow E \leq D+D^2. \end{aligned} \tag{1.11}$$

Therefore, requiring

$$D := \frac{2C_1K\sqrt{s}\log(4s)\sqrt{\log(8N)\log(9m)}}{\sqrt{m}} \leq \eta_1, \tag{1.12}$$

for some $\eta_1 \in (0,1)$, yields the desired estimate $E \leq \eta_1+\eta_1^2$.

Part III) Estimate in probability The second step is to estimate the deviation of δ_s from its expectation in probability. The principal tool employed will be the following deviation inequality for suprema of empirical processes above their mean, sometimes referred to as *Talagrand's inequality*. This result corresponds to [FR13, Theorem 8.42].

Lemma 1.18 (Talagrand's inequality)**.** *Let $\mathcal{G}$ be a countable set of functions $G : \mathbb{C}^n \to \mathbb{R}$. Let $Y_1,\ldots,Y_m$ be independent random vectors in $\mathbb{C}^n$ such that $\mathbb{E}[G(Y_i)] = 0$ and $G(Y_i) \leq R$ almost surely for all $i \in [m]$ and for all $G \in \mathcal{G}$ for some constant $R > 0$. Introduce*

$$Z = \sup_{G\in\mathcal{G}}\sum_{i=1}^{m}G(Y_i).$$

Let $\sigma_i^2 > 0$ such that $\mathbb{E}[G(Y_i)^2] \leq \sigma_i^2$ for all $G \in \mathcal{G}$ and $i \in [m]$. Then, for all $t > 0$,

$$\mathbb{P}\{Z \geq \mathbb{E}[Z] + t\} \leq \exp\left(- \frac{t^2/2}{\sigma^2 + 2R\mathbb{E}[Z] + tR/3} \right),$$

where $\sigma^2 = \sum_{i=1}^{m} \sigma_i^2$.

First, with simple algebraic manipulations it is possible to show the identity

$$m\delta_s = \sup_{(\mathbf{z},\mathbf{w}) \in Q_{s,N}^*} \sum_{i=1}^{m} G_{\mathbf{z},\mathbf{w}}(\mathbf{x}_i),$$

where $G_{\mathbf{z},\mathbf{w}}(\mathbf{x}) := \mathrm{Re}\langle (\mathbf{x}\mathbf{x}^{\mathsf{H}} - \mathbf{I})\mathbf{z}, \mathbf{w}\rangle$, for every $\mathbf{x} \in \mathbb{C}^N$, $Q_{s,N}^*$ is a countable dense subset of

$$Q_{s,N} := \bigcup_{\mathcal{S} \subseteq [N], |\mathcal{S}| \leq s} Q_{\mathcal{S},N}$$

with

$$Q_{\mathcal{S},N} := \{(\mathbf{z},\mathbf{w}) \in \mathbb{C}^N \times \mathbb{C}^N : \|\mathbf{z}\|_2 = \|\mathbf{w}\|_2 = 1,\ \mathrm{supp}(\mathbf{z}), \mathrm{supp}(\mathbf{w}) \subseteq \mathcal{S}\}.$$

Since $\mathbb{E}[G_{\mathbf{z},\mathbf{w}}(\mathbf{x}_i)] = 0$, for every $(\mathbf{z},\mathbf{w}) \in Q_{s,N}^*$, in order to apply Lemma 1.18, we need to provide upper bounds to $G_{\mathbf{z},\mathbf{w}}(\mathbf{x}_i)$ and $\mathbb{E}[|G_{\mathbf{z},\mathbf{w}}(\mathbf{x}_i)|^2]$.

Fix a pair $(\mathbf{z},\mathbf{w}) \in Q_{\mathcal{S},N}$, with $|\mathcal{S}| = s$. Then, we have

$$\begin{aligned} |G_{\mathbf{z},\mathbf{w}}(\mathbf{x}_i)| &\leq |\langle (\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}} - \mathbf{I})\mathbf{z}, \mathbf{w}\rangle| \leq \|\mathbf{z}\|_2 \|\mathbf{w}\|_2 \|(\mathbf{x}_i)_{\mathcal{S}}(\mathbf{x}_i)_{\mathcal{S}}^{\mathsf{H}} - \mathbf{I}_{\mathcal{S},\mathcal{S}}\|_2 \\ &\leq \|(\mathbf{x}_i)_{\mathcal{S}}(\mathbf{x}_i)_{\mathcal{S}}^{\mathsf{H}} - \mathbf{I}_{\mathcal{S}}\|_1 = \max_{j \in \mathcal{S}} \sum_{k \in \mathcal{S}} |\beta_j(\mathbf{t}_i)\overline{\beta_k(\mathbf{t}_i)} - \delta_{jk}| \\ &\leq sK^2, \end{aligned}$$

where we used the Cauchy-Schwarz inequality, the fact that $\|\mathbf{M}\|_2 \leq \|\mathbf{M}\|_1$, for any self-adjoint matrix $\mathbf{M}$, the definition of $\|\cdot\|_1$ for matrices and the boundedness condition with constant $K \geq 1$. Moreover, $(\mathbf{x}_i)_{\mathcal{S}}$ is the restriction of $\mathbf{x}_i$ to $\mathcal{S}$ and $\mathbf{I}_{\mathcal{S},\mathcal{S}}$ denotes the principal submatrix of $\mathbf{I}$ relative to the indices in $\mathcal{S}$.

Now, we estimate

$$\begin{aligned} \mathbb{E}[|G_{\mathbf{z},\mathbf{w}}(\mathbf{x}_i)|^2] &\leq \mathbb{E}[|\langle (\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}} - \mathbf{I})\mathbf{z}, \mathbf{w}\rangle|^2] \\ &\leq \mathbb{E}[\|((\mathbf{x}_i)_{\mathcal{S}}(\mathbf{x}_i)_{\mathcal{S}}^{\mathsf{H}} - \mathbf{I})\mathbf{z}\|_2^2] \\ &= \mathbb{E}[\|(\mathbf{x}_i)_{\mathcal{S}}\|_2^2 |\langle \mathbf{x}_i, \mathbf{z}\rangle|^2] - 2\mathbb{E}[|\langle \mathbf{x}_i, \mathbf{z}\rangle|^2] + 1, \end{aligned}$$

by algebraic manipulations, employing the Cauchy-Schwarz inequality and exploiting that $\|\mathbf{u}\mathbf{u}^{\mathsf{H}}\|_2 = \|\mathbf{u}\|_2^2$, for every $\mathbf{u} \in \mathbb{C}^N$. Now, observe that

$$\|(\mathbf{x}_i)_{\mathcal{S}}\|_2^2 = \sum_{j \in \mathcal{S}} |\beta_j(\mathbf{t}_i)|^2 \leq sK^2,$$

and that

$$\mathbb{E}[|\langle \mathbf{x}_i, \mathbf{z} \rangle|^2] = \sum_{j \in \mathcal{S}} \sum_{k \in \mathcal{S}} z_j \overline{z}_k \mathbb{E}[\beta_j(\mathbf{t}_i)\overline{\beta_k(\mathbf{t}_i)}] = \|\mathbf{z}\|_2^2 = 1.$$

As a consequence,

$$\mathbb{E}[|G_{\mathbf{z},\mathbf{w}}(\mathbf{x}_i)|^2] \leq sK^2.$$

Finally, applying Lemma 1.18 yields

$$\begin{aligned}\mathbb{P}\{\delta_s \geq \eta_1 + \eta_1^2 + \eta_2\} &\leq \mathbb{P}\{\delta_s \geq \mathbb{E}[\delta_s] + \eta_2\} = \mathbb{P}\{m\delta_s \geq \mathbb{E}[m\delta_s] + m\eta_2\} \\ &\leq \exp\left(- C_2(\eta_1) \frac{m\eta_2^2}{K^2 s} \right),\end{aligned}$$

for a suitable positive constant $C_2(\eta_1) \leq 3/32$, that is less than ε provided that

$$m \geq C_3 \eta_2^{-2} K^2 s \log(\varepsilon^{-1}), \tag{1.13}$$

with $C_3 = 32/3$.

Part IV) Conclusion To summarize, recalling (1.12) and (1.19), we proved that $\delta_s \leq \eta_1 + \eta_1^2 + \eta_2$ with probability at least $1 - \varepsilon$ provided that the conditions in Table 1.1-(c) be fulfilled. Using elementary algebraic arguments (see [FR13, Remark 12.33]), it can be proved that they are in turn implied by the conditions in Table 1.1-(a), with $\varepsilon = N^{-\log^3(s)}$, corresponding to what is claimed in the thesis.

□

The sufficient conditions in Theorem 1.16 that guarantee the RIP for $m^{-\frac{1}{2}}\mathbf{A}$ with high probability can be restated in several ways (see [FR13, Remark 12.33]). We summarize the different set of hypotheses in Table 1.1.

The great power of the BOS theory is its huge generality. Indeed, many situations that frequently occur in CS can be restated under the BOS framework, such as random subsampling from an orthonormal system of $\mathbb{C}^N$.

Discrete orthonormal systems Consider the case $\mathbf{\Psi} = \mathbf{I}$. Given a unitary matrix $\mathbf{U} = [\mathbf{u}_1|\cdots|\mathbf{u}_N] \in \mathbb{C}^{N \times N}$, if we set $\mathcal{D} = [N]$ and consider the discrete uniform measure on $\mathcal{D}$, namely

$$\mathbb{P}(q) := q/N, \quad \forall q \in [N],$$

	Sufficient condition	$\mathbb{P}\{m^{-\frac{1}{2}}\mathbf{A} \in \mathrm{RIP}(\delta, s)\}$
(a)	$\begin{cases} m \gtrsim \delta^{-2} K^2 s \log^3(s) \log(N) \\ s \geq K^2 \delta^{-2} \log(N) \end{cases}$	$\geq 1 - N^{-\log^3(s)}$
(b)	$m \gtrsim K^2 \delta^{-2} s \log^4(N)$	$\geq 1 - N^{-\log^3(N)}$
(c)	$\begin{cases} \frac{m}{\log(9m)} \gtrsim \eta_1^{-2} K^2 s \log^2(4s) \log(8N) \\ m \gtrsim \eta_2^{-2} K^2 s \log(\varepsilon^{-1}) \\ \delta = \eta_1 + \eta_1^2 + \eta_2 \end{cases}$	$\geq 1 - \varepsilon$

Table 1.1: Different sets of sufficient conditions for Theorem 1.16 to hold.

then the functions $\beta_j : [N] \to \mathbb{C}$, defined as $\beta_j(q) := \sqrt{N}[\mathbf{u}_j]_q$, form a BOS if

$$\max_{j \in [N]} \|\beta_j\|_\infty = \sqrt{N} \max_{q,j \in [N]} |U_{qj}| \leq K. \tag{1.14}$$

In this case, $\mathbf{\Phi}$ is built taking m columns from $\mathbf{U}$ drawn independently at random from $[N]$ according to the discrete uniform measure $\mathbb{P}$ and the resulting sensing matrix is $\mathbf{A} = \mathbf{\Phi}^{\mathsf{H}}$. The fact that random matrices with this structure (rescaled by a factor $m^{-1/2}$) satisfy the RIP with high probability is a consequence of Theorem 1.16, but was originally proved in [CR07]. Condition (1.14) essentially prevents the columns of $\mathbf{U}$ from begin too concentrated on just few entries. Let us analyze two diametrically opposed cases: (1) if a column $\mathbf{u}_k$ has just one non-zero entry (the information is maximally concentrated), the normalization condition $\|\mathbf{u}_k\|_2 = 1$ forces the left-hand side in (1.14) to be equal to $\sqrt{N}$; (2) if a column $\mathbf{u}_k$ is a constant vector (the information is totally scattered), then the left-hand side is equal to 1.

A remarkable example is $\mathbf{U} = \mathbf{F}$, the Fourier matrix defined in (1.3). Indeed, $|F_{qj}| = 1/\sqrt{N}$, for every $q, j \in [N]$.

By noticing that

$$\max_{q,j \in [N]} |U_{qj}| = \mu(\mathbf{U}, \mathbf{I}),$$

the boundedness condition (1.14) can be interpreted as a mutual incoherence between $\mathbf{U}$ and the canonical basis $\mathbf{I}$.

Mutually incoherent bases Another important application of the BOS theory is the case of two mutually incoherent bases. Let $\mathbf{U} = [\mathbf{u}_1|\cdots|\mathbf{u}_N], \mathbf{V} = [\mathbf{v}_1|\cdots|\mathbf{v}_N] \in \mathbb{C}^{N \times N}$ be two unitary and mutually incoherent matrices, i.e., such that

$$\mu(\mathbf{U}, \mathbf{V}) \leq \frac{K}{\sqrt{N}},$$

for some constant $K > 0$. Then, the matrix $\mathbf{W} := \mathbf{V}^{\mathsf{H}}\mathbf{U}$ is unitary, and we can apply the argument used before. In fact, it suffices to notice that

$$\sqrt{N} \max_{q,j\in[N]} |W_{qj}| = \sqrt{N} \max_{q,j\in[N]} |\langle \mathbf{v}_q, \mathbf{u}_j \rangle| = \sqrt{N}\mu(\mathbf{V},\mathbf{U}) \leq K.$$

Therefore, if we set $\boldsymbol{\Psi} = \mathbf{U}$ and build $\boldsymbol{\Phi}$ by taking m columns from $\mathbf{V}$ chosen independently at random using the uniform density, the resulting rescaled sensing matrix $m^{-1/2}\mathbf{A} = m^{-1/2}\boldsymbol{\Phi}^{\mathsf{H}}\boldsymbol{\Psi}$ satisfies the RIP with high probability thanks to Theorem 1.16.

1.2.5 Sampling strategies based on the local coherence

There is a nice application of the BOS theory, recently published in [KW14], that provides a general strategy to build sampling matrices, based on the concept of *local coherence*. This notion will be extremely useful for the theoretical study of CORSING, presented in Chapter 3.

Definition 1.19. Given two unitary matrices $\mathbf{U} = [\mathbf{u}_1|\cdots|\mathbf{u}_N], \mathbf{V} = [\mathbf{v}_1|\cdots|\mathbf{v}_N] \in \mathbb{C}^{N\times N}$, the *local coherence* of $\mathbf{U}$ with respect to $\mathbf{V}$ is a vector of $\mathbb{R}^N$ whose q-th component is[2]

$$\mu_q^{loc}(\mathbf{U},\mathbf{V}) := \max_{j\in[N]} |\langle \mathbf{u}_j, \mathbf{v}_q \rangle|^2, \quad \text{for } q \in [N].$$

Notice that, in contrast to the mutual coherence $\mu(\mathbf{U},\mathbf{V})$, the local coherence $\mu_q^{loc}(\mathbf{U},\mathbf{V})$ is not commutative in $\mathbf{U}$ and $\mathbf{V}$. Based on this definition and on Theorem 1.16, we provide a RIP theorem for locally incoherent bases. We also provide its simple and insightful proof [KW14].

Theorem 1.20. *Let $\mathbf{U},\mathbf{V} \in \mathbb{C}^{N\times N}$ be unitary matrices and $\delta \in (0,1)$. Assume the local coherence to have a componentwise upper bound $\boldsymbol{\nu} \in \mathbb{R}^N$*

$$\mu_q^{loc}(\mathbf{U},\mathbf{V}) \leq \nu_q, \quad \forall q \in [N],$$

and let $s, m \in \mathbb{N}$ be such that

$$s \gtrsim \log(N), \quad m \gtrsim \delta^{-2}\|\boldsymbol{\nu}\|_1 s \log^3(s)\log(N).$$

Define $\boldsymbol{\Psi} = \mathbf{U}$ and build $\boldsymbol{\Phi}$ by choosing m columns from $\mathbf{V}$, corresponding to the i.i.d. indices $\tau_1,\ldots,\tau_m$, independently at random, according to the probability

$$\mathbb{P}\{\tau_i = q\} = p_q := \frac{\nu_q}{\|\boldsymbol{\nu}\|_1}, \quad \forall q \in [N],\ \forall i \in [m].$$

[2]The local coherence defined here corresponds to the square of that defined in [KW14]. This notational choice will be useful in order to enlighten the notation in Chapter 3.

Consider the sensing matrix $\mathbf{A} = \mathbf{\Phi}^{\mathsf{H}}\mathbf{\Psi}$ *and the diagonal matrix* $\mathbf{D} \in \mathbb{C}^{m\times m}$ *with entries*

$$D_{ik} = \frac{\delta_{ik}}{\sqrt{mp_{\tau_i}}}, \quad \forall i,k \in [m].$$

Then,

$$\mathbb{P}\{\mathbf{DA} \in RIP(s,\delta)\} \geq 1 - N^{-\log^3(s)}.$$

Proof. First, we notice that the matrix $\mathbf{W} := \mathbf{V}^{\mathsf{H}}\mathbf{U}$ is unitary. Then, we set $\mathcal{D} = [N]$ and define the functions $\beta_j : [N] \to \mathbb{C}$ as

$$\beta_j(q) := \frac{1}{\sqrt{p_q}} W_{qj}, \quad \forall q,j \in [N].$$

Now, we show that $\{\beta_j\}$ is a BOS with respect to the probability measure $\mathbb{P}$. Indeed, the orthogonality follows from

$$\begin{aligned}\int_{\mathcal{D}} \beta_j \overline{\beta}_k \,\mathrm{d}\,\mathbb{P} &= \sum_{q=1}^{N} \beta_j(q)\overline{\beta_k(q)}p_q = \sum_{q=1}^{N}\left(\frac{1}{\sqrt{p_q}}W_{qj}\right)\left(\frac{1}{\sqrt{p_q}}\overline{W}_{qk}\right)p_q \\ &= \sum_{q=1}^{N} W_{qj}\overline{W}_{qk} = [\mathbf{W}^{\mathsf{H}}\mathbf{W}]_{k,j} = \delta_{k,j}, \quad \forall j,k \in [N],\end{aligned}$$

whereas, regarding the boundedness, we observe that

$$|\beta_j(q)|^2 = \frac{\|\boldsymbol{\nu}\|_1}{\nu_q}|\langle \mathbf{v}_q, \mathbf{u}_j\rangle|^2 \leq \|\boldsymbol{\nu}\|_1 =: K^2, \quad \forall q,j \in [N].$$

The thesis is now a direct consequence of Theorem 1.16. □

1.2.6 A guiding example: Haar *vs* Fourier

We provide here a particular choice for $\mathbf{\Psi}, \mathbf{\Phi} \in \mathbb{C}^{N\times N}$ that is very popular in CS, namely $\mathbf{\Psi}$ collects the Haar wavelets basis and $\mathbf{\Phi}$ is a subset of the Fourier basis. This pair of bases has been of strong inspiration for the development of CORSING (see Chapter 2). For further details we refer to [KW14].

The discrete Haar basis Fix $p \in \mathbb{N}$ and set $N = 2^p$. Then, the univariate Haar orthonormal basis $\mathbf{H}$ of $\mathbb{C}^N$ is built as follows. Let $\mathbf{c}$ be the constant vector of $\mathbb{C}^N$ whose components are equal to $2^{-p/2}$. Define the step function $\mathbf{h}_{0,0} = \mathbf{h}$ as

$$[\mathbf{h}]_j = \begin{cases} 2^{-p/2} & \text{if } 1 \leq j \leq 2^{p-1} \\ -2^{-p/2} & \text{if } 2^{p-1} < j \leq 2^p, \end{cases}$$

where $j \in \mathbb{N}$, with $1 \le j \le 2^p$ and build the corresponding dyadic translations $\mathbf{h}_{\ell,k}$ as

$$[\mathbf{h}_{\ell,k}]_j = 2^{\frac{\ell}{2}} [\mathbf{h}]_{2^\ell j - 2^p k} = \begin{cases} 2^{\frac{\ell-p}{2}} & \text{if } k2^{p-\ell} < j \le (k+\frac{1}{2})2^{p-\ell} \\ -2^{\frac{\ell-p}{2}} & \text{if } (k+\frac{1}{2})2^{p-\ell} < j \le (k+1)2^{p-\ell} \\ 0 & \text{otherwise,} \end{cases}$$

where $(\ell, k) \in \mathbb{N}^2$, with $0 < \ell < p$, $0 \le k < 2^\ell$ and $j \in \mathbb{N}$, for $1 \le j \le 2^p$. Then, the Haar basis of $\mathbb{C}^N$ is given by

$$\{\mathbf{c}\} \cup \{\mathbf{h}_{\ell,k} : 0 \le \ell < p,\ 0 \le k < 2^\ell\}.$$

For example, the Haar basis of $\mathbb{C}^4$ is

$$\mathbf{H} = \begin{bmatrix} 1/2 & 1/2 & 1/\sqrt{2} & 0 \\ 1/2 & 1/2 & -1/\sqrt{2} & 0 \\ 1/2 & -1/2 & 0 & 1/\sqrt{2} \\ 1/2 & -1/2 & 0 & -1/\sqrt{2} \end{bmatrix}.$$

Haar *vs* Fourier Unfortunately, the mutual coherence is a useless tool when considering the Haar and the Fourier discrete systems. Indeed, we have

$$\mu(\mathbf{H}, \mathbf{F}) = \langle \mathbf{c}, \mathbf{f}_0 \rangle = 1.$$

Nevertheless, this case can be perfectly analyzed employing the local coherence. First, we estimate the scalar products through explicit computations (see [KW14, Lemma 1])

$$|\langle \mathbf{f}_r, \mathbf{h}_{\ell,k} \rangle|^2 \le \min\left(\frac{6 \cdot 2^{\frac{\ell}{2}}}{|r|}, 3\pi 2^{-\frac{\ell}{2}} \right)^2, \quad \forall r \ne 0,\ \forall \ell, k.$$

Then, employing the inequality $\min(a,b)^2 \le ab$ for every $a, b > 0$, we obtain the following upper bound to the local coherence (we employ the translated test indices $r \in \mathbb{Z}, -N/2 < r \le N/2$)

$$\mu_r^{loc}(\mathbf{H}, \mathbf{F}) \le \frac{18\pi}{|r|} =: \nu_r, \quad \forall r \in \mathbb{Z},\ -\frac{N}{2} < r \le \frac{N}{2},\ r \ne 0.$$

Now, consider $\mathbf{\Psi} = \mathbf{H}$, and build $\mathbf{\Phi}$ as a random selection of m Fourier vectors from $\mathbf{F}$, drawn according to the following probability density

$$\mathbb{P}(r) = C_N \min\left(C_0, \frac{1}{|r|} \right), \quad \forall r \in \mathbb{Z},\ -\frac{N}{2} < r \le \frac{N}{2},$$

where C_N is a normalization constant and C_0 is introduced to avoid the singularity at $r = 0$. Then, as a direct consequence of Theorem 1.20, the sensing matrix $\mathbf{A} = \mathbf{\Phi}^{\mathsf{H}}\mathbf{\Psi}$ satisfies the RIP(δ, s) with overwhelming probability provided that

$$m \gtrsim \delta^{-2} s \log^3(s) \log^2(N), \quad s \gtrsim \log(N),$$

after a suitable diagonal preconditioning.

In [KW14], this result is also proved in the two-dimensional case, where both the bases are generalized through tensorization.

1.2.7 RIP for generic matrices

We prove a generalization of Theorem 1.20. In particular, given a matrix $\mathbf{B}$ and a lower (respectively, upper) bound to the minimum (respectively, maximum) eigenvalue of $\mathbf{B}^{\mathsf{H}}\mathbf{B}$, we can set up a suitable sampling strategy that guarantees the RIP for a random selection of m rows from $\mathbf{B}$ with high probability, adopting a suitable preconditioning. The proof is a modification of that of Theorem 1.16 and of Theorem 1.20.

The results in this section have been obtained in collaboration with Holger Rauhut and Sjoerd Dirksen, during a visit of the author to RWTH Aachen University in September 2015.

Theorem 1.21. *Consider $\mathbf{B} \in \mathbb{R}^{M\times N}$, with $M \geq N$, and suppose that there exist two constants $0 < r \leq R$ such that*

$$0 < r \leq \lambda_{\min}(\mathbf{B}^{\mathsf{H}}\mathbf{B}) \leq \lambda_{\max}(\mathbf{B}^{\mathsf{H}}\mathbf{B}) \leq R.$$

Moreover, assume that there exists a vector $\boldsymbol{\nu} \in \mathbb{R}^M$ such that

$$\max_{j\in[N]} |B_{qj}|^2 \leq \nu_q, \quad \forall q \in [M].$$

Then, for every $\delta \in (1 - r/R, 1)$, there exists a constant C such that, provided

$$m \geq \widetilde{C} s \log^3(s) \log(N)$$

and $s \geq \widetilde{C} \log(N)$, where

$$\widetilde{C} = C \max\{\|\boldsymbol{\nu}\|_1, R\} R^{-1} (\delta - 1 + r/R)^{-2},$$

choosing $\tau_1, \dots, \tau_m$ i.i.d. from $[M]$ according to the probability $p_q = \nu_q/\|\boldsymbol{\nu}\|_1$, it holds

$$\mathbb{P}\{\mathbf{DA} \in RIP(\delta, s)\} \geq 1 - N^{-\log^3(s)},$$

where

$$A_{ij} = B_{\tau_i, j}, \quad D_{ik} = \frac{\delta_{ik}}{\sqrt{mRp_{\tau_i}}}, \quad \forall i, k \in [m], \forall j \in [N].$$

Proof. The proof is divided in four parts, analogously to that of Theorem 1.16. First, we observe that, for any square symmetric matrix $\mathbf{M}$, it holds

$$|||\mathbf{M}|||_s \leq \rho(\mathbf{M}), \tag{1.15}$$

$\rho(\mathbf{M})$ being the spectral radius of $\mathbf{M}$.

Part I) Characterization of the RIP constant First, denoting by $\widetilde{\mathbf{A}} := \mathbf{DA}$, we have the decomposition

$$\widetilde{\mathbf{A}}^{\mathsf{H}}\widetilde{\mathbf{A}} = \mathbf{A}^{\mathsf{H}}\mathbf{D}^2\mathbf{A} = m^{-1}\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}},$$

with $\mathbf{x}_i := (Rp_{\tau_i})^{-1/2}\mathbf{B}^{\mathsf{H}}\mathbf{e}_{\tau_i}$. In particular, for every $i \in [m]$, it holds

$$\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}] = \sum_{q=1}^{M} p_q(p_qR)^{-1}\mathbf{B}^{\mathsf{H}}\mathbf{e}_q\mathbf{e}_q^{\mathsf{H}}\mathbf{B} = R^{-1}\mathbf{B}^{\mathsf{H}}\mathbf{B} =: \mathbf{E},$$

and, thanks to the normalization chosen, we notice that

$$0 < r/R \leq \lambda_{\min}(\mathbf{E}) \leq \lambda_{\max}(\mathbf{E}) \leq 1. \tag{1.16}$$

Employing the triangle inequality, we obtain

$$\delta_s = |||\widetilde{\mathbf{A}}^{\mathsf{H}}\widetilde{\mathbf{A}} - \mathbf{I}|||_s \leq |||\widetilde{\mathbf{A}}^{\mathsf{H}}\widetilde{\mathbf{A}} - \mathbf{E}|||_s + |||\mathbf{E} - \mathbf{I}|||_s.$$

We bound the deterministic term, using (1.15) and (1.16), as

$$|||\mathbf{E} - \mathbf{I}|||_s \leq \rho(\mathbf{E} - \mathbf{I}) = 1 - \lambda_{\min}(\mathbf{E}) \leq 1 - r/R.$$

Now, defining $\delta_s^* := |||\widetilde{\mathbf{A}}^{\mathsf{H}}\widetilde{\mathbf{A}} - \mathbf{E}|||_s$ and $\delta^* := \delta - 1 + r/R$, the goal is to find sufficient conditions such that the upper bound $\delta_s^* \leq \delta^*$ holds with high probability. Indeed, $\delta_s^* \leq \delta^*$ implies $\delta_s \leq \delta$. Finally, notice that $\delta \geq 1 - r/R$.

Now, defining

$$K := \sqrt{\frac{\max\{\|\boldsymbol{\nu}\|_1, R\}}{R}} \geq 1,$$

the random vectors $\mathbf{x}_i$ satisfy the uniform upper bound

$$\begin{aligned}\|\mathbf{x}_i\|_\infty^2 &= (Rp_{\tau_i})^{-1}\|\mathbf{B}^{\mathsf{H}}\mathbf{e}_{\tau_i}\|_\infty^2 = (Rp_{\tau_i})^{-1}\max_{j\in[N]}|B_{\tau_i,j}|^2 \\ &\leq R^{-1}\frac{\|\boldsymbol{\nu}\|_1}{\nu_{\tau_i}}\nu_{\tau_i} = \frac{\|\boldsymbol{\nu}\|_1}{R} \leq K^2.\end{aligned} \tag{1.17}$$

Part II) Estimate in expectation Define $E := \mathbb{E}[\delta_s^*]$. Then, analogously to the proof of Theorem 1.16, we rewrite δ_s^* as

$$\delta_s^* = \frac{1}{m}|||\sum_{i=1}^{m}(\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}} - \mathbf{E})|||_s,$$

and employ Lemma 1.17, obtaining

$$\begin{aligned} E &\leq \frac{2C_1G(K,s,m,N)}{\sqrt{m}}\mathbb{E}_{\mathbf{X}}[\sqrt{|||\frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}|||_s}] \\ &\leq \frac{2C_1G(K,s,m,N)}{\sqrt{m}}\mathbb{E}_{\mathbf{X}}[\sqrt{|||\frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}} - \mathbf{E})|||_s + |||\mathbf{E}|||_s}] \\ &\leq D\sqrt{E+1}, \end{aligned}$$

where $D = 2C_1G(K,s,m,N)/\sqrt{m}$ and we used that $|||\mathbf{E}|||_s \leq 1$. After few elementary algebraic manipulations analogous to (1.11), we obtain $E \leq D + D^2$. Hence, imposing $D \leq \eta_1$, with $\eta_1 \in (0,1)$, yields $E \leq \eta_1 + \eta_1^2$.

Part III) Estimate in probability First, we observe that

$$m\delta_s^* := \sup_{(\mathbf{z},\mathbf{w})\in Q_{s,N}^*}\sum_{i=1}^{m}G_{\mathbf{z},\mathbf{w}}(\mathbf{x}_i),$$

where $G_{\mathbf{z},\mathbf{w}}(\mathbf{x}) := \mathrm{Re}\langle(\mathbf{x}\mathbf{x}^{\mathsf{H}} - \mathbf{E})\mathbf{z},\mathbf{w}\rangle$. Now, we check the hypotheses needed in order to apply Lemma 1.18, fixing a pair $(\mathbf{z},\mathbf{w}) \in Q_{\mathcal{S},N}$, with $|\mathcal{S}| = s$.

It can be easily checked that $\mathbb{E}[G_{\mathbf{z},\mathbf{w}}(\mathbf{x}_i)] = 0$.

In order to check the boundedness condition, recalling (1.17), we observe that, for every $j,k \in \mathcal{S}$, it holds

$$|E_{jk}| = R^{-1}|[\mathbf{B}^{\mathsf{H}}\mathbf{B}]_{jk}| = R^{-1}\sum_{q=1}^{M}|\overline{B}_{qj}B_{qk}| \leq R^{-1}\sum_{q=1}^{M}\nu_q = R^{-1}\|\boldsymbol{\nu}\|_1 \leq K^2,$$

and

$$|[(\mathbf{x}_i)_{\mathcal{S}}(\mathbf{x}_i)_{\mathcal{S}}]_{jk}| = |[\mathbf{x}_i]_j[\mathbf{x}_i]_k| \leq \|\mathbf{x}_i\|_\infty^2 \leq K^2.$$

Then, analogously to the proof of Theorem 1.16, we obtain

$$|G_{\mathbf{z},\mathbf{w}}(\mathbf{x}_i)| \leq \|(\mathbf{x}_i)_{\mathcal{S}}(\mathbf{x}_i)_{\mathcal{S}}^{\mathsf{H}} - \mathbf{E}_{\mathcal{S},\mathcal{S}}\|_1 = \max_{j\in\mathcal{S}}\sum_{k\in\mathcal{S}}|[(\mathbf{x}_i)_{\mathcal{S}}(\mathbf{x}_i)_{\mathcal{S}}]_{jk} - E_{jk}| \leq 2sK^2,$$

where $\mathbf{E}_{\mathcal{S},\mathcal{S}}$ denotes the principal submatrix of $\mathbf{E}$ relative to the indices in $\mathcal{S}$.

Now, we check the condition on the second moment

$$\begin{aligned}\mathbb{E}[|G_{\mathbf{z},\mathbf{w}}(\mathbf{x}_i)|^2] &\le \mathbb{E}[\|((\mathbf{x}_i)_{\mathcal{S}}\mathbf{x}_i^{\mathsf{H}} - \mathbf{E})\mathbf{z}\|_2^2] \\ &= \mathbb{E}[\|(\mathbf{x}_i)_{\mathcal{S}}\mathbf{x}_i^{\mathsf{H}}\mathbf{z}\|_2^2 - 2\langle(\mathbf{x}_i)_{\mathcal{S}}\mathbf{x}_i^{\mathsf{H}}\mathbf{z}, \mathbf{E}\mathbf{z}\rangle + \|\mathbf{E}\mathbf{z}\|_2^2] \\ &\le \mathbb{E}[\|(\mathbf{x}_i)_{\mathcal{S}}\|_2^2|\langle\mathbf{x}_i,\mathbf{z}\rangle|^2] + 2\mathbb{E}[\|(\mathbf{x}_i)_{\mathcal{S}}\|_2|\langle\mathbf{x}_i,\mathbf{z}\rangle|]\|\mathbf{E}\|_2 + \|\mathbf{E}\|_2^2. \quad (1.18)\end{aligned}$$

We also notice that $\|(\mathbf{x}_i)_{\mathcal{S}}\|_2^2 \le sK^2$ and that

$$\|\mathbf{E}\|_2 = \rho(\mathbf{E}^{\mathsf{H}}\mathbf{E}) = \rho(\mathbf{E}^2) = \rho(\mathbf{E})^2 \le 1,$$

where we exploit that $\mathbf{E}$ is symmetric positive definite. Moreover,

$$\mathbb{E}[|\langle\mathbf{x}_i,\mathbf{z}\rangle|^2] = \mathbb{E}[\overline{(\mathbf{x}_i^{\mathsf{H}}\mathbf{z})}\mathbf{x}_i^{\mathsf{H}}\mathbf{z}] = \mathbf{z}^{\mathsf{H}}\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}]\mathbf{z} = \mathbf{z}^{\mathsf{H}}\mathbf{E}\mathbf{z} \le \rho(\mathbf{E}) \le 1.$$

Thus, recalling (1.18), we obtain that

$$\begin{aligned}\mathbb{E}[|G_{\mathbf{z},\mathbf{w}},(\mathbf{x}_i)|^2] &\le sK^2\mathbb{E}[|\langle\mathbf{x}_i,\mathbf{z}\rangle|^2] + \sqrt{s}K\mathbb{E}[|\langle\mathbf{x}_i,\mathbf{z}\rangle|] + 1 \\ &\le sK^2 + \sqrt{s}K + 1 \le sK^2\left(1 + \frac{1}{\sqrt{s}K} + \frac{1}{sK^2}\right) \le 3sK^2,\end{aligned}$$

where we employed Jensen's inequality to bound $\mathbb{E}[|\langle\mathbf{x}_i,\mathbf{z}\rangle|] \le 1$ and the fact that $K \ge 1$.

Finally, applying Lemma 1.18 yields

$$\begin{aligned}\mathbb{P}\{\delta_s^* \ge \eta_1 + \eta_1^2 + \eta_2\} \le \mathbb{P}\{\delta_s^* \ge \mathbb{E}[\delta_s^*] + \eta_2\} &= \mathbb{P}\{m\delta_s^* \ge \mathbb{E}[m\delta_s^*] + m\eta_2\} \\ &\le \exp\left(-C_2(\eta_1)\frac{m\eta_2^2}{K^2 s}\right),\end{aligned}$$

for a suitable positive constant $C_2(\eta_1) \le 1/32$, that is less than ε provided

$$m \ge C_3\eta_2^{-2}K^2 s\log(\varepsilon^{-1}), \quad (1.19)$$

with $C_3 = 32$. Notice that we are adding a factor 3 with respect to the analogous constant in Theorem 1.16.

Part IV) Conclusion The concluding remarks are analogous to those of Theorem 1.16.

□

Chapter 2

CORSING: Towards a theoretical understanding

In this chapter, we present the CORSING (COmpRessed SolvING) method. The exposition will be mostly empirical and heuristic, following the same process that brought us towards the discovery and the formalization of this new method. A significant part of this chapter corresponds to the contents in [BMP15].

The approach followed will lead us towards a theoretical understanding of CORSING, that will be presented in Chapter 3.

Outline of the chapter In Section 2.1, we review the classical Petrov-Galerkin method for advection-diffusion-reaction problems, collecting the most important elements from the Babuška-Nečas theory. In Section 2.2, we present the CORSING method, in its first formulation [BMP15]. Finally, an extensive numerical assessment is presented for the one-dimensional case (Section 2.3) and the two-dimensional case (Section 2.4)

2.1 The Petrov-Galerkin method

2.1.1 Weak problems in Hilbert spaces

Consider a *weak problem* of the form

$$\text{find } u \in U: \quad a(u,v) = \mathcal{F}(v), \quad \forall v \in V, \tag{2.1}$$

where U and V are Hilbert spaces equipped with norms $\|\cdot\|_U$, $\|\cdot\|_V$, respectively, $a : U \times V \to \mathbb{R}$ is a bilinear form and $\mathcal{F} \in V^*$, V^* being the dual space of V.

In particular, we focus on the scalar homogeneous *advection-diffusion-reaction* (ADR) problem

$$\begin{cases} -\operatorname{div}(\eta \nabla u) + \mathbf{b} \cdot \nabla u + \rho u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{2.2}$$

where η, $\mathbf{b}$, ρ and f are given functions defined on a sufficiently smooth open domain $\Omega \subset \mathbb{R}^d$, with $d \in \mathbb{N}$, and u is the unknown scalar field defined on $\overline{\Omega}$.

Problem (2.2) admits the generalized weak formulation (2.1), where U and V are suitable Hilbert spaces, a priori distinct, $a : U \times V \to \mathbb{R}$ and $\mathcal{F} : V \to \mathbb{R}$ are the bilinear and the linear forms defined by

$$a(u,v) = (\eta \nabla u, \nabla v) + (\mathbf{b} \cdot \nabla u, v) + (\rho u, v), \quad \mathcal{F}(v) = (f, v), \quad \forall u \in U, \forall v \in V,$$

$(\cdot,\cdot)$ denoting the $L^2(\Omega)$-inner product. Standard notation is employed for all the Lebesgue and Sobolev spaces and their norms [LM72].

The particular choice $U = V$ in (2.1) identifies the classical weak formulation associated with a standard Galerkin formulation whose well-posedness is guaranteed by Lax-Milgram's Lemma, where the sufficient hypotheses required on $a(\cdot,\cdot)$ are *continuity* and *coercivity*.

Theorem 2.1 (Lax-Milgram's Lemma)**.** *Consider problem* (2.1)*, with* $U = V$*. If* $a(\cdot,\cdot)$ *is* continuous*, i.e.,*

$$\exists \beta > 0 : \quad |a(u,v)| \leq \beta \|u\|_U \|v\|_U, \quad \forall u, v \in U, \tag{2.3}$$

and coercive*, i.e.,*

$$\exists \alpha > 0 : \quad a(u,u) \geq \alpha \|u\|_U^2, \quad \forall u \in U,$$

then there exists a unique solution $u \in U$ *to* (2.1) *that satisfies the a priori estimate*

$$\|u\|_U \leq \frac{1}{\alpha} \|\mathcal{F}\|_{U^*}.$$

In the more general case, $U \neq V$, the coercivity of $a(\cdot,\cdot)$ is replaced by the well-known *inf-sup condition*, that will play a key role in this thesis. We have the following generalization of the Lax-Milgram Lemma, due to Nečas [Neč62].

Theorem 2.2 (Nečas)**.** *Consider the problem* (2.1) *with* $a(\cdot,\cdot)$ *continuous. Then,* (2.1) *admits a unique solution* $u \in U$ *if and only if* $a(\cdot,\cdot)$ *satisfies*

$$\exists \alpha > 0 : \quad \inf_{u \in U} \sup_{v \in V} \frac{a(u,v)}{\|u\|_U \|v\|_V} \geq \alpha, \tag{2.4}$$

$$\sup_{u \in U} a(u,v) > 0, \quad \forall v \in V \setminus \{0\}. \tag{2.5}$$

Moreover, the following a priori estimate holds

$$\|u\|_U \leq \frac{1}{\alpha} \|\mathcal{F}\|_{V^*}.$$

Proof. A complete proof can be found in [QV08, Theorem 5.1.2] or [EG13, Theorem 2.6]. The existence and uniqueness essentially rely on the Riesz representation theorem, whereas the a priori estimate is an immediate consequence of (2.4), indeed

$$\alpha\|u\|_U \leq \sup_{v\in V} \frac{a(u,v)}{\|v\|_V} = \sup_{v\in V} \frac{\mathcal{F}(v)}{\|v\|_V} = \|\mathcal{F}\|_{V^*}.$$

□

Condition (2.4) is usually called *inf-sup condition*, but is also referred to as *Ladyženskaja-Babuška-Brezzi* (LBB) condition and will be of vital importance for the theoretical analysis of CORSING, performed in Chapter 3.

The inf-sup condition can also be employed with $U = V$, when coercivity does not hold (and, thus, Theorem 2.1 cannot be applied). This is the case of the *Stokes problem*, where $U = V = H_0^1(\Omega) \times (L^2(\Omega)/\mathbb{R})$ (see Section 4.1).

2.1.2 From weak problems to linear systems

A very popular, and nowadays classical, way to numerically approximate a solution to (2.1) is the so-called *Petrov-Galerkin* (PG) method. This approach is a generalization of the Galerkin method (also known as Bubnov-Galerkin or Ritz-Galerkin method), and it is sometimes also referred to as the nonstandard Galerkin method.

A century of history The history of the PG method dates back to the beginning of the 20$^{\text{th}}$ century, when W. Ritz set up a mathematical method to study the deformation of an elastic plate under an external force [Rit08] and to compute the *Chaldni figures* [Chl87] produced by the sand on a metal plate, when excited using the bow of a violin [Rit09]. Ritz's method was immediately employed in Russia, in order to solve hard engineering problems, by remarkable scientists such as S.P. Timoshenko [Tim13], I. Bubnov [Bub13], B.G. Galerkin [Gal15] and G.I. Petrov [Pet40], who first proposed the variant presented here. The PG method was then formalized during the early 1970's by A.K. Aziz and I. Babuška [AB72]. To learn more details about the thrilling history of the PG method, see the nice review [GW12].

The essence and the power of the PG method is very well summarized by S.P. Timoshenko [Tim13]:

> "We will not address the mathematical aspects of this method: a remarkable publication of a Swiss scientist, Mr. Walter Ritz, was dedicated to this subject. Transforming the problem of integrating the equations into a problem of evaluating integrals, Mr. W. Ritz has

> shown for a large class of problems, that by increasing the parameters a_1, a_2, a_3,..., one can find the exact solution of the problem."[1]

The PG method essentially transforms the weak problem (2.1) into a linear system of equations. This is achieved in four steps: (1) introduce two finite-dimensional spaces, called trial and test spaces; (2) consider a finite dimensional approximation to (2.1) restricted to the trial and test spaces; (3) choose two (possibly different) bases for the trial and test spaces; (4) evaluate the bilinear form $a(\cdot,\cdot)$ and the linear operator $\mathcal{F}$ on these basis functions.

The finite dimensional weak problem Consider two finite dimensional subspaces $U^N \subseteq U$ of dimension N and $V^m \subseteq V$ of dimension m, called *trial space* and *test space*, respectively. Then, the idea is to consider a finite dimensional approximation of problem (2.1), defined as follows[2]

$$\text{find } \widehat{u} \in U^N : \quad a(\widehat{u}, v) = \mathcal{F}(v), \quad \forall v \in V^m. \tag{2.6}$$

In this chapter, we consider $U = V = H_0^1(\Omega)$, but U^N and V^m, in general, do not coincide, yielding a PG formulation. When $N = m$ and $U^N = V^m$, one has a Galerkin formulation.

In the following, the solution $\widehat{u}$ to the finite dimensional weak problem (2.6), will be sometimes denoted as $\widehat{u}_m^N$, when the dimension of the trial space and of the test space need to be tracked.

The stability and the convergence of the formulation (2.6) are analyzed in the following theorem due to A.K. Aziz and I. Babuška [AB72].

Theorem 2.3 (Aziz-Babuška)**.** *Consider two linear subspaces $U^N \subseteq U$ and $V^m \subseteq V$. If the bilinear form $a(\cdot,\cdot)$ fulfills the continuity condition* (2.3) *and satisfies the conditions*

$$\exists \widetilde{\alpha} > 0 : \quad \inf_{u \in U^N} \sup_{v \in V^m} \frac{a(u,v)}{\|u\|_U \|v\|_V} \geq \widetilde{\alpha}, \tag{2.7}$$

$$\sup_{u \in U^N} a(u,v) > 0, \quad \forall v \in V^m \setminus \{0\}, \tag{2.8}$$

then problem (2.6) *admits a unique solution that fulfills the a priori estimate*

$$\|\widehat{u}\|_U \leq \frac{1}{\widetilde{\alpha}} \|\mathcal{F}\|_{V^*},$$

[1] $a_1, a_2, a_3, \ldots$ are the degrees of freedom of the linear system, i.e., the coefficients with respect to the trial basis functions.

[2] In the literature the subspaces U^N and V^m are usually called U_h and V_h, and the discrete approximate solution is usually denoted u_h, where h is a parameter related to the mesh that discretizes the physical domain Ω, e.g., the maximum among the elements' diameters. In this exposition we prefer to get rid of this notation, since our discretizations can not be summarized by a single scalar parameter h.

and the error estimate

$$\|u - \widehat{u}\|_U \leq \left(1 + \frac{\beta}{\widetilde{\alpha}}\right) \inf_{w \in U^N} \|u - w\|_U, \tag{2.9}$$

where u is the solution to (2.1).

Proof. The existence, uniqueness and the a priori estimate are a consequence of Theorem 2.2. In order to prove the error estimate, first we notice that (2.7) implies

$$\widetilde{\alpha}\|\widehat{u} - w\|_U \leq \sup_{v \in V^m} \frac{a(\widehat{u} - w, v)}{\|v\|_V} = \sup_{v \in V^m} \frac{a(u - w, v)}{\|v\|_V} \leq \beta\|u - w\|_U, \quad \forall w \in U^N, \tag{2.10}$$

where we exploited the property $a(\widehat{u} - u, v) = 0$, for every $v \in V^m$ (usually called *Galerkin orthogonality*) and the continuity of $a(\cdot, \cdot)$. Now, the triangle inequality implies

$$\|u - \widehat{u}\|_U \leq \inf_{w \in U^N} (\|u - w\|_U + \|\widehat{u} - w\|_U).$$

Combining the last inequality with relation (2.10) yields (2.9). □

Theorem 2.3 shows that the approximation error committed by solving (2.6) is near-optimal, in the sense that it is bounded, up to a constant factor, by the best approximation error of u in the space U^N. Estimate (2.9) is usually referred to as *Céa's Lemma*.

Moreover, Theorem 2.3 highlights the importance of the discrete inf-sup condition (2.7): the goal is to build suitable finite dimensional subspaces U^N and V^m such that the constant $\widetilde{\alpha}$ does not become too small. The undesirable case of a small $\widetilde{\alpha}$ forces the constant factor in (2.9) to become large, thus not guaranteeing that the approximate solution $\widehat{u}$ is sufficiently close to the true solution u.

Remark 2.1.1 (Further extensions). There are several possible variations on this scenario: for example, we could consider the approximation errors due to numerical integration, yielding an approximate bilinear form $\widehat{a}(\cdot, \cdot)$ and an approximate linear operator $\widehat{\mathcal{F}}$; or we could substitute the original norms $\|\cdot\|_U$ and $\|\cdot\|_V$ with some discrete variants $\|\cdot\|_{U^N}$ and $\|\cdot\|_{V^m}$. These issues will not be addressed here and for a further discussion the reader is referred to [QV08, EG13]. □

The PG discretization Assume U^N and V^m to be generated by a basis of *trial functions* $\{\psi_1, \ldots, \psi_N\}$ and *test functions* $\{\varphi_1, \ldots, \varphi_m\}$, respectively. Namely,

$$U^N = \text{span}\{\psi_1, \ldots, \psi_N\}, \quad V^m = \text{span}\{\varphi_1, \ldots, \varphi_m\}. \tag{2.11}$$

Then, since $V^m = \text{span}\{\varphi_i\}$, the weak formulation (2.6) is then equivalent to

$$\text{find } \widehat{u} \in U^N : \quad a(\widehat{u}, \varphi_i) = \mathcal{F}(\varphi_i), \quad \text{for } i = 1, \dots, m. \tag{2.12}$$

If we expand $\widehat{u} \in U^N$ as a linear combination of the trial functions $\{\psi_j\}$, problem (2.12) can be written as the linear system

$$\mathbf{A}\widehat{\mathbf{u}} = \mathbf{f}, \tag{2.13}$$

where $\mathbf{A} \in \mathbb{R}^{m \times N}$ is the *generalized stiffness matrix* that represents the bilinear form $a(\cdot, \cdot)$ with respect to the bases $\{\psi_j\}$, $\{\varphi_i\}$, and $\mathbf{f} \in \mathbb{R}^m$ is the *generalized load vector*, namely,

$$A_{ij} = a(\psi_j, \varphi_i), \quad f_i = \mathcal{F}(\varphi_i), \tag{2.14}$$

with $i = 1, \dots, m$ and $j = 1, \dots, N$. The unknown vector $\widehat{\mathbf{u}} \in \mathbb{R}^N$ contains the coefficients of the discrete solution $\widehat{u}$ expressed in terms of the trial basis $\{\psi_j\}$.

The solution $\widehat{\mathbf{u}}$ to (2.13), will be sometimes denoted as $\widehat{\mathbf{u}}_m^N$, in order to keep track of the trial and test space dimensions.

full-PG *vs* corsed-PG approximation We denote formulation (2.6) associated with the choice (2.11) and $m = N$ by full-PG, corresponding to the classical approach. Goal of this work is to provide a computationally efficient technique to approximate the full-PG solution by picking $m \ll N$, i.e., using far fewer tests than trials. The resulting approximation is denoted by corsed-PG. This new approach leads us to deal with a highly underdetermined system (2.13), that will be solved employing tools from sparse recovery, such as the ℓ_0- and the ℓ_1-minimization. Moreover, at this stage, we cannot assume the stiffness matrix $\mathbf{A}$ to have a particular sparsity pattern, $\mathbf{A}$ being possibly a full matrix. To define a corsed-PG approximation, we exploit ideas and techniques inspired by Compressed Sensing.

2.2 CORSING: COmpRessed SolvING

Let us now enter the core of this chapter: the explanation of the proposed approximation strategy, referred to as COmpRessed SolvING or, more briefly, CORSING.

2.2.1 Description of the methodology

The CORSING method consists of two distinct phases: the *assembly* and the *recovery* phase. In the first one, the generalized stiffness matrix and the generalized load vector are built. The second one deals with the actual computation of the corsed-PG solution, outcome of the CORSING approach.

We now describe the two phases in detail.

Assembly phase The assembly phase essentially forms the generalized stiffness matrix and load vector in (2.14). In turn, this phase is divided in three steps:

1. choose two sets of N independent vectors in $H_0^1(\Omega)$ for the full-PG formulation: the trial functions $\{\psi_1,\dots,\psi_N\}$, and the test functions $\{\varphi_1,\dots,\varphi_N\}$;

2. choose an integer $m < N$ (desirably $m \ll N$) and select a subset of m test functions $\{\varphi_{\tau_1},\dots,\varphi_{\tau_m}\}$, where $\tau_i \in \{1,\dots,N\}$ for $i = 1,\dots,m$;

3. build the generalized stiffness matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ and load vector $\mathbf{f} \in \mathbb{C}^m$, defined as
$$A_{ij} = a(\psi_j, \varphi_{\tau_i}) \quad f_i = \mathcal{F}(\varphi_{\tau_i}).$$

The assembly phase is, in general, the most costly one (see Section 2.3.1).

Recovery phase Goal of the recovery phase is to compute the corsed-PG solution $\widehat{\mathbf{u}}_m^N$ to the underdetermined linear system $\mathbf{Au} = \mathbf{f}$ (i.e., the *corsed-PG solution* $\widehat{u}_m^N$ of (2.2) via the basis $\{\psi_j\}$). The solution is computed through either problem (P_0) or (P_1), defined in (1.5) and (1.6), respectively.

Notice that the global sensing matrix characterizing the standard CS approach is here replaced by the generalized stiffness matrix associated with the PG formulation (2.6). Concerning the selection of the trial functions at the step 1. of the assembly phase, one should essentially try to get a discrete solution as sparse as possible or, at least, compressible with respect to that basis, namely the sparsity prior for the corsed-PG solution $\widehat{\mathbf{u}}_m^N$ should be guaranteed. As for the test functions, at the step 2., the choice of the test indices $\{\tau_1,\dots,\tau_m\}$ can be carried out in either a deterministic or a randomized way. We denote by D-CORSING the first approach, and by R-CORSING the randomized strategy. In particular, we adopt the two following extraction procedures:

- D-CORSING: set $\tau_i = i$ for $i = 1,\dots,m$;

- R-CORSING: by using Algorithm 2.1, this procedure selects m different numbers $\{\tau_1,\dots,\tau_m\}$ out of the set $\{1,\dots,N\}$, each number i having a probability proportional to a given weight w_i of being drawn without repetitions.

Notice that R-CORSING requires a more involved numerical assessment due to the randomized nature of the approach. Indeed, to analyze the results using statistical tools, we need to perform multiple runs of the same experiment. Numerical experiments show that a uniform random selection of the τ_i's (i.e., $w_i = \text{const}$, $\forall i$ in Algorithm 2.1) does not correctly work, in general. Vice versa,

Algorithm 2.1

procedure TestSelection($N, m, \{w_1, \dots, w_N\}$)
  $\mathcal{U} \leftarrow \{1, \dots, N\}$ ▷ The urn initially contains all the indices
  $\mathcal{T} \leftarrow \emptyset$ ▷ The selected indices set is initially empty
  for $i = 1, \dots, m$ **do**
    define $\mathbb{P}(k) = w_k \Big/ \sum_{k' \in \mathcal{U}} w_{k'}$, $\forall k \in \mathcal{U}$ ▷ Probability distribution on $\mathcal{U}$
    randomly select $\tau_i \in \mathcal{U}$ according to $\mathbb{P}$
    $\mathcal{U} \leftarrow \mathcal{U} \setminus \{\tau_i\}$ ▷ Remove τ_i from the urn
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{\tau_i\}$ ▷ Add τ_i to the selected indices
  end for
  return $\mathcal{T}$
end procedure

as shown in Section 2.3, a non-uniform randomization can improve the performance of the CORSING or even become crucial to get a reliable solution in particular cases.[3]

In the full-PG case no CORSING occurs and we solve the square system (2.13) for $m = N$, via the the MATLAB® \ (backslash) command.

Remark 2.2.1. (Practical usage of OMP-BOX and SPGL1) The employed MATLAB® packages OMP-BOX and SPGL1 actually solve the minimization problems

$$(\mathrm{P}_q^\varepsilon) \quad \min_{\mathbf{u} \in \mathbb{C}^N} \|\mathbf{u}\|_q \quad \text{s.t. } \|\mathbf{A}\mathbf{u} - \mathbf{f}\|_2 \leq \varepsilon, \quad q = 0, 1$$

instead of (P_0) and (P_1), where $\varepsilon > 0$ is a given tolerance. This is due to the fact that both OMP-BOX and SPGL1 are iterative solvers, and it is numerically impossible requiring $\|\mathbf{A}\mathbf{u} - \mathbf{f}\|_2 = 0$ as a stopping criterion. In every experiment performed in this chapter, the tolerance on the ℓ_2-norm of the residual is set to $\varepsilon =$ 1e-08. Moreover, we always normalize the columns of $\mathbf{A}$ with respect to the ℓ_2-norm before employing the solvers. After the solution of the normalized system is computed, we apply the substitution $u_j \mapsto u_j / \|\mathbf{a}_j\|_2$ for every $j = 1, \dots, N$, where $\mathbf{a}_j$ is the j-th column of $\mathbf{A}$. □

2.2.2 Assembling the stiffness matrix

Usually, building the stiffness matrix is the most costly operation, since the stiffness matrices used in the CORSING approach are in general full. Hence, dealing with the numerical quadratures could be, in general, a challenging issue. This inconvenience could be overcome in several possible ways.

[3] To our knowledge, Algorithm 2.1, despite being already known in the literature, is not given an identifying name. The officially acknowledged most similar algorithm that we have found so far is the Independent Chip Model (ICM), used in poker tournaments. See also the MathOverflow question `http://mathoverflow.net/questions/160738`.

Symbolic approach Compute symbolically $a(\psi_j, \varphi_q)$, and then just evaluate it for $j \in [N]$ and $q = \tau_i$, with $i \in [m]$. Of course, the integrals needed to obtain these expressions are not always explicitly computable. Nevertheless, this strategy is very effective, e.g., for equations with constant coefficients or with analytically defined coefficients.

Pre-computing Pre-compute the elements of the *whole* stiffness matrix $\mathbf{B} \in \mathbb{R}^{N\times N}$, defined as $B_{qj} := a(\psi_j, \varphi_q)$, $\forall j \in [N], \forall q \in [N]$, and then, after the test selection, extract the rows of $\mathbf{B}$ corresponding to the indices $\tau_1, \dots, \tau_m$ and plug them in $\mathbf{A}$. This approach requires $\mathcal{O}(N^2)$ memory, which is typically a large amount.

Fast transforms If the trial basis possesses some remarkable structure, as it does in the examples presented in this chapter, the numerical integration process can be implemented using fast transforms, such as FFT, DST, DCT or the FWT (see, e.g., [Mal99]). Using this approach, the computational cost of the assembly can be considerably reduced, or even avoided, implementing a *matrix-free* version of the algorithm.

Interpolation Another possibility is to perform an interpolation of the test functions φ_q with respect to the basis $\{\psi_j\}_{j\in[N]}$, namely

$$\varphi_q \approx \sum_{k=1}^{N} \phi_{q,k}\psi_k, \quad \forall q \in [N].$$

Then, build the stiffness matrix $\mathbf{A}^\psi \in \mathbb{R}^{N\times N}$ associated with the Galerkin discretization performed with respect to the trial functions, defined as

$$A^\psi_{jk} := a(\psi_k, \psi_j), \quad \forall j, k \in [N].$$

Then, each row of $\mathbf{A}$ can be built as a suitable linear combination of rows of $\mathbf{A}^\psi$. Again, this would require a significant amount of memory.

As a general remark, the approximation error introduced by the last two approaches should be estimated on theoretical grounds (see Remark 2.1.1).

In particular, in this thesis we employ the symbolic approach.

2.3 CORSING in action

The choice of the trial and test functions in the assembly phase may be quite arbitrary, in principle, except for ensuring the well-posedness of (2.6). In particular, we select hat and sine functions. We explore their role as both trials

and tests. These two bases have been chosen in order to fulfill two main requirements. On the one hand, they both are able to capture sparsity in $H_0^1(\Omega)$, this being the function space associated with the weak formulation of problem (2.2). On the other hand, and this is a more heuristic motivation, they belong to qualitatively different "worlds", namely, the hat functions are sparse in the *spatial* domain, the sines are sparse in the *frequency* domain. This duality between space and frequency is a key concept to CS, and it has been widely used since its discovery [CRT06].

The CORSING setting poses more constraints with respect to the standard CS. First, CS is usually cast in the finite dimensional space $\mathbb{C}^N$, whereas CORSING relies on the infinite dimensional function space $H_0^1(\Omega)$. About infinite dimensional spaces, CS has been recently extended to the sampling of a continuous signal that is sparse in a wavelet basis, by resorting to few random Fourier measurements [AHPR13, AH15]. In these works, the sampling problem takes place in a Hilbert space. Second, in the CORSING case, a generic bilinear form $a(\cdot,\cdot)$, not necessarily symmetric, replaces the Euclidean inner product and the boundary conditions have to be included in the sampling problem. The choice of the hat and sine functions matches these requirements as shown below.

We start with the simple 1D case, by choosing $\Omega = (0,1)$.

Hat functions The first basis, corresponding to the spatial domain, is the hierarchical multiscale basis over the interval $[0,1]$, consisting of the mother hat function

$$\mathcal{H}(x) = \begin{cases} x & \text{if } 0 \le x < \frac{1}{2} \\ 1-x & \text{if } \frac{1}{2} \le x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

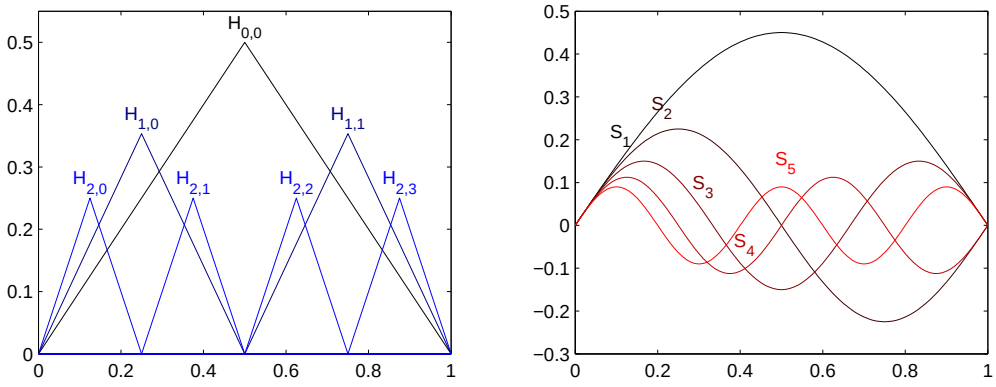
and of its scaled dyadic translations $\mathcal{H}_{\ell,k}(x) = 2^{-\ell/2}\,\mathcal{H}(2^\ell x - k)$, defined for $\ell \in \mathbb{N}$ and $k = 0,\dots,2^\ell - 1$ [Dah97]. The normalization constant guarantees all these functions to have a unit $H^1(\Omega)$-seminorm. Moreover, $\mathcal{H}_{\ell,k}$ is locally supported on the interval $(k2^{-\ell},(k+1)2^{-\ell})$. We denote the hierarchical basis of level $L \ge 0$ with

$$\mathcal{H}^L = \{\mathcal{H}_{\ell,k} : 0 \le \ell \le L\}. \tag{2.15}$$

It can be checked that

$$\operatorname{span}(\mathcal{H}^L) \equiv \left\{u \in X^1_{2^{-(L+1)}} : u(0) = u(1) = 0\right\}, \tag{2.16}$$

with $\dim(\operatorname{span}(\mathcal{H}^L)) = 2^{L+1} - 1$, and where X_h^1 is the space of continuous piecewise linear functions over the grid of uniform step h on the interval $[0,1]$. The basis $\mathcal{H}^3$ is shown in Figure 2.1, (a).



(a) (b)

Figure 2.1: The basis $\mathcal{H}^3$, (a); the basis $\mathcal{S}^5$, (b).

Remark 2.3.1. The first order derivative (in a weak sense) of $\mathcal{H}$ is the Heaviside step function

$$\psi^{\mathrm{H}}(x) = \begin{cases} 1 & \text{if } 0 \le x < \frac{1}{2} \\ -1 & \text{if } \frac{1}{2} \le x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, the functions $\psi^{\mathrm{H}}_{\ell,k}(x) = \mathcal{H}'_{\ell,k}(x) = 2^{\ell/2}\psi^{\mathrm{H}}(2^{\ell}x - k)$, together with the constant function identically equal to 1, form the well known *Haar wavelet basis* of $L^2(\Omega)$. This property will be useful when approximating the one-dimensional Poisson problem. □

As mentioned before, the choice of the trial and test bases is a key issue in CORSING, and it will be investigated from a theoretical viewpoint in Chapter 3, using the notion of local a-coherence.

Remark 2.3.2 (Classical FE bases). It is not convenient to employ a classical piecewise polynomial FE basis [QV08] as trials, since it is not able to provide a sparse representation of the exact solution u except when it has a localized support. On the contrary, the hierarchical structure of $\mathcal{H}^L$ makes it perfectly able to sparsify u even though the support of u is the whole Ω. □

Sine functions The second basis, associated with the frequency domain, is given by

$$\mathcal{S}_r(x) = \frac{\sqrt{2}}{\pi r}\sin(\pi r x), \quad x \in [0,1],$$

for $r \in \mathbb{N} \setminus \{0\}$, where the normalization constant ensures that $\mathcal{S}_r$ has a unit $H^1(\Omega)$-seminorm. The sine basis of dimension R, for some integer $R \ge 1$, is

denoted by

$$\mathcal{S}^R = \{\mathcal{S}_r : 1 \le r \le R\}. \tag{2.17}$$

We plot the basis $\mathcal{S}^5$ in Figure 2.1, (b).

Remark 2.3.3. The set $\mathcal{S}^\infty$ is a complete orthonormal system of $H_0^1(0,1)$ with respect to the inner product $\int_0^1 u'v' \,\mathrm{d}x$. This property can be proved by exploiting the completeness of $\{e^{i\pi r x}\}_{r\in\mathbb{Z}}$ in $L^2(-1,1)$ and using the odd extension from $[0,1]$ to $[-1,1]$. □

Remark 2.3.4. Boundary conditions different from Dirichlet's lead to a different choice of the trial and test functions. □

Now, we assess the performance of CORSING applied to some simple one-dimensional problems. After the discussion of these basic examples, we will deal with more challenging test cases.

2.3.1 The 1D Poisson problem

In order to show the reliability and the robustness of CORSING, we consider both the constrained minimization problems (P_0) and (P_1). In particular, we resort to the MATLAB® packages OMP-BOX [Rub09, RZE08] and SPGL1 [vdBF08, vdBF07], respectively.[4]

We focus on the 1D Poisson problem

$$\begin{cases} -u'' = f \quad \text{in } \Omega = (0,1) \\ u(0) = u(1) = 0, \end{cases} \tag{2.18}$$

where the source term f is chosen such that the exact solution be

$$u_t^*(x) = (e^{tx} - 1)(1 - x), \quad \forall x \in \overline{\Omega},$$

with t a positive parameter. The shape of u_t^* can be tuned by varying t, i.e., u_t^* exhibits a thinner and thinner boundary layer as t increases (see Fig. 2.2).

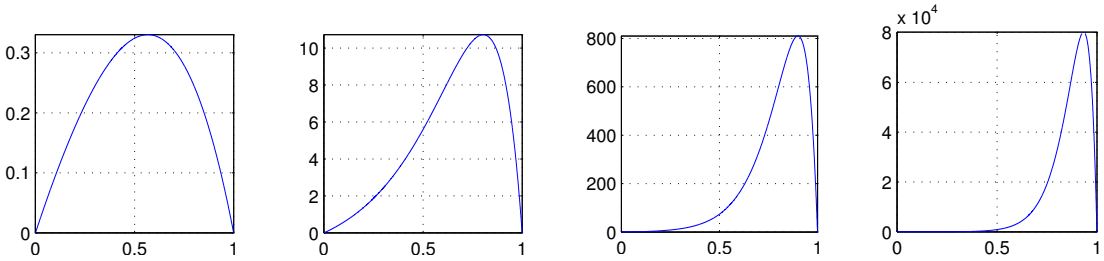
Before dealing with the numerical validation of CORSING, we provide the asymptotic best approximation error estimates for u_t^* in the spaces spanned by the bases $\mathcal{H}^L$ and $\mathcal{S}^N$.

Proposition 2.4. *Given $t > 0$, there exist two constants $C, D > 0$ such that*

$$\inf_{v\in\mathrm{span}(\mathcal{H}^L)} \|v - u_t^*\|_{L^2(\Omega)} \le CN^{-2}, \qquad \inf_{v\in\mathrm{span}(\mathcal{S}^N)} \|v - u_t^*\|_{L^2(\Omega)} \le DN^{-2.5},$$

$$\inf_{v\in\mathrm{span}(\mathcal{H}^L)} |v - u_t^*|_{H^1(\Omega)} \le CN^{-1}, \qquad \inf_{v\in\mathrm{span}(\mathcal{S}^N)} |v - u_t^*|_{H^1(\Omega)} \le DN^{-1.5},$$

[4] All the experiments have been performed using MATLAB® R2013a 64-bit (version 8.1.0.604) on a MacBook Pro equipped with a 3GHz Intel Core i7 processor and 8GB of RAM.



(a) u_1^* (b) u_5^* (c) u_{10}^* (d) u_{15}^*

Figure 2.2: The function u_t^* for different values of t.

with $N = 2^{L+1} - 1$, *and with* $\mathcal{H}^L$ *and* $\mathcal{S}^N$ *the bases defined in* (2.15) *and* (2.17), *respectively. Moreover, the constants* C *and* D *depend only on* t, *and exhibit the following asymptotic behavior*

$$C \sim \sqrt{t}e^t, \quad D \sim te^t. \tag{2.19}$$

Proof. Let us start with the estimates related to $\text{span}(\mathcal{H}^L)$. Due to (2.16), we exploit the interpolation error estimate in [Qua14, Theorem 4.2] to get, for $k = 0, 1$,

$$|\Pi_N^1 v - v|_{H^k(\Omega)} \lesssim |v|_{H^2(\Omega)} N^{-2+k}, \quad \forall v \in H^2(\Omega),$$

with $\Pi_N^1 : H^2(\Omega) \to X_{1/(N+1)}^1$ the standard piecewise linear Lagrange interpolation operator, $H^0(\Omega) = L^2(\Omega)$, and where it is understood that $\lesssim$ hides a constant independent of N. We observe that the interpolation error associated with $v = u_t^*$ grows as t increases. Indeed we have

$$|u_t^*|_{H^2(\Omega)} = \sqrt{\tfrac{5}{2}te^{2t} - t(t^2 - 3t + \tfrac{5}{2})}$$

and, consequently, for $N \to +\infty$, there exists a constant $C > 0$ such that

$$\inf_{v \in \text{span}(\mathcal{H}^L)} |v - u_t^*|_{H^k(\Omega)} \leq |\Pi_N^1 u_t^* - u_t^*|_{H^k(\Omega)} \leq CN^{-2+k}, \quad \text{for } k = 0, 1,$$

where C behaves like $|u_t^*|_{H^2(\Omega)} \sim \sqrt{t}e^t$, for $t \to +\infty$.

Concerning the approximation in the space $\mathcal{S}^N$, we notice that the family $\{\pi k \mathcal{S}_k\}_{k \in \mathbb{N}}$ is an orthonormal complete system of $L^2(\Omega)$ with respect to the $L^2(\Omega)$-scalar product. Employing Parseval's identity, the squared $L^2(\Omega)$-norm

of the best approximation error in $\mathcal{S}^N$ associated with u_t^* is

$$\left\| u_t^* - \sum_{k=1}^{N} (u_t^*, \pi k \mathcal{S}_k) \pi k \mathcal{S}_k \right\|_{L^2(\Omega)}^2 = \left\| \sum_{k=N+1}^{+\infty} (u_t^*, \pi k \mathcal{S}_k) \pi k \mathcal{S}_k \right\|_{L^2(\Omega)}^2 = \sum_{k=N+1}^{+\infty} |(u_t^*, \pi k \mathcal{S}_k)|^2.$$

In order to estimate this series, it can be checked, via a symbolic computation, that the k-th Fourier coefficient is

$$|(u_t^*, \pi k \mathcal{S}_k)|^2 = \frac{2\, t^2 \left(t^3 - 2\,\pi^2 k^2 + \pi^2 t\, k^2 + 2\,(-1)^k \pi^2 k^2 \mathrm{e}^t\right)^2}{\pi^2 k^2 \left(t^2 + \pi^2 k^2\right)^4},$$

i.e., for $k \to +\infty$, there exists $\widetilde{D} > 0$ such that $|(u_t^*, \pi k \mathcal{S}_k)|^2 \le \widetilde{D} k^{-6}$. Thanks to the monotonicity of the function x^{-6} for $x > 0$, the series with generic term k^{-6} can be bounded from above by

$$\sum_{k=N+1}^{+\infty} k^{-6} \le \int_N^{+\infty} x^{-6}\, \mathrm{d}x \lesssim N^{-5}.$$

Finally, we have

$$\inf_{v \in \mathrm{span}(\mathcal{S}^N)} \|v - u_t^*\|_{L^2(\Omega)} \le D N^{-2.5},$$

where the constant D is asymptotic to te^t, for $t \to +\infty$.

In order to estimate the best approximation error in the $H^1(\Omega)$-seminorm, we observe that since $u_t^* \in \mathcal{C}^\infty(\Omega) \cap H_0^1(\Omega)$, its odd extension $u_t^{*,\mathrm{odd}}$ to $[-1,1]$ belongs to $\mathcal{C}^1_{\mathrm{p},2}\big|_{[-1,1)}$, where, for a generic $k \in \mathbb{N}$, we define

$$\mathcal{C}^k_{\mathrm{p},2} = \{g \in \mathcal{C}^k(\mathbb{R}) \text{ and } g(x+2) = g(x) \quad \forall x \in \mathbb{R}\}. \tag{2.20}$$

Moreover, its second derivative is (at least) absolutely integrable. Consequently, the Fourier series of $(u_t^{*,\mathrm{odd}})'$, that is obtained as the even extension of $(u_t^*)'$ to $[-1,1]$, is convergent in $L^2(-1,1)$ and coincides with the term-by-term derivative of the Fourier series of the even extension to $[-1,1]$ of $(u_t^*)'$ (see [Tol12, Section 5.8, Theorem 2]). Thus, we have

$$\inf_{v \in \mathrm{span}(\mathcal{S}^N)} |v - u_t^*|_{H^1(\Omega)}^2 = \left\| (u_t^*)' - \sum_{k=1}^{N} (u_t^*, \pi k \mathcal{S}_k) \pi k \mathcal{S}_k' \right\|_{L^2(\Omega)}^2 = \left\| \sum_{k=N+1}^{+\infty} (u_t^*, \pi k \mathcal{S}_k) \pi k \mathcal{S}_k' \right\|_{L^2(\Omega)}^2.$$

Using the orthonormality and completeness of the system $\{\sqrt{2}\cos(k\pi x)\}_{k\geq 0}$ in $L^2(\Omega)$ and with computations similar to the $L^2(\Omega)$-analysis, we obtain

$$\inf_{v\in \mathrm{span}(\mathcal{S}^N)} |v - u_t^*|_{H^1(\Omega)} \leq DN^{-1.5}.$$

□

Remark 2.3.5. An alternative (and more general) proof of the estimates in the space $\mathcal{S}^N$ can be based on the decay of the Fourier coefficients with respect to their frequency index. Actually, it can be proved that, for any piecewise periodic $\mathcal{C}^1_{\mathrm{p},2}$-function, say g, with at most a finite number of jumps on $[-1,1)$, the corresponding Fourier coefficients, $\widehat{g_k}$, satisfy $|\widehat{g_k}| \lesssim 1/k$.[5] Since $(u_t^{*,\mathrm{odd}})''$ enjoys the properties of g, the Fourier coefficients of $u_t^{*,\mathrm{odd}}$ behave like $1/k^3$. This in turn yields that the $L^2(-1,1)$-norm of the approximation error due to truncation of the Fourier series of $u_t^{*,\mathrm{odd}}$ is $\lesssim N^{-2.5}$.

Alternatively one could rely on [CQ82, Theorem 1.1] involving the Sobolev spaces $H^s(-1,1)$ of non-integer order. Nevertheless, $u_t^{*,\mathrm{odd}} \in H^s(-1,1)$ for $0 \leq s < 2.5$. Thus we would obtain suboptimal convergence rates, in comparison to the ones in Proposition 2.4.

These two alternative arguments allow the generalization of Proposition 2.4 to functions different from u_t^* when $\mathcal{S}^N$ is the trial basis. □

After providing the best-approximation results, we now consider the full-PG formulation in the spaces $\mathrm{span}(\mathcal{S}^N)$ and $\mathrm{span}(\mathcal{H}^L)$, with $N = 2^{L+1} - 1$, and we prove that it is well-posed. To show this, we employ the *inf-sup* condition (see Section 2.1). This property implies the existence and uniqueness of the solution $\widehat{u}_N^N$ of the full-PG method applied to (2.18) with the sine functions as trials and the hat functions as tests, as stated in the following

Proposition 2.5. *Let $L \in \mathbb{N}$ and define $N = 2^{L+1} - 1$. Then, the finite dimensional spaces $U^N = \mathrm{span}(\mathcal{S}^N)$ and $V^N = \mathrm{span}(\mathcal{H}^L)$ satisfy the inf-sup condition with respect to the bilinear form $a(\cdot,\cdot)$ associated with problem* (2.18)*. Namely, there exists a constant $\widetilde{\alpha} > 0$, not depending on N, such that*

$$\inf_{u\in U^N} \sup_{v\in V^N} \frac{a(u,v)}{|u|_{H^1(\Omega)}|v|_{H^1(\Omega)}} \geq \widetilde{\alpha}, \tag{2.21}$$

with $a(u,v) = \int_0^1 u'v'\,dx$ and $\widetilde{\alpha} = 2/\pi$.

Proof. We recall that $U^N = \mathrm{span}\{\psi_j : 1 \leq j \leq N\}$ and $V^N = \mathrm{span}\{\varphi_i : 1 \leq i \leq N\}$, with

$$\psi_j(x) = \sin(\pi j x), \quad \forall x \in [0,1]$$

[5] See Paul Garrett's answer to the MathOverflow question `http://mathoverflow.net/questions/182684`.

and

$$\varphi_i(x) = \begin{cases} (x - x_{i-1})/h & \text{if } x \in [x_{i-1}, x_i) \\ (x_{i+1} - x)/h & \text{if } x \in [x_i, x_{i+1}] \\ 0 & \text{otherwise,} \end{cases}$$

where $h = 1/(N+1)$ and $x_i = ih$. First, we notice that both $\{\psi_j\}$ and $\{\varphi_i\}$ are not normalized since $\widetilde{\alpha}$ in (2.21) is independent of any scaling of both u and v. Moreover, condition (2.21) is equivalent to the algebraic condition

$$\forall \mathbf{u} \in \mathbb{R}^N \setminus \{\mathbf{0}\},\ \exists \mathbf{v} \in \mathbb{R}^N \setminus \{\mathbf{0}\} \text{ s.t.} \quad \mathbf{v}^\intercal \mathbf{A}\mathbf{u} \geq \widetilde{\alpha} (\mathbf{u}^\intercal \mathbf{D}\mathbf{u})^{\frac{1}{2}} (\mathbf{v}^\intercal \mathbf{T}\mathbf{v})^{\frac{1}{2}}, \tag{2.22}$$

where $u = \sum_{j=1}^{N} u_j \psi_j$, $v = \sum_{i=1}^{N} v_i \varphi_i$, $D_{ij} = \delta_{ij}(\pi j)^2/2$, with δ_{ij} the Kronecker symbol, $\mathbf{A}$ is the stiffness matrix, i.e., $A_{ij} = a(\psi_j, \varphi_i)$, and

$$T_{ij} = \begin{cases} 2/h & \text{if } i = j \\ -1/h & \text{if } |i-j| = 1 \\ 0 & \text{otherwise.} \end{cases}$$

For every $\mathbf{u} \in \mathbb{R}^N \setminus \{\mathbf{0}\}$, we show that the ansatz

$$\mathbf{v} = \mathbf{S}\mathbf{u}, \tag{2.23}$$

where $S_{ij} = \sin(ij\pi h)$, represents the good candidate for satisfying the inf-sup condition. Observe that it holds

$$\mathbf{A} = \mathbf{T}\mathbf{S}. \tag{2.24}$$

Indeed, we have that

$$\begin{aligned} A_{ij} = a(\psi_j, \varphi_i) &= \frac{1}{h}(-\sin(\pi j x_{i-1}) + 2\sin(\pi j x_i) - \sin(\pi j x_{i+1})) \\ &= \frac{1}{h}(-S_{i-1,j} + 2S_{ij} - S_{i+1,j}) = T_{i,i-1}S_{i-1,j} + T_{i,i}S_{i,j} + T_{i,i+1}S_{i+1,j} \\ &= [\mathbf{T}\mathbf{S}]_{ij}, \end{aligned}$$

where it is understood that these equalities formally hold also for $i, j \in \{1, N\}$, by letting $S_{0,j} = S_{N+1,j} = 0$ for $j = 1, \ldots, N$.

Now, employing (2.23) and (2.24), (2.22) can be equivalently written as

$$\mathbf{u}^\intercal (\mathbf{S}^\intercal \mathbf{T}\mathbf{S})\mathbf{u} \geq \widetilde{\alpha}^2 \mathbf{u}^\intercal \mathbf{D}\mathbf{u}, \quad \forall \mathbf{u} \in \mathbb{R}^N \setminus \{0\}. \tag{2.25}$$

In order to determine $\widetilde{\alpha} > 0$, we first exploit the symmetry of $\mathbf{S}$ and the property that the columns of $\mathbf{S}$ form a basis of eigenvectors of the matrix $\mathbf{T}$, to show

that the matrix $\widetilde{\mathbf{D}} = \mathbf{S}^\intercal\mathbf{T}\mathbf{S} = \mathbf{S}\mathbf{T}\mathbf{S}$ on the left-hand side of (2.25) is diagonal. In particular, $\widetilde{D}_{ij} = \delta_{ij}\, 2/h^2 \sin^2(h\pi j/2)$. Then, we consider the minimization problem

$$\widetilde{\alpha}^2 = \min_{\mathbf{u}\in\mathbb{R}^N\setminus\{\mathbf{0}\}} \frac{\mathbf{u}^\intercal\widetilde{\mathbf{D}}\mathbf{u}}{\mathbf{u}^\intercal\mathbf{D}\mathbf{u}}.$$

This is equivalent to finding the minimum generalized eigenvalue associated with the matrix pencil $\widetilde{\mathbf{D}} - \lambda\mathbf{D}$, i.e.,

$$\begin{aligned}\widetilde{\alpha}^2 &\equiv \min\{\lambda\in\mathbb{R} : \det(\widetilde{\mathbf{D}} - \lambda\mathbf{D}) = 0\} = \min_{j\in[N]} \frac{\widetilde{D}_{jj}}{D_{jj}} \\ &= \min_{j\in[N]} \left[\sin\left(\frac{h\pi j}{2}\right)/\left(\frac{h\pi j}{2}\right)\right]^2 = \left[\sin\left(\frac{N}{(N+1)}\frac{\pi}{2}\right)/\left(\frac{N}{(N+1)}\frac{\pi}{2}\right)\right]^2 \geq \frac{4}{\pi^2},\end{aligned}$$

where the last inequality follows from the observation that $g(t) = \sin(t)/t$, for $t\in(0,\pi/2)$ can be bounded from below by $g(\pi/2) = 2/\pi$. □

According to Theorem 2.3, an immediate consequence of Proposition 2.5 is the following corollary. Notice that hypothesis (2.8) of Theorem 2.3 holds, since the stiffness matrix $\mathbf{A}$ in (2.24) is nonsingular.

Corollary 2.6. *In the same framework as in Proposition 2.5, the following estimate holds*

$$|u - \widehat{u}_N^N|_{H^1(\Omega)} \leq \left(1 + \frac{\beta}{\widetilde{\alpha}}\right) \inf_{w\in U^N} |u - w|_{H^1(\Omega)},$$

where β is the continuity constant of $a(\cdot,\cdot)$ with respect to the $H^1(\Omega)$-seminorm of the arguments and $\widetilde{\alpha}$ is defined as in (2.21).

Analogous results hold by swapping the trial and test spaces.

Proposition 2.7. *The inf-sup condition in Proposition 2.5 holds also for $U^N = \mathrm{span}(\mathcal{H}^L)$ and $V^N = \mathrm{span}(\mathcal{S}^N)$, with $N = 2^{L+1} - 1$, with the same value of $\widetilde{\alpha}$.*

Proof. The proof of Proposition 2.5 can be mimicked, working on $\mathbf{A}^\intercal$. □

The same statement as in Corollary 2.6 holds for $U^N = \mathrm{span}(\mathcal{H}^L)$ and $V^N = \mathrm{span}(\mathcal{S}^N)$, with $N = 2^{L+1} - 1$, with the same value of $\widetilde{\alpha}$.

Employing an argument similar to Theorem 2.3, we can prove a recovery result for the CORSING procedure, that generalizes Proposition 1.5 to the function space setting. First, we define the space of s-sparse functions in U^N

$$U_s^N := \left\{ w\in U^N : \exists\mathbf{w}\in\mathbb{R}^N \text{ s.t. } w = \sum_{j=1}^N w_j\psi_j \text{ and } \|\mathbf{w}\|_0 \leq s \right\}.$$

Proposition 2.8. *Let $s, m, N \in \mathbb{N}$ such that $0 < 2s \leq m \leq N$. Suppose that the corsed-PG solution fulfills the exact constraint $\mathbf{A}\widehat{\mathbf{u}} = \mathbf{f}$ and that $\|\widehat{\mathbf{u}}\|_0 \leq s$. Then, if the following $2s$-sparse inf-sup condition is fulfilled*

$$\inf_{u \in U_{2s}^N} \sup_{v \in V^m} \frac{a(u,v)}{|u|_{H^1(\Omega)}|v|_{H^1(\Omega)}} \geq \widetilde{\alpha}, \tag{2.26}$$

for some positive constant $\widetilde{\alpha} > 0$, the following estimate holds

$$|u - \widehat{u}|_{H^1(\Omega)} \leq \left(1 + \frac{\beta}{\widetilde{\alpha}}\right) \inf_{w \in U_s^N} |u - w|_{H^1(\Omega)},$$

where β is the continuity constant of $a(\cdot,\cdot)$ with respect to the $H^1(\Omega)$-seminorm of the arguments.

Proof. Fix $w \in U_s^N$. Then, we have $\widehat{u} - w \in U_{2s}^N$. Hence, applying (2.26) and exploiting the Galerkin orthogonality

$$a(\widehat{u} - u, v) = 0, \quad \forall v \in V^m$$

due to (2.1) and (2.6), there exists $v \in V^m$ such that

$$|\widehat{u} - w|_{H^1(\Omega)} \leq \frac{1}{\widetilde{\alpha}} \frac{a(\widehat{u} - w, v)}{|v|_{H^1(\Omega)}} = \frac{1}{\widetilde{\alpha}} \frac{a(u - w, v)}{|v|_{H^1(\Omega)}} \leq \frac{\beta}{\widetilde{\alpha}} |u - w|_{H^1(\Omega)},$$

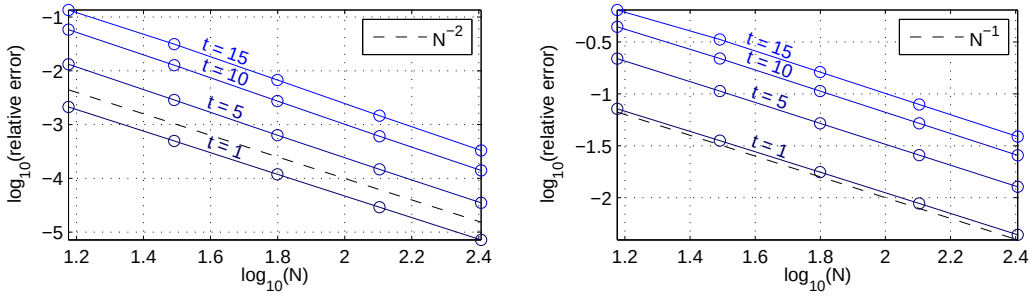
after exploiting the continuity of $a(\cdot,\cdot)$. Finally, the triangle inequality implies

$$|u - \widehat{u}|_{H^1(\Omega)} \leq |u - w|_{H^1(\Omega)} + |\widehat{u} - w|_{H^1(\Omega)} \leq \left(1 + \frac{\beta}{\widetilde{\alpha}}\right) |u - w|_{H^1(\Omega)}$$

so that the thesis follows from the arbitrariness of w. □

Remark 2.3.6. The hypotheses of Proposition 2.8 are quite restrictive: the linear constraint $\mathbf{A}\widehat{\mathbf{u}} = \mathbf{f}$ is assumed to be exactly fulfilled and, at the same time, the sparsity of the corsed-PG solution is constrained. Consider, for example, the OMP algorithm. Although it controls the maximum sparsity level, it produces - in general - a nonzero residual. The role played by the inf-sup property in the analysis of CORSING will be thoroughly investigated in the next chapter. □

Now, we will test the CORSING procedure described in Section 2.2 using the bases $\mathcal{H}^L$ and $\mathcal{S}^N$.



(a) $L^2(\Omega)$-norm (b) $H^1(\Omega)$-seminorm

Figure 2.3: full-PG error analysis on the model problem (2.18) in the $\mathcal{HS}$ case for different values of t.

Hats *vs* sines

First, let us consider the functions in $\mathcal{H}^L$ as trials and the functions in $\mathcal{S}^R$ as tests, in short the $\mathcal{HS}$ setting. We adopt the lexicographic ordering over the set $\mathcal{H}^L$

$$\begin{array}{c|ccccccccc} (\ell,k) & (0,0) & (1,0) & (1,1) & (2,0) & (2,1) & (2,2) & (2,3) & \cdots \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \cdots \\ j & 1 & 2 & 3 & 4 & 5 & 6 & 7 & \cdots \end{array}$$

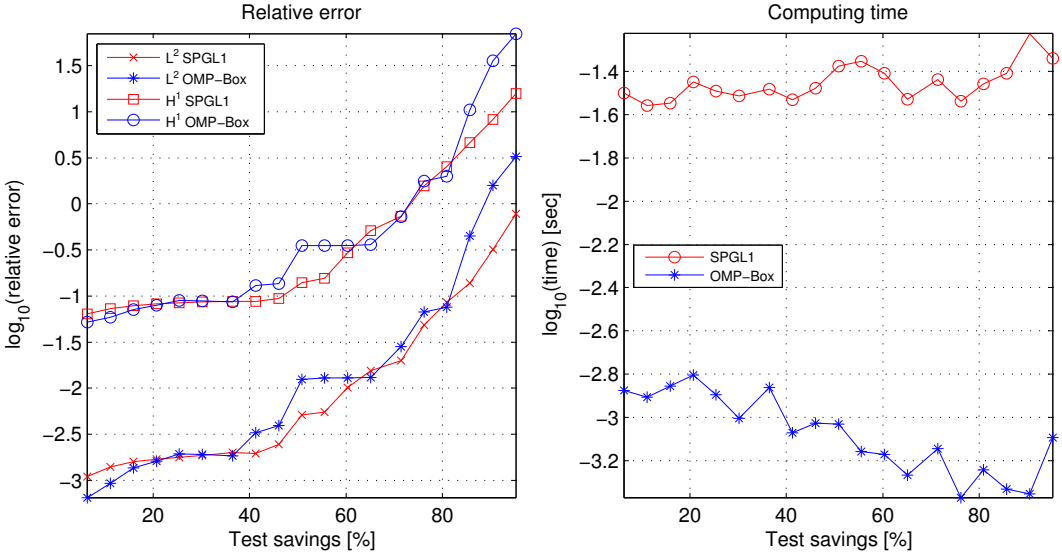
defined by the relation $j(\ell,k) = 2^\ell + k$ and with inverse mapping

$$(\ell,k)(j) = (\ell(j),k(j)) = \left(\lfloor \log_2(j) \rfloor, j - 2^{\lfloor \log_2(j) \rfloor}\right),$$

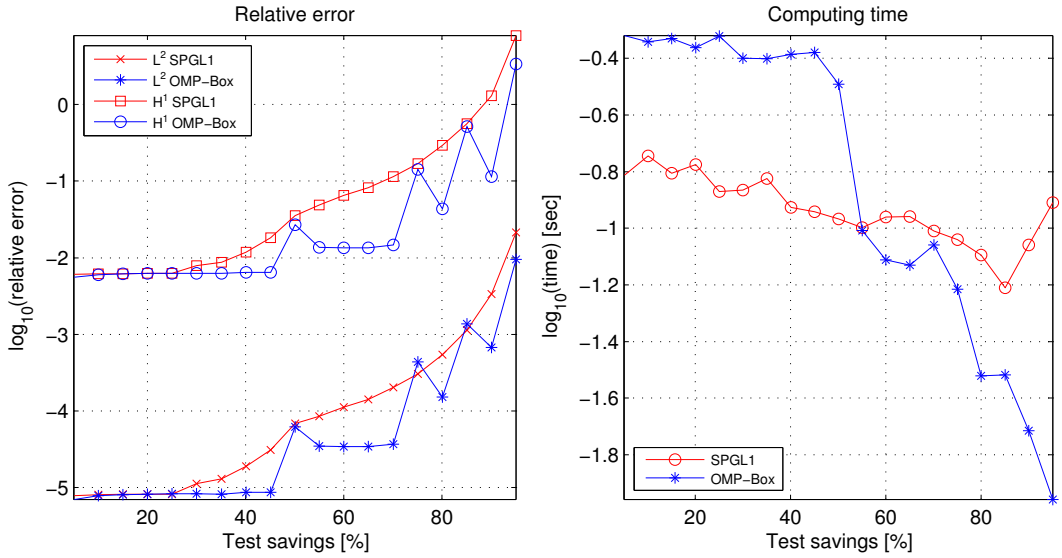
$\lfloor \cdot \rfloor$ denoting the floor function. With this convention, and according to the notation introduced in Section 2.2, the first combination of trials and tests assessed is

$$\psi_j = \mathcal{H}_{\ell(j),k(j)}, \qquad \varphi_i = \mathcal{S}_i\,.$$

Convergence of full-PG We numerically check the convergence and robustness of the full-PG method. Actually, we reach the best approximation error as stated in Proposition 2.4. We solve (2.18) for five different choices of the maximum level, i.e., $L = 3,4,5,6,7$, corresponding to the linear system (2.13) of dimension $N = 15,31,63,127,255$. In particular, this system is solved using the MATLAB® \ (backslash) command. Moreover, the four values of the parameter $t = 1,5,10,15$ are considered. For any combination of values for L and t, we show in Figure 2.3 the relative error associated with $\widehat{u}_N^N$ with respect to the $L^2(\Omega)$-norm and the $H^1(\Omega)$-seminorm. The errors follow the behavior predicted in Proposition 2.4 with respect to N. Moreover, the asymptotic constant monotonically grows as a function of t, according to (2.19).



(a) level $L = 5$ (resulting trials: $N = 63$)



(b) level $L = 9$ (resulting trials: $N = 1023$)

Figure 2.4: Numerical performance of D-CORSING in the $\mathcal{HS}$ case on the model problem (2.18) with exact solution u_5^*. Maximum level $L = 5$ (a) and $L = 9$ (b). Relative error (left) and computing times of the recovery phase (right).

Assessment of D-CORSING We check the performance of D-CORSING on the model problem (2.18) with exact solution u_5^*. In order to quantify the compression level of the discretized model, we define a new index, i.e., the *Test Savings*

$$\mathrm{TS} = 100\,\frac{N-m}{N}\,\%.$$

With a view to a computationally efficient approximation of (2.18), a large value of TS is of course desirable. We carry out two numerical experiments, setting the maximum level L to 5 in the first case, and to 9 in the second case. In both cases, the maximum level L (and, consequently, the number of trials equal to $N = 63$ and $N = 1023$, respectively) is kept fixed, while the number m of tests decreases such that TS varies from 5% (low compression) to 95% (high compression). For each value of m, the underdetermined system (2.13) is solved by means of both (P_0) and (P_1), using the solvers OMP-BOX and SPGL1, respectively. The results are shown in Figure 2.4. We evaluate the relative errors associated with the corsed-PG solution $\widehat{u}_m^N$ with respect to the $L^2(\Omega)$-norm and $H^1(\Omega)$-seminorm. The relative errors of the (P_0) and (P_1) approaches are comparable. Indeed, in both cases, the error follows the trend characterizing the full-PG as TS approaches 0% (compare the left panel of Fig. 2.4 with Fig. 2.3). The loss of accuracy with respect to the full-PG error is particularly small (less than one order of magnitude) for $\mathrm{TS} \lesssim 60\%$.

The recovery computing times of OMP-BOX are lower than those required by SPGL1 for small sized problems ($L = 5$), whereas for larger problems ($L = 9$) the opposite occurs, up to a maximum TS value. This behavior is supported by further choices of L not shown in this work.

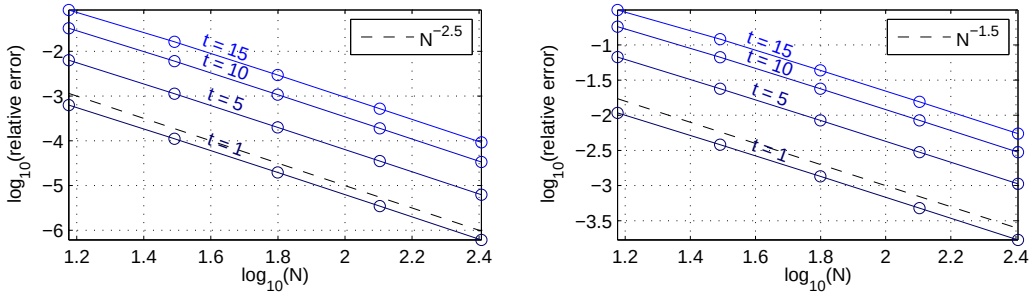
We also observe that the slope of the curve related to the OMP-BOX recovery computing time is much more emphasized in case (b). This can be explained by noticing that the computational cost associated with the OMP algorithm is $\mathcal{O}(smN)$, where $s = \|\widehat{\mathbf{u}}_m^N\|_0$ is the sparsity of the resulting corsed-PG solution (see [Ela10, Section 3.1.2]). Moreover, we have also experimentally checked that both OMP-BOX and SPGL1 furnish a corsed-PG solution with $s \approx m$ (we also refer to Figure 2.14). Henceforth, the resulting computational cost associated with OMP-BOX is approximately $\mathcal{O}(m^2N)$. On the contrary, it is less evident how to quantify from a theoretical viewpoint the computational effort demanded by the ℓ_1-minimization performed by SPGL1, even though experimentally it does not seem to be as heavily affected by m as OMP-BOX.

Sines *vs* hats

In the second set of experiments, we set

$$\psi_j = \mathcal{S}_j, \qquad \varphi_i = \mathcal{H}_{\ell(i),k(i)},$$

and we denote this framework by $\mathcal{SH}$.



(a) $L^2(\Omega)$-norm (b) $H^1(\Omega)$-seminorm

Figure 2.5: full-PG error analysis on the model problem (2.18) in the $\mathcal{SH}$ case for different values of t.

Convergence of full-PG Analogously to Section 2.3.1, the first test that we perform aims at checking the convergence of full-PG applied to the model problem (2.18) with exact solution $u_1^*, u_5^*, u_{10}^*, u_{15}^*$ and for $N = 15, 31, 63, 127, 255$. The results are shown in Figure 2.5. The theoretical results in Proposition 2.4 are confirmed, as the relative errors measured in the $L^2(\Omega)$-norm and the $H^1(\Omega)$-seminorm exhibit the expected trend and sensitivity to t.

Assessment of D-CORSING Numerical testing shows that D-CORSING is not robust in the $\mathcal{SH}$ case. This is due to the massive presence of the *aliasing* phenomenon. In Figure 2.6, we have a clear example of such an issue: the number of trials is $N = 63$, while the two values $m = 29$ and $m = 30$ are considered. In the case of OMP-BOX, surprisingly, $\widehat{u}_{29}^{63}$ is a good approximation of u_5^*, while $\widehat{u}_{30}^{63}$ is totally noisy. This shows that a sequential selection of the levels of $\mathcal{H}^L$ does not necessarily capture the high frequency components of the solution. On the contrary, if we apply R-CORSING, the quality of the corsed-PG solution highly increases for the same choices of N and m, and the aliasing phenomenon completely disappears (see Figure 2.6, right). In particular, we plot ten corsed-PG solutions corresponding to ten random experiments. The same behavior holds for the SPGL1 solver, whose performance is even worse than in the OMP-BOX case (see Figure 2.6, bottom-left).

Assessment of R-CORSING Due to the results of the last paragraph, hereafter we employ the R-CORSING for the $\mathcal{SH}$ case. The weights used in Algorithm 2.1 are

$$w_i = 2^{-\ell(i)}. \tag{2.27}$$

So far, this is an empirical choice suggested by an extensive numerical trial-and-error procedure (see Remark 2.3.7). A rigorous mathematical recipe for

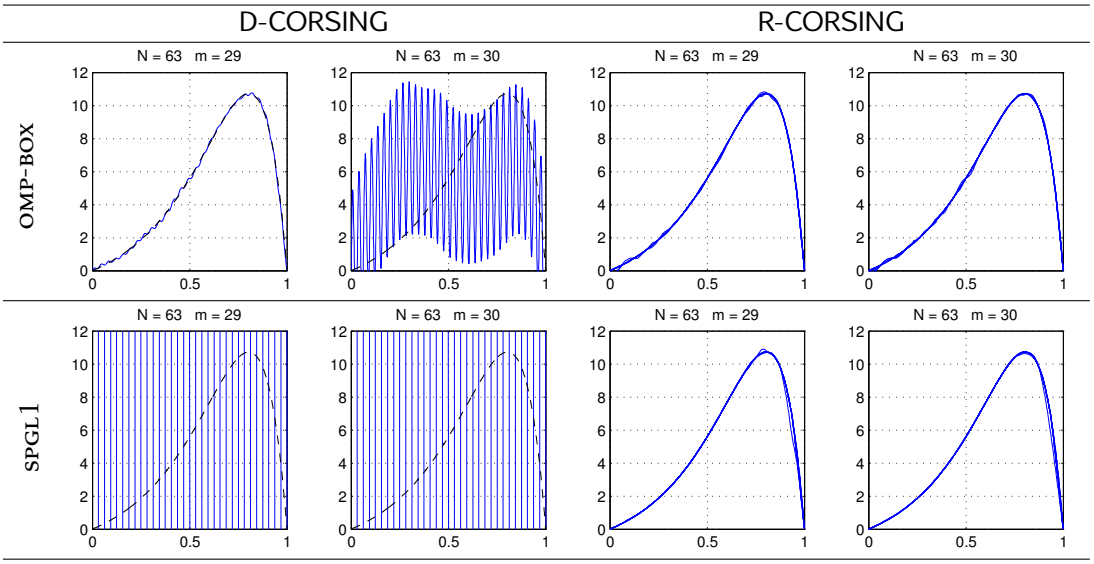


Figure 2.6: Aliasing phenomenon for D-CORSING (left); R-CORSING (right) in the $\mathcal{SH}$ case: exact solution (dashed line), corsed-PG solution (solid line).

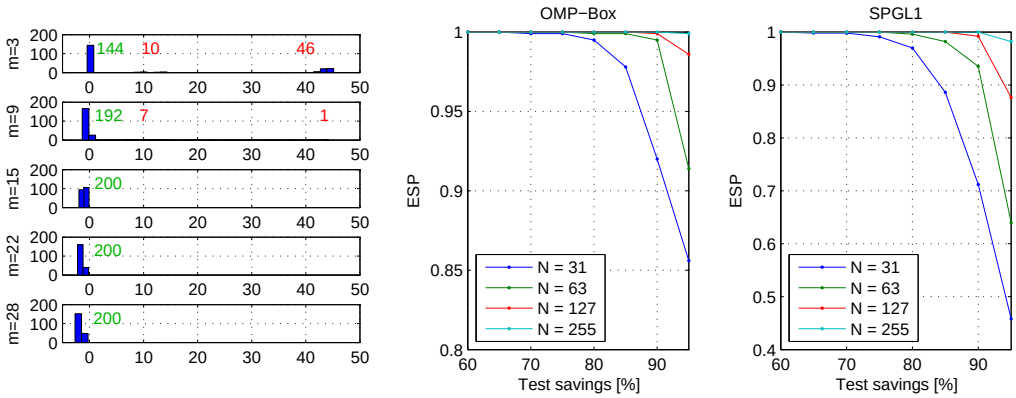
the selection of the weights will be provided in Chapter 3.

It turns out that, for large values of TS, the behavior of R-CORSING is quite chaotic. In particular, we detect the presence of a *clusterization* phenomenon. For example, if we run 200 random experiments for $N = 63$ and $m = 3, 9, 15, 22, 28$, we get the results in Figure 2.7 (a), where a histogram representing the relative error for $\widehat{u}_m^N$ with respect to the $L^2(\Omega)$-norm, for each random test, is provided. The numbers on top of the histograms indicate how many experiments are contained in the corresponding cluster. The green number is associated with successful experiments, i.e., the cluster with the lowest relative error, whereas the red numbers count the failed experiments. We notice that, for the smallest values of m, the relative error tends to cluster in separate groups. The number of these groups decreases, as expected, as m gets higher and higher, until it narrows down to one.

In order to better analyze these results, we compute the *Empirical Success Probability* index, defined as

$$\text{ESP} = \frac{\#\text{ experiments in the first cluster}}{\#\text{ experiments}}, \tag{2.28}$$

whose desirable value is 1. This quantity is plotted as a function of TS in Figure 2.7, (b). In particular, we select four values of N, i.e., $31, 63, 127, 255$, and we compare the performance of OMP-BOX, (b)-left, with SPGL1, (b)-right, for TS ranging from 60% to 95%. For each value of TS and N, 1000 random experiments are performed. The range of TS starts from 60% since we have observed



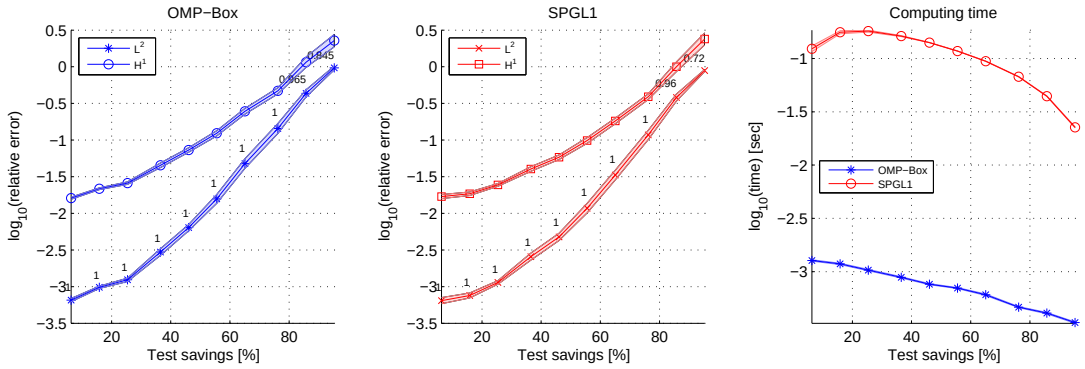
(a) Clusterization (b) ESP as a function of TS

Figure 2.7: Statistical analysis of R-CORSING in the $\mathcal{SH}$ case: histograms of the relative error with respect to the $L^2(\Omega)$-norm for $N = 63$ and for different values of m (a); sensitivity of the quantity ESP to TS for different values of N (b), and for the OMP-BOX (left) and SPGL1 (right) approach.

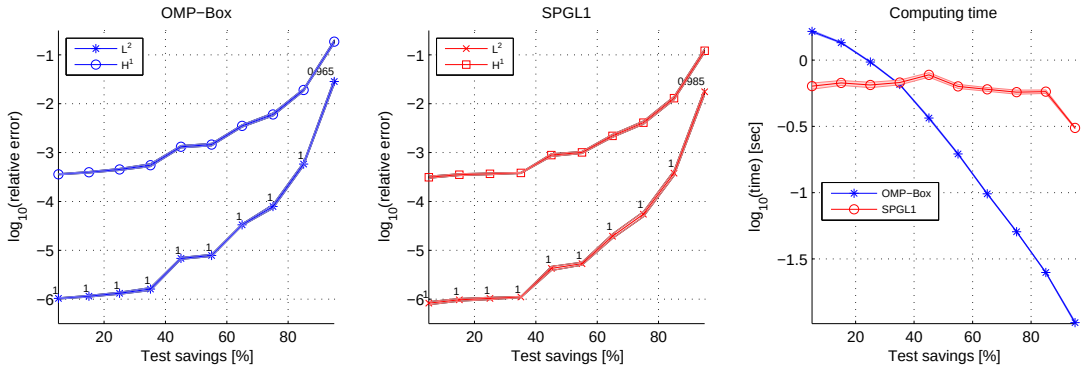
that ESP = 1 for TS $\lesssim$ 60%, for every N. For a fixed TS, ESP increases monotonically with N. This behavior is what usually occurs in the CS setting. Thus, larger values of N allow one to obtain a higher compression. Overall, OMP-BOX performs better than SPGL1. Notice that, for example, for this range of TS, ESP is always greater than 0.85 for OMP-BOX.

Now, we focus on the performance of R-CORSING on the model problem (2.18), with exact solution u_5^*, by duplicating the two experiments of the $\mathcal{HS}$ framework, fixing $N = 63$ and $N = 1023$ (see Figure 2.8, (a) and (b), respectively). The number of tests m varies such that TS ranges from 5% to 95%. For each combination of N and m, we perform 200 random experiments, using both OMP-BOX and SPGL1, and computing the relative error associated with the $L^2(\Omega)$-norm and the $H^1(\Omega)$-seminorm. The relative errors are represented as marked strips: the markers identify the mean of the errors which belong to the cluster of the successful experiments, the thickness represents the corresponding 95% confidence interval, while the numbers provide the value of ESP (they are not printed twice since the values are the same for both norms). The mean and the associated confidence interval are computed using the MATLAB® command `ttest` on the $\log_{10}$ of the data belonging to the first cluster.

From the accuracy viewpoint, OMP-BOX and SPGL1 exhibit similar performances. Concerning the values of ESP, both solvers ensure probability 1 for a large range of TS, i.e., TS $\lesssim$ 75% for $N = 63$, and TS $\lesssim$ 85% for $N = 1023$. Analogously to the $\mathcal{HS}$ setting, SPGL1 tends to be much faster than OMP-BOX for large values of N and as TS decreases (see Figure 2.8, right).

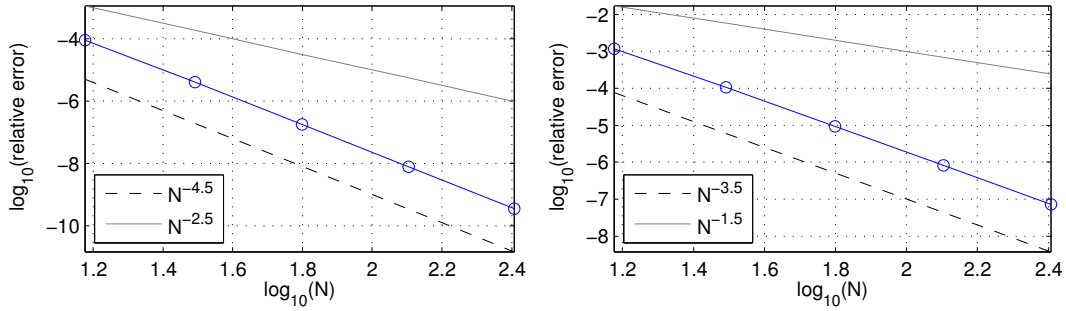


(a) $N = 63$



(b) $N = 1023$

Figure 2.8: Numerical performance of R-CORSING in the $\mathcal{SH}$ case on the model problem (2.18) with exact solution u_5^*: error strips for $N = 63$ (a), and $N = 1023$ (b), using OMP-BOX (left) and SPGL1 (center); corresponding computing times (right).



(a) $L^2(\Omega)$-norm (b) $H^1(\Omega)$-seminorm

Figure 2.9: full-PG error analysis in the $\mathcal{SH}$ case on the model problem (2.18) with exact solution w defined in (2.29): $L^2(\Omega)$-norm (a) and $H^1(\Omega)$-seminorm (b). The light lines refer to the trends predicted in Proposition 2.4.

Convergence of full-PG and R-CORSING for regular exact solutions The CORSING approach in the $\mathcal{SH}$ case turns out to be really effective when the odd extension of the solution has high regularity. Consider, for example, the function

$$w(x) = x^3(1-x)^3, \quad \forall x \in [0,1] \tag{2.29}$$

and set $W = w^{\text{odd}}$ its odd extension on $[-1,1]$. It is easy to check that $W \in \mathcal{C}^3_{\text{p},2}$, defined in (2.20). Moreover, the fourth-order derivative $W^{(4)}$ of W is infinitely differentiable except for a finite set of jumps on $[-1,1)$. Differentiating four times the Fourier series of W term-by-term, we get that the Fourier coefficients satisfy

$$|\widehat{(W^{(4)})}_k| \sim |\widehat{W}_k||k|^4.$$

Exploiting the same argument as in Remark 2.3.5 on $W^{(4)}$, we have the asymptotic decay $|\widehat{(W^{(4)})}_k| \sim |k|^{-1}$ and, henceforth, $|\widehat{W}_k| \sim |k|^{-5}$. With considerations analogous to the ones in the proof of Proposition 2.4, we obtain

$$\inf_{v \in \text{span}(\mathcal{S}^N)} \|w - v\|_{H^k(\Omega)} = \mathcal{O}(N^{-4.5+k}) \quad \text{for } k = 0, 1,$$

thus predicting a convergence rate higher than those in Proposition 2.4.

In Figure 2.9, we numerically check that full-PG ensures this best-approximation trend, by computing the relative error with respect to the $L^2(\Omega)$-norm and $H^1(\Omega)$-seminorm on problem (2.18) with exact solution w. The order of convergence is $\mathcal{O}(N^{-4.5})$ for the $L^2(\Omega)$-norm and $\mathcal{O}(N^{-3.5})$ for the $H^1(\Omega)$-seminorm (marked lines in Figure 2.9).

As expected, the regularity of W positively affects the performance of R-CORSING as well, as shown in Figure 2.10, where the same quantities as in

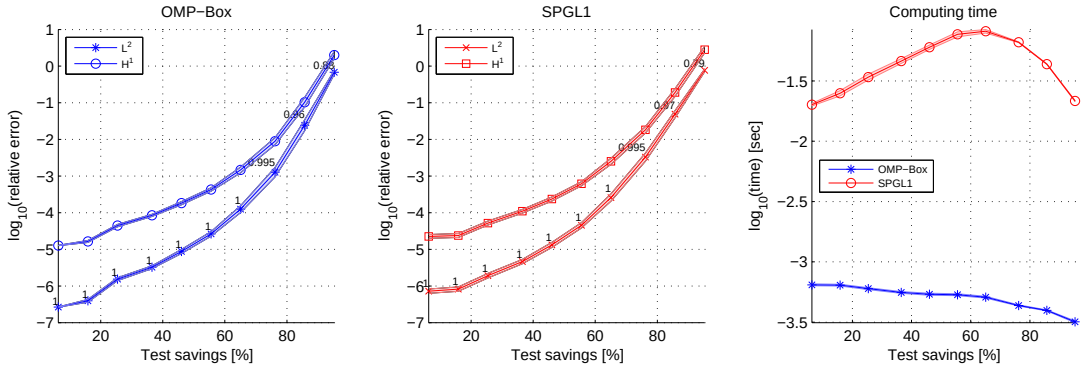


Figure 2.10: Numerical performance of R-CORSING on the model problem (2.18) with exact regular solution w defined in (2.29): error strips for $N = 63$ using OMP-BOX (left) and SPGL1 (center); corresponding computing times (right).

Figure 2.8 are shown. In particular, we choose $N = 63$, and we carry out 200 runs for each value of TS.

These results can be carried over to the more general case where w in (2.29) is replaced by

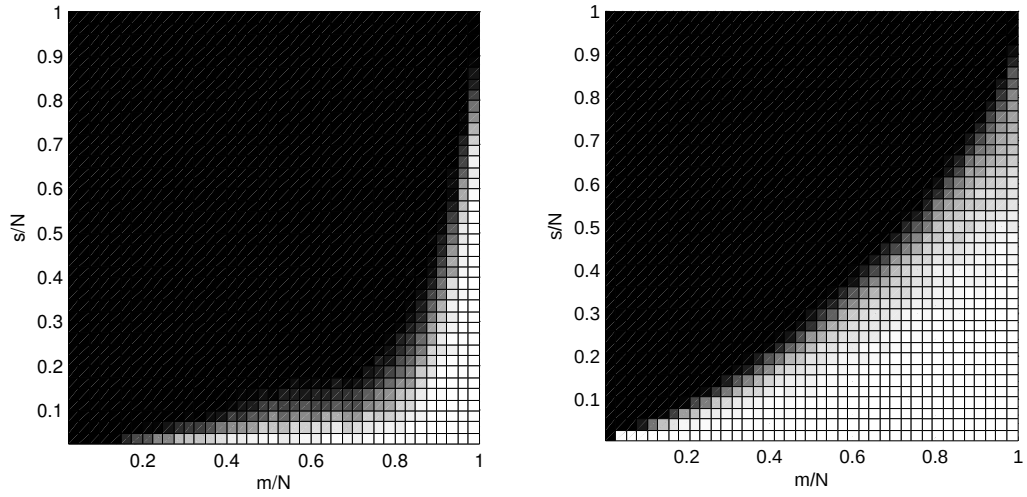
$$w(x) = x^r(1-x)^r,$$

and $W = w^{\text{odd}}$ is its odd extension, for any integer r. It can be checked that the i-th order derivative of w vanishes at $x = 0$ and $x = 1$, for $i \leq r$, whereas it is non-zero when $i > r$. Thus, $W \in \mathcal{C}^r_{\text{p},2}$ if r is odd and $W \in \mathcal{C}^{r-1}_{\text{p},2}$ if r is even. The resulting convergence rates are $\mathcal{O}(N^{-(r+1.5-k)})$ for r odd, and $\mathcal{O}(N^{-(r+0.5-k)})$ for r even, with respect to the $H^k(\Omega)$-norm, with $k = 0, 1$.

D-CORSING $\mathcal{HS}$ vs R-CORSING $\mathcal{SH}$

We now compare D-CORSING and R-CORSING, under some special conditions. We first consider the case when the solution to the differential problem is exactly sparse, i.e., it coincides with an element of the trial space, and is a linear combination of few trial functions. Although this case rarely occurs in actual situations, it is very useful for assessing a sort of consistency of CORSING. Then, we check the performance of CORSING when the solution to the differential problem is characterized by a minimal regularity, i.e., it is only in $H^1(\Omega)$. Finally, we assess the accuracy of CORSING, namely, the dependence of the approximation error on m, in comparison with the best m-term approximation error in the trial space.

CORSING robustness on sparse solutions We assess the CORSING ability to recover *sparse* solutions. For this purpose, we fix $N = 255$ and we denote by s



(a) D-CORSING $\mathcal{HS}$ (b) R-CORSING $\mathcal{SH}$

Figure 2.11: ESP as a function of m/N and s/N.

the sparsity of the solution to problem (2.18), namely, the number of trial functions involved in the definition of u. Successively, we vary independently m and s between 1 and N. For each pair, (m, s), we perform 100 runs of CORSING with an s-sparse randomly generated exact solution. In particular, the indices of the non-zero coefficients are picked via a uniform probability while the values of these coefficients follow a standard normal distribution. In the case of R-CORSING, for each pair, (m, s), the m test functions are randomly selected once for each run, according to the weights in (2.27). For both CORSING approaches, we employ only the (P_0) solver.

We expect that, for $m < s$, CORSING is hardly able to provide us with the exact solution. The number of measurements has to match at least the number of non-zero components.

CORSING robustness is assessed by computing the ESP index, shown in Figure 2.11, where the white cells are associated with the value 1, whereas black boxes correspond to the value 0.[6] Ideally, the black and white regions should be separated by the diagonal of the square with an optimal value of the white area equal to 0.5. By comparing D-CORSING with R-CORSING, we can appreciate the benefits due to the randomization of the CORSING procedure. In fact, the area of the white zone is 0.13 in Figure 2.11, (a), while it reaches the value 0.29 in

[6] In this case, the ESP definition (2.28) needs a clarification: due to the exact s-sparsity of the solution, the first cluster has been identified by means of a strong condition, namely only the corsed-PG solutions characterized by an error less than 1e-12 with respect to the H^1-seminorm have been considered correctly recovered.

Figure 2.11, (b).

Remark 2.3.7 (Tuning of the test selection weights). Now that we presented both the error analysis with respect to TS (Figure 2.8) and the assessment of the CORSING on sparse solutions (Figure 2.11), we are in a position to provide further considerations regarding the choice of the weights w_i in Algorithm 2.1. Let us briefly explain the procedure followed to derive recipe (2.27). First, we suppose the weights to follow a decay law of the form

$$w_i = 2^{-C\ell(i)}$$

for $C = 0, 0.25, \ldots, 1.75, 2$. For each value of C, we perform the same ESP analysis as in Figure 2.11, (b) on exactly sparse solutions. In particular, the corresponding areas of the white region are

C	0	0.25	0.5	0.75	1	1.25	1.5	1.75	2
Area	0.31	0.32	0.31	0.30	0.29	0.24	0.21	0.19	0.18

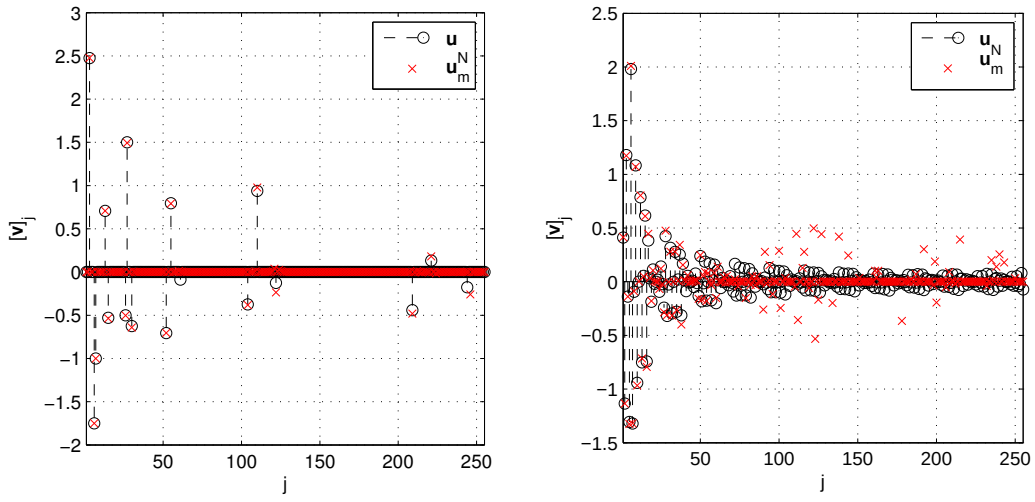
This would apparently lead us to the choice $C = 0.25$, corresponding to the maximum area. Nevertheless, if we perform an error analysis as in Figure 2.8 on u_5^* for each value of K, it turns out that the mean error associated with the corsed-PG solutions in the first cluster is lower as C approaches the value 2. Due to this controversial situation we opted for a trade-off value, i.e., $C = 1$ as in (2.27). For a rigorous investigation about the weights selection, we refer to Chapter 3. □

CORSING robustness on low regular solutions Let us consider problem (2.18) with exact solution

$$u(x) = \begin{cases} 11x - 7 & \text{if } \dfrac{7}{11} \le x < \dfrac{8}{11} \\ -\dfrac{11}{2}x + 5 & \text{if } \dfrac{8}{11} \le x < \dfrac{10}{11} \\ 0 & \text{otherwise,} \end{cases} \tag{2.30}$$

which belongs to the space $H^s(\Omega)$, for every $s < 3/2$. We compare the performance of D-CORSING $\mathcal{HS}$ and R-CORSING $\mathcal{SH}$, paying particular attention to the coefficients of the corsed-PG solution. Besides computing the errors with respect to the $L^2(\Omega)$-norm and $H^1(\Omega)$-seminorm, we are also interested in assessing how the vector $\widehat{\mathbf{u}}_m^N$ approximates vector $\mathbf{u}$ in $\mathbb{R}^N$.

We take $N = 255$, TS = 50% (corresponding to $m = 127$), and consider the minimization (P_0) only. In the R-CORSING case, only one single run is carried out. We consider the parameter $s = \|\widehat{\mathbf{u}}_m^N\|_0$ as furnished in output by the OMP-BOX package, based on the `sparse` MATLAB® format. Actually, s is computed via



(a) D-CORSING $\mathcal{HS}$ (b) R-CORSING $\mathcal{SH}$

Figure 2.12: Comparison between $\widehat{\mathbf{u}}_{127}^{255}$ and $\mathbf{u}$ for the low regular exact solution (2.30).

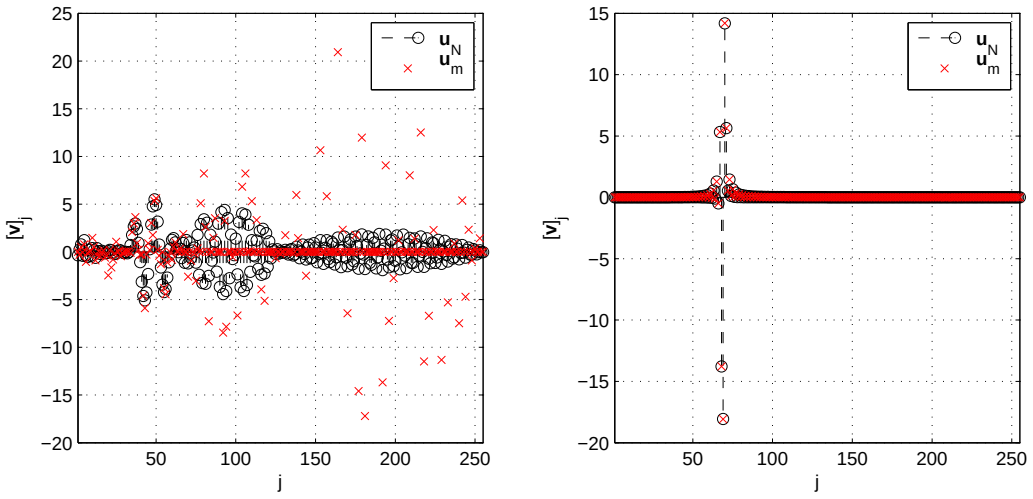
Method	s	$\|u - \widehat{u}_m^N\|_{L^2(\Omega)}$	$\|u - \widehat{u}_m^N\|_{H^1(\Omega)}$	$\|\mathbf{u} - \widehat{\mathbf{u}}_m^N\|_2$
D-CORSING $\mathcal{HS}$	106	6.1e-04	6.6e-01	3.5e-01
R-CORSING $\mathcal{SH}$	127	5.5e-03	1.6e+00	1.6e+00

Table 2.1: Comparison between D-CORSING $\mathcal{HS}$ and R-CORSING $\mathcal{SH}$ with u defined in (2.30).

the command nnz. This value can be assumed as a measure of the computational cost of the CORSING procedure, because it corresponds to the number of components of $\widehat{\mathbf{u}}_m^N$ activated by the greedy algorithm OMP.

The results are shown in Figure 2.12 and in Table 2.1. Since the exact solution (2.30) has low regularity, the best approximation error in $\mathcal{S}^N$ decays slowly [CQ82] and the resulting vector $\mathbf{u}$ is poorly compressible. In fact, the largest components in absolute value of $\mathbf{u}$ are well captured by the corsed-PG solution, whereas the long tail of the smaller coefficients is not captured at all, and causes some noise in the high frequencies of the corsed-PG solution $\widehat{\mathbf{u}}_m^N$ (Figure 2.12, (b)). On the contrary, the low regularity of u does not affect the performance of D-CORSING $\mathcal{HS}$. The main components of the sparse vector $\mathbf{u}$ are almost perfectly captured by $\widehat{\mathbf{u}}_m^N$, as highlighted in Figure 2.12, (a).

Table 2.1 also confirms that D-CORSING $\mathcal{HS}$ outperforms R-CORSING $\mathcal{SH}$. We point out that even though the number, s, of components activated by OMP is similar in the two cases, the number of meaningful components of $\widehat{\mathbf{u}}_m^N$, i.e., with absolute value grater than 10^{-3}, is very different, being 38 in the $\mathcal{HS}$ case, and 119 in the $\mathcal{SH}$ case.



(a) D-CORSING $\mathcal{HS}$ (b) R-CORSING $\mathcal{SH}$

Figure 2.13: Comparison between $\widehat{\mathbf{u}}_{127}^{255}$ and $\mathbf{u}$ for the exact solution defined in (2.31).

Remark 2.3.8. An example analogous to that just examined can be carried out by considering an exact solution that contains few significative frequencies, e.g.,

$$u(x) = \sin(69\pi(x - \tfrac{1}{4}))x(1-x). \tag{2.31}$$

The results in Figure 2.13 show that R-CORSING $\mathcal{SH}$ is a better choice than D-CORSING $\mathcal{HS}$, since in the D-CORSING $\mathcal{HS}$ case the solution is not sparse at all. □

CORSING vs best m-term approximation error We assess the behavior of CORSING with respect to the best m-term approximation error. For this purpose, we recall some results of nonlinear approximation theory following [DeV98].

Let $U = \text{span}\{\psi_j\}_{j\in\mathbb{N}}$, and

$$\sigma_s(u)_{H^1} = \inf_{w_s \in U_s} |u - w_s|_{H^1(\Omega)} \tag{2.32}$$

be the best s-term approximation error in the $H^1(\Omega)$-seminorm, where

$$U_s = \bigcup_{J\subseteq\mathbb{N}, |J|\leq s} \text{span}\{\psi_j\}_{j\in J}$$

is the set of vectors that are linear combinations of at most s trials. In the case when the trials $\{\psi_j\}$ are orthonormal with respect to a scalar product, the "inf" in (2.32) turns out to be a "min" and is realized by w_s, the vector associated with

the s largest coefficients of u, expanded in terms of $\{\psi_j\}$. Without the orthonormality of the basis, the computation of the best s-term approximation error could be a challenging issue [DeV98]. Exploiting the orthonormality property in CORSING, $\sigma_s(u)_{H^1}$ is easily computable identifying the space U with $\mathcal{H}^L$ or $\mathcal{S}^N$, for $L, N \to \infty$, since the basis functions $\{\psi_j\}$ are orthonormal with respect to the inner product, $\int_\Omega u'v' \, \mathrm{d}x$, inducing the $H^1(\Omega)$-seminorm.

To actually compare CORSING with the best m-term approximation error, we consider problem (2.18) with exact solution u_1^*. To estimate $\sigma_m(u)_{H^1}$, we compute the first $\mathcal{N}$ coefficients of u_1^* with respect to the basis $\{\psi_j\}$ for some $\mathcal{N} \gg N$, and then we evaluate the $H^1(\Omega)$-seminorm of the difference between u_1^* and the function spanned by the m trials associated with the m largest coefficients out of the $\mathcal{N}$. The coefficients with respect to the hat trial functions are computed symbolically, while those associated with the sine functions through the MATLAB® command `dst`.

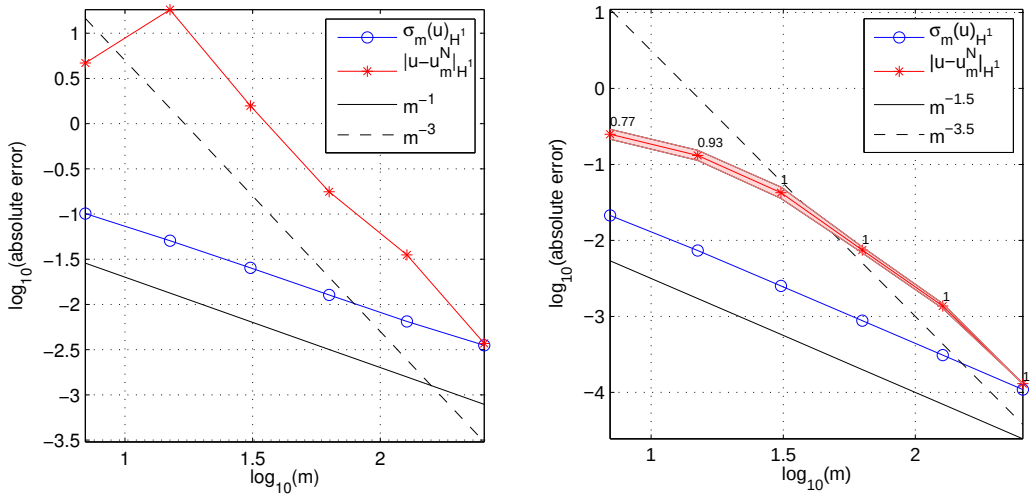
We assume $N = 255$ and $m = 7, 15, 31, 63, 127, 255$, and employ the (P_0) solver. In the R-CORSING case, 100 runs of the same test are performed for each value of m. The error $|\widehat{u}_m^N - u|_{H^1(\Omega)}$ is compared with $\sigma_s(u)_{H^1}$ with $s = m$. This choice is due to the fact that, as it can be checked numerically, the sparsity, $s = \|\widehat{\mathbf{u}}_m^N\|_0$, of the corsed-PG solution $\widehat{\mathbf{u}}_m^N$ is always very close to m. In this particular test case, for D-CORSING, $\|\widehat{\mathbf{u}}_m^N\|_0 = m$ for every value of m, whereas with R-CORSING the mean values of $\|\widehat{\mathbf{u}}_m^N\|_0$ (rounded to the nearest integer) are $7, 15, 30, 60, 119, 255$ for $m = 7, 15, 31, 63, 127, 255$, respectively.

In Figure 2.14, we compare $|\widehat{u}_m^N - u|_{H^1(\Omega)}$ with $\sigma_m(u)_{H^1}$. We observe that the CORSING error reaches the best m-term approximation error only for $m = N$ in both CORSING approaches. However, the decay rate of $|\widehat{u}_m^N - u|_{H^1(\Omega)}$ is faster than $\sigma_m(u_1^*)_{H^1}$, especially in the $\mathcal{HS}$ case.

Comparison with an SVD-based approach

In order to certify, to some extent, the new proposed approach, we compare CORSING with a reduction strategy based on the SVD factorization [GL13]. This choice is motivated by some recent model order reduction techniques, such as the Proper Orthogonal Decomposition (see, e.g., [KV02]), which exploit the SVD factorization.

Here, we compare the CORSING method with an SVD-based reduction technique, hereafter denoted by SVD-Reduction. Like CORSING, this approach is split into an assembly and a recovery phase. The first phase essentially coincides with the assembly phase of CORSING, setting $m = N$ (i.e., we build the full-PG stiffness matrix $\mathbf{A}$) followed by a further step, where we compute the SVD factorization of such a matrix, $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\intercal$ with $\mathbf{U}, \mathbf{V} \in \mathbb{C}^{N\times N}$ unitary matrices and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_N) \in \mathbb{C}^{N\times N}$ collecting the singular values σ_i of $\mathbf{A}$, in decreasing order. We finally compute the m-th order truncation of the SVD



(a) D-CORSING $\mathcal{HS}$ (b) R-CORSING $\mathcal{SH}$

Figure 2.14: Error analysis of CORSING and comparison with the best m-term approximation error.

factorization, i.e., we replace $\mathbf{A}$ with

$$\widetilde{\mathbf{A}} = \widetilde{\mathbf{U}}\widetilde{\Sigma}\widetilde{\mathbf{V}}^{\mathsf{T}},$$

where $\widetilde{\mathbf{U}}, \widetilde{\mathbf{V}} \in \mathbb{C}^{N\times m}$ contain the first m columns of $\mathbf{U}$ and $\mathbf{V}$, respectively, and $\widetilde{\Sigma}$ is the leading principal $m \times m$ submatrix of Σ. As an alternative procedure for the assembly phase, we refer to [Ose10, Gra10].

The recovery phase aims at computing an approximation $\widetilde{\mathbf{u}}_m^N$ to the full-PG solution $\mathbf{u}_N^N$, by resorting to the Moore-Penrose pseudo-inverse $\widetilde{\mathbf{A}}^+$ of $\widetilde{\mathbf{A}}$ as

$$\widetilde{\mathbf{u}}_m^N = \widetilde{\mathbf{A}}^+\mathbf{f} = \widetilde{\mathbf{V}}\widetilde{\Sigma}^+\widetilde{\mathbf{U}}^{\mathsf{T}}\mathbf{f},$$

where $\mathbf{f} \in \mathbb{C}^N$ is the full load vector, $\widetilde{\Sigma}^+ = \mathrm{diag}(\sigma_1^+, \dots, \sigma_m^+)$ with $\sigma_i^+ = 1/\sigma_i$ if $\sigma_i \neq 0$ and zero otherwise [GL13].

We apply both CORSING and SVD-Reduction to problem (2.18) with exact solution u_5^*, considering both the $\mathcal{HS}$ and $\mathcal{SH}$ settings and by employing OMP-BOX and SPGL1. The results are shown in Table 2.2, where we gather the relative error in the $L^2(\Omega)$-norm, the value of ESP for R-CORSING, and the recovery time t_{rec}, for different levels of test savings, and for $N = 1023$. The values are computed as the mean over 200 runs.

The SVD reduction leads to the most accurate results in all cases thanks to the effective low-rank approximation properties of the SVD decomposition, and it seems to be less sensitive to the compression level. On the other hand,

TS	OMP-BOX			SPGL1			SVD		
	rel. err.	ESP	t_{rec}	rel. err.	ESP	t_{rec}	rel. err.	t_{rec}	t_{svd}
$\mathcal{HS}$ (D-CORSING vs SVD-Reduction)									
5%	7.0e-06	-	4.9e-01	7.9e-06	-	1.3e-01	2.1e-06	6.5e-02	
25%	8.3e-06	-	4.8e-01	8.2e-06	-	1.2e-01	2.1e-06	5.3e-02	
45%	8.7e-06	-	4.3e-01	3.1e-05	-	9.7e-02	2.2e-06	6.2e-02	8.5e-01
65%	3.5e-05	-	7.2e-02	1.4e-04	-	9.1e-02	2.9e-06	2.8e-02	
85%	1.4e-03	-	3.3e-02	1.1e-03	-	5.5e-02	1.7e-05	1.3e-02	
$\mathcal{SH}$ (R-CORSING vs SVD-Reduction)									
5%	1.0e-06	1.00	1.7e+00	9.3e-07	1.00	7.2e-01	2.0e-07	6.5e-02	
25%	1.5e-06	1.00	9.7e-01	1.1e-06	1.00	6.6e-01	3.0e-07	4.9e-02	
45%	7.3e-06	1.00	3.7e-01	5.3e-06	1.00	8.0e-01	6.4e-07	6.8e-02	8.7e-01
65%	3.7e-05	1.00	9.9e-02	2.5e-05	1.00	6.5e-01	2.0e-06	4.4e-02	
85%	7.0e-04	1.00	2.6e-02	5.7e-04	1.00	6.0e-01	1.6e-05	1.8e-02	

Table 2.2: Quantitative comparison between CORSING and SVD-Reduction.

TS	CORSING		SVD-Reduction		
	A	**f**	**A**	**f**	SVD
$\mathcal{HS}$ (D-CORSING vs SVD-Reduction)					
5%	8.7e-02	5.3e-04			
25%	6.8e-02	6.2e-04			
45%	5.2e-02	3.7e-04	9.5e-02	6.2e-04	8.5e-01
65%	3.2e-02	3.2e-04			
85%	1.6e-02	3.6e-04			
$\mathcal{SH}$ (R-CORSING vs SVD-Reduction)					
5%	9.7e-02	3.0e-03			
25%	8.2e-02	2.3e-03			
45%	5.9e-02	1.8e-03	1.3e-01	3.0e-03	8.7e-01
65%	3.8e-02	1.3e-03			
85%	2.3e-02	8.4e-04			

Table 2.3: Computing times for CORSING and SVD-Reduction for the assembly phase.

the recovery computing times are in general comparable. Nonetheless, the bottleneck of the SVD is the computing time of the factorization (denoted by t_{svd} in Table 2.2) (the times in the table refer to the MATLAB® command svd). In particular, the asymptotic trend of t_{svd} is $\mathcal{O}(N^3)$, on the order of minutes already for $N \simeq 4000$. In Table 2.3, we investigate in more detail the computing time of the whole assembly phase, by comparing CORSING with SVD-Reduction. The times required by the assembling of $\mathbf{A}$ and $\mathbf{f}$ is substantially comparable for all the approaches. However, SVD-Reduction has a non-neglibile computational burden (i.e., at least one order larger with respect to the assembling times of $\mathbf{A}$ and $\mathbf{f}$) due to the SVD algorithm.

Moreover, the memory needed by SVD-Reduction to store $\widetilde{\mathbf{U}}, \widetilde{\mathbf{V}}, \widetilde{\Sigma}$ is even double with respect to the memory requirement of CORSING, which stores only matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$. Finally, we notice that CORSING can be implemented in a matrix-free version, because the solvers OMP-BOX and SPGL1 only require a matrix-vector multiplication. On the contrary, this is not the case of the SVD-Reduction.

After the extensive numerical assessment of CORSING on the one-dimensional Poisson problem just carried out, we slightly increase the difficulty of the test case, dealing with an advection-diffusion equation.

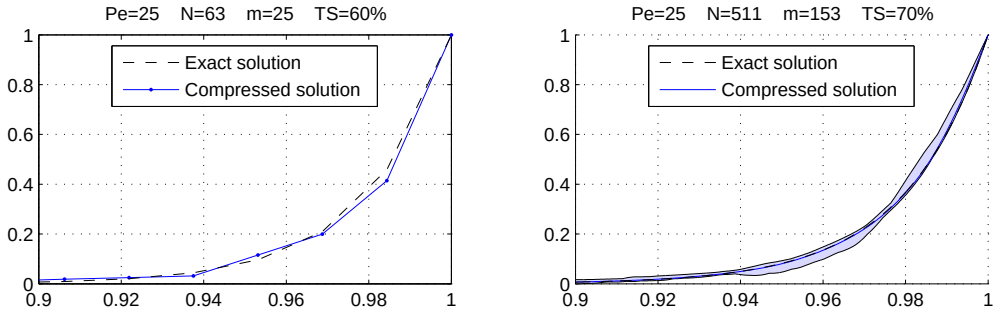
2.3.2 A 1D advection-diffusion problem

We are now interested in testing CORSING on the following advection-diffusion problem

$$\begin{cases} -\eta u'' + bu' = 0 \quad \text{in } \Omega = (0,1) \\ u(0) = 0,\ u(1) = 1, \end{cases} \tag{2.33}$$

completed with non-homogeneous Dirichlet boundary conditions, and studying the behavior of the corsed-PG solution $\widehat{u}_m^N$ in the presence of a high global Péclet number $\mathbb{P}\text{e} = b/(2\eta) \gg 1$, with b the advective field. In particular, the solution to (2.33) exhibits a layer of thickness $\mathcal{O}(\eta/b)$ at the boundary $x = 1$. The non-homogeneous condition at $x = 1$ is dealt with a standard lifting, which allows us to employ both the $\mathcal{HS}$ and the $\mathcal{SH}$ settings, in a straightforward way.

CORSING validation on an advection-dominated problem The results are shown in Figure 2.15 for two choices of N and m (i.e., of TS) for $\mathbb{P}\text{e} = 25$, where a zoom in on the numerical solution in the range $0.9 \le x \le 1$ is highlighted. In all cases, the method SPGL1 is used. D-CORSING in the $\mathcal{HS}$ case is considered in (a), while in (b) we show, for the $\mathcal{SH}$ case, a strip delimited by the minimum and the maximum for every x of the R-CORSING solution over 200 random runs, along with the associated mean. The D-CORSING solution exhibits a quite standard



(a) $\mathcal{HS}$ case (b) $\mathcal{SH}$ case

Figure 2.15: CORSING validation on the advection-diffusion problem (2.33): D-CORSING in the $\mathcal{HS}$ framework (a), R-CORSING in the $\mathcal{SH}$ setting (b).

behavior according to the chosen discretization step $h = 1/64$. On the other hand, the strip related to the R-CORSING solution is rather thin, despite the high value of the test savings (70%). This corroborates the reliability of the R-CORSING approach even for a large compression level and for an advection-dominated problem.

CORSING vs FE We assess the accuracy of CORSING applied to problem (2.33) with $\mathbb{P}e = 200$ (corresponding to $\eta = 1$ and $b = 400$) and compare its performance to those of a FE method, employing standard P1 elements on a uniform grid.

We fix $N = 2047$ and let TS vary from 10% to 80%. The CORSING results are compared with those obtained via FE for a step size equal to $1/(m+1)$, to preserve the number of tests. Only OMP-BOX is employed in the recovery phase and the relative error with respect to the $H^1(\Omega)$-seminorm is considered. For R-CORSING, the values are computed as a mean over 100 runs. Moreover, we also assess the performance of R-CORSING $\mathcal{HS}$, using weights $w_i = 1/i$ (empirically tuned as explained in Remark 2.3.7).

The results are shown in Table 2.4. We observe that the accuracy of D-CORSING $\mathcal{HS}$ and R-CORSING $\mathcal{SH}$ is comparable with that of the FE solution for moderate values of TS, whereas R-CORSING $\mathcal{HS}$ is able to outperform FE, especially when TS becomes large. We can still appreciate the benefits due to randomization by comparing the results of D-CORSING $\mathcal{HS}$ and R-CORSING $\mathcal{HS}$. It is remarkable that the R-CORSING $\mathcal{HS}$ accuracy is constant (up to the second significant digit) with respect to TS.

As a final remark, we stress that the $\mathcal{HS}$ approach is more effective than the $\mathcal{SH}$ one because the sparsity of the exact solution with respect to the corresponding trial basis is emphasized in the former case. In Figure 2.16, we plot the coefficients of the exact lifted solution with respect to $\mathcal{H}^N$ and $\mathcal{S}^N$. In the

TS	m	FE	D-CORSING $\mathcal{HS}$	R-CORSING $\mathcal{HS}$		R-CORSING $\mathcal{SH}$	
		H^1-rel. err.	H^1-rel. err.	H^1-rel. err.	ESP	H^1-rel. err.	ESP
10%	1842	5.0e-02	6.0e-02	6.0e-02	1.00	3.4e-02	1.00
20%	1638	5.5e-02	8.0e-02	6.0e-02	1.00	4.2e-02	1.00
30%	1433	6.1e-02	9.2e-02	6.0e-02	1.00	5.4e-02	1.00
40%	1228	6.8e-02	1.6e-01	6.0e-02	1.00	1.0e-01	1.00
50%	1024	7.6e-02	4.0e-01	6.0e-02	1.00	1.3e-01	1.00
60%	819	8.6e-02	4.0e-01	6.0e-02	1.00	1.7e-01	1.00
70%	614	9.6e-02	2.6e+01	6.0e-02	1.00	3.2e-01	1.00
80%	409	1.0e-01	2.2e+01	6.0e-02	1.00	6.9e-01	1.00

Table 2.4: Quantitative comparison between CORSING and FE accuracy.

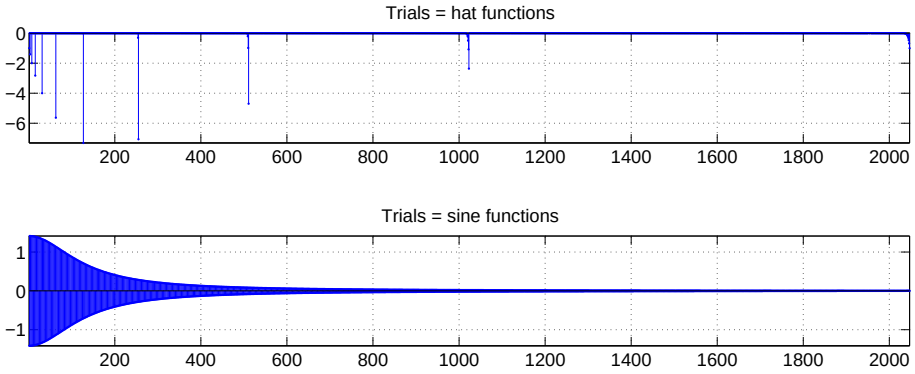


Figure 2.16: Coefficients of the lifted exact solution to problem (2.33), with $\eta = 1$ and $b = 400$, with respect to the basis $\mathcal{H}^{2047}$ (top) and $\mathcal{S}^{2047}$ (bottom).

first case (top) the vector is clearly sparse, whereas in the second case (bottom) the components exhibit only a strong decay.

2.4 Extension to the 2D case

The generalization of CORSING to the 2D case is not straightforward. In particular, we shall tackle in more detail the selection of the trial and test functions.

Analogously to the 1D case, we select two distinct bases, one associated with the space domain, the other with the frequency domain. In this section, the domain is the unit square $\Omega = (0,1)^2$.

Pyramids In order to discretize the spatial domain, we consider a hierarchical multi-scale basis of pyramids, defined as follows. The reference pyramid is

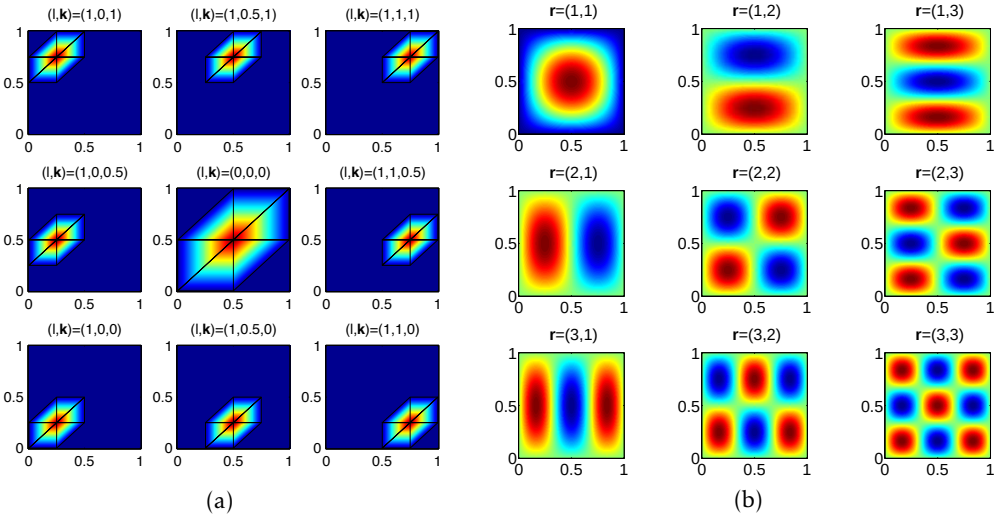
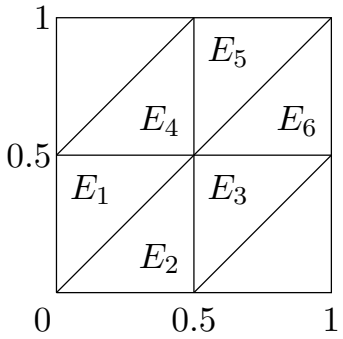


Figure 2.17: The basis $\mathcal{P}^1$ (a); the basis $\mathcal{S}^3$ (b).

$$\mathcal{P}(\mathbf{x}) := \begin{cases} x_1 & \text{if } \mathbf{x} \in E_1 \\ x_2 & \text{if } \mathbf{x} \in E_2 \\ \frac{1}{2} - x_1 + x_2 & \text{if } \mathbf{x} \in E_3 \\ \frac{1}{2} + x_1 - x_2 & \text{if } \mathbf{x} \in E_4 \\ 1 - x_2 & \text{if } \mathbf{x} \in E_5 \\ 1 - x_1 & \text{if } \mathbf{x} \in E_6 \\ 0 & \text{otherwise.} \end{cases}$$



The dyadic translation of level ℓ and multi-index $\mathbf{k} = (k_1, k_2)$ is the pyramid

$$\mathcal{P}_{\ell,\mathbf{k}}(\mathbf{x}) = \mathcal{P}(2^\ell \mathbf{x} - \mathbf{k}), \tag{2.34}$$

for $\ell \in \mathbb{N}$ and $\mathbf{k} \in \frac{1}{2}\mathbb{Z}^2$, such that $0 \le k_1, k_2 < 2^\ell$ with $(\{k_1\}, \{k_2\}) \neq (\frac{1}{2}, \frac{1}{2})$, where $\{\lambda\}$ denotes the fractional part of $\lambda \in \mathbb{R}$. We notice that $|\mathcal{P}_{\ell,\mathbf{k}}|_{H^1(\Omega)} = 1$, $\forall \ell, \mathbf{k}$.
For a fixed maximum level L, we denote this basis by

$$\mathcal{P}^L = \{\mathcal{P}_{\ell,\mathbf{k}} : 0 \le \ell \le L\}.$$

The cardinality of $\mathcal{P}^L$ is equal to $(2^{L+1} - 1)^2$, after discretizing the domain Ω with a three-directional structured mesh of uniform size $h = 1/2^{L+1}$. Each level ℓ contains $|\mathcal{P}^\ell| - |\mathcal{P}^{\ell-1}| = 2^{2\ell}3 - 2^{\ell+1}$ elements. In Figure 2.17, (a), we show the elements of $\mathcal{P}^1$.

2D Sines The basis associated with the frequency domain consists of the tensor product of sinusoidal functions, i.e.,

$$\mathcal{S}_{\mathbf{r}}(\mathbf{x}) = \frac{2}{\pi\|\mathbf{r}\|_2}\sin(r_1\pi x_1)\sin(r_2\pi x_2), \tag{2.35}$$

with $\mathbf{r} = (r_1, r_2) \in \mathbb{N}^2$, and $1 \le r_1, r_2 \le R$ for some integer $R \ge 1$. The normalization constant ensures that $|\mathcal{S}_{\mathbf{r}}|_{H^1(\Omega)} = 1$. This basis is denoted by

$$\mathcal{S}^R := \{\mathcal{S}_{\mathbf{r}} : 0 \le \mathbf{r} \le R\}.$$

The elements of $\mathcal{S}^3$ are plotted in Figure 2.17, (b).[7]

In the 2D case, the ordering of the basis functions plays a crucial role. Indeed, this choice affects the D-CORSING and the R-CORSING strategies, that both depend on how the trials ψ_j and the tests φ_i are ordered.

For $\mathcal{P}^L$, we adopt the lexicographic ordering on the multi-index $(\ell, \mathbf{k}) = (\ell, k_1, k_2)$, i.e.,

$$(0,0,0),\ (1,0,0),\ (1,\tfrac{1}{2},0),\ (1,\tfrac{1}{2},1),\ (1,1,0),\ldots,(L,2^L-1,2^L-1).$$

For $\mathcal{S}^R$, we use a diagonal arrangement on $\mathbf{r} = (r_1, r_2)$, i.e.,

$$(1,1),\ (1,2),\ (2,1),\ (1,3),\ (2,2),\ (3,1),\ldots,(R,R).$$

In practice, the multi-index (r_1, r_2) is ordered such that the sum $r_1 + r_2$ is increasing, and, for a fixed sum, the lexicographic order is used.

We now apply the CORSING procedure as described in Section 2.2. Analogously to the 1D case, we cast D-CORSING in a $\mathcal{PS}$ ($\mathcal{P}$yramid vs $\mathcal{S}$ine) setting, i.e., we pick

$$\psi_j = \mathcal{P}_{\ell(j),\mathbf{k}(j)} \qquad \text{and} \qquad \varphi_i = \mathcal{S}_{\mathbf{r}(i)},$$

whereas, due to aliasing, R-CORSING is employed in a $\mathcal{SP}$ ($\mathcal{S}$ine vs $\mathcal{P}$yramid) setting such that

$$\psi_j = \mathcal{S}_{\mathbf{r}(j)} \qquad \text{and} \qquad \varphi_i = \mathcal{P}_{\ell(i),\mathbf{k}(i)}.$$

In the R-CORSING case, the weights for the test selection procedure are empirically chosen as

$$w_i = 2^{-\ell(i)}.$$

to favour the lower levels.

[7] In principle, this notation could generate some ambiguity with the family of one-dimensional sine functions. Nevertheless, in the next developments, the dimension considered will be totally clear from the context.

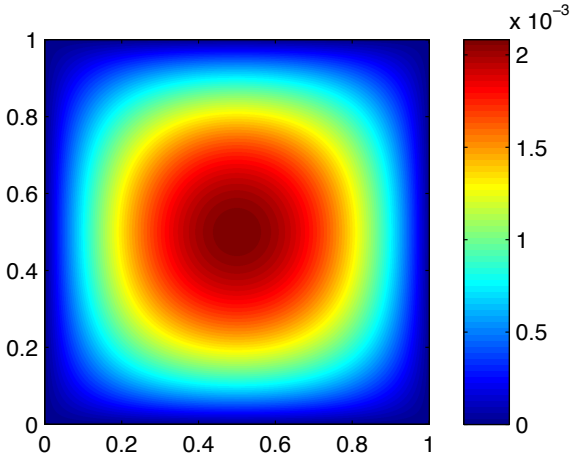


Figure 2.18: The exact solution u^* to (2.36).

Remark 2.4.1. The MATLAB® implementation of the 2D CORSING requires some care, especially in assembling the stiffness matrix $\mathbf{A}$. For this purpose, we employ a symbolic approach where explicit formulas for A_{ij} are first computed via the symbolic toolbox and then evaluated using a vectorization programming to avoid loops which unavoidably slow down the performance of the MATLAB® scripts (see also Section 2.2.2). □

2.4.1 The model 2D Poisson problem

First, we focus on the Poisson problem, with Dirichlet homogeneous boundary conditions

$$\begin{cases} -\Delta u = f & \text{in } \Omega = (0,1)^2 \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{2.36}$$

with

$$f(\mathbf{x}) = \frac{1}{15}[x_1(1-x_1) + x_2(1-x_2)].$$

The exact solution to (2.36) is the bubble function

$$u^*(\mathbf{x}) = \tfrac{1}{30}(x_1 - x_1^2)(x_2 - x_2^2) \tag{2.37}$$

plotted in Figure 2.18. This example is taken from [JMPY10].

In Figure 2.19, we show some results of D-CORSING applied to problem (2.36) in the $\mathcal{PS}$ case. The number of trials is fixed to $N = 961$, corresponding to $L = 4$ maximum hierarchical levels. Then, the number m of tests is chosen such that TS assumes the values 85%, 90% and 95%. For each combination of N and m, both SPGL1 and OMP-BOX are used. The color plots of the resulting corsed-PG

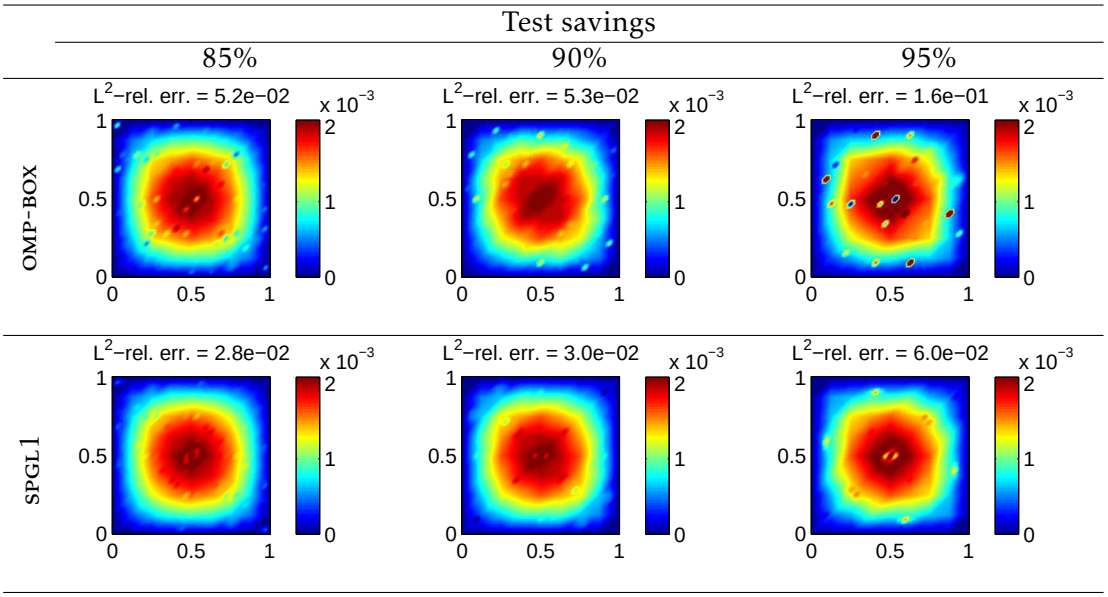


Figure 2.19: Numerical performance of D-CORSING in the $\mathcal{PS}$ case: $L^2(\Omega)$-norm relative error and color plot of the corsed-PG solution using OMP-BOX (top), and SPGL1 (bottom).

solutions are provided along with the values of the relative errors with respect to the $L^2(\Omega)$-norm. The results are promising, especially for SPGL1, considering the high level of compression. Indeed, we have a sufficiently accurate approximation of the true solution for TS $\lesssim 90\%$, that corresponds to using at least $m = 96$ tests out of 961 available functions. The OMP-BOX solver is also able to capture the main features of u^*, except for some localized noise. As expected, the error increases as TS grows, and it is larger in the case of OMP-BOX.

We check now the performance of R-CORSING in the $\mathcal{SP}$ setting, resorting to $\mathcal{S}^R$ with $R = 31$, and with $N = 31^2 = 961$ trials (see Figure 2.20). In order to assess the influence of randomization, we carry out three random experiments for each of the three choices of TS. OMP-BOX shows the best performance, whereas SPGL1 seems more sensitive to randomization as well as to the compression level. Moreover, OMP-BOX provides more accurate corsed-PG solutions than SPGL1 does.

A possible justification of the different results in Figures 2.19 and 2.20 is the different sparsity of the coefficients of the exact solution (2.37) with respect to the pyramid or sine basis. If the trials are the sine functions, the resulting vector $\widehat{\mathbf{u}}_N^N$, with $N = 961$, associated with the $\mathcal{SP}$ approach is much sparser than the vector characterizing the $\mathcal{PS}$ expansion (see Figure 2.21).

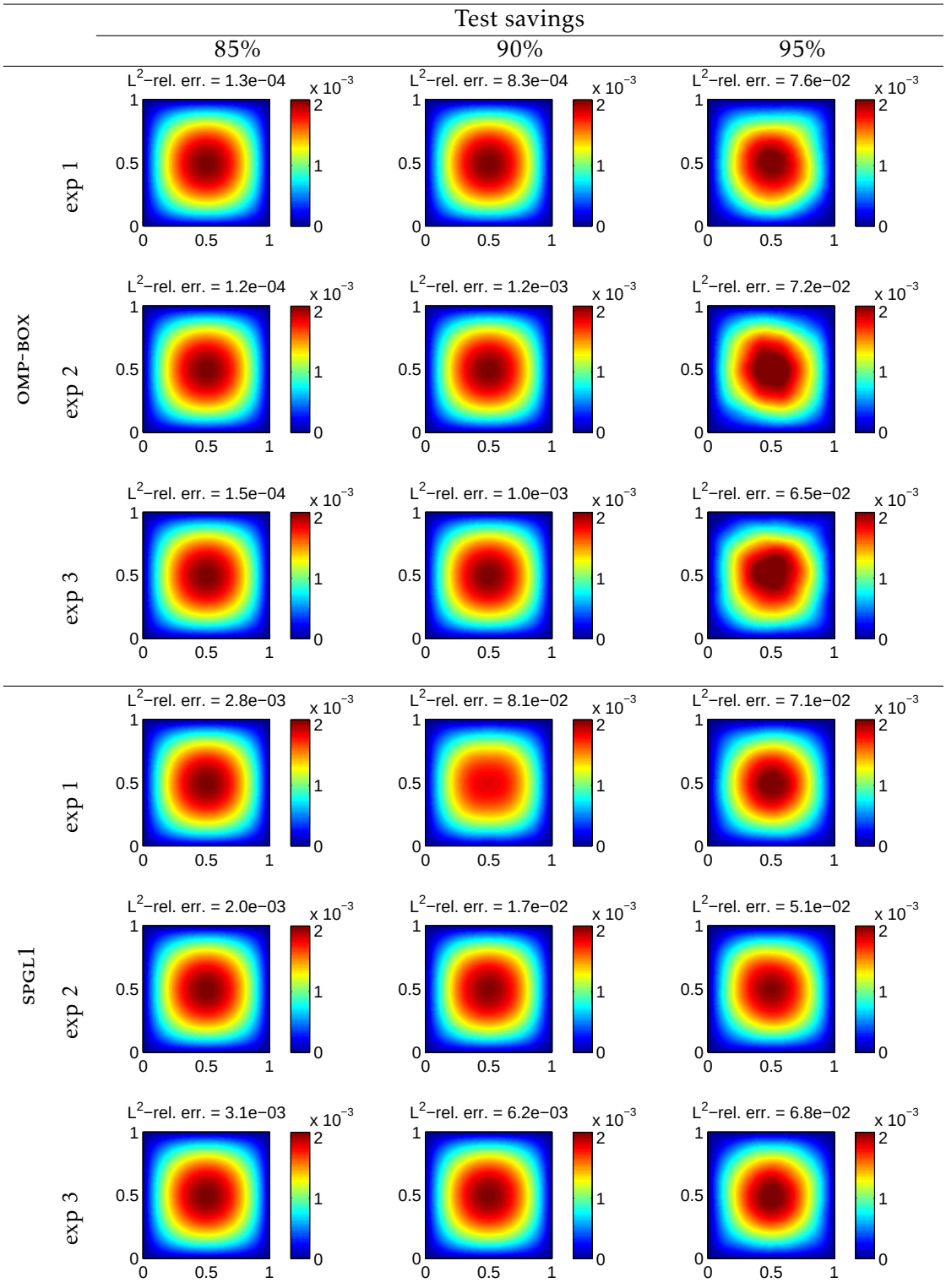


Figure 2.20: Numerical performance of R-CORSING in the $\mathcal{SP}$ case: $L^2(\Omega)$-norm relative error and color plot of the corsed-PG solution using OMP-BOX (top panel), and SPGL1 (bottom panel).

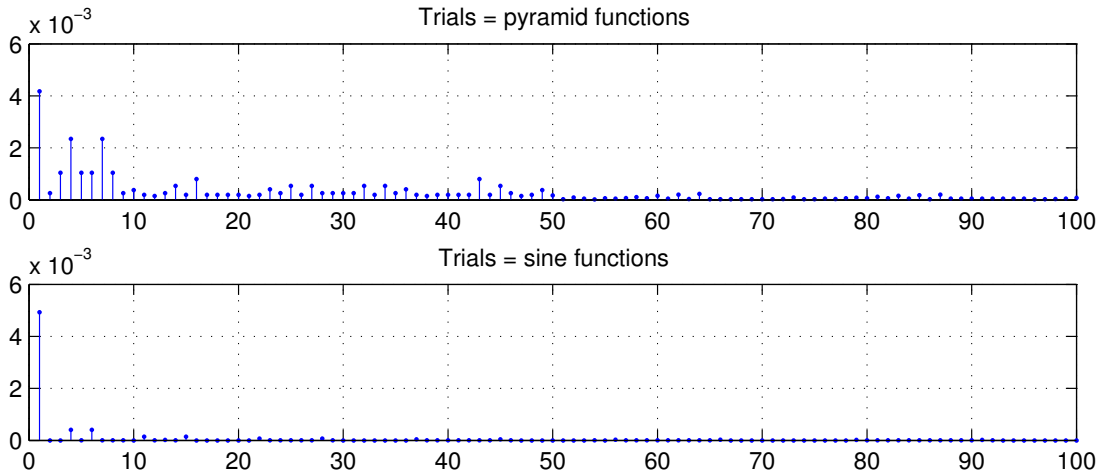


Figure 2.21: Absolute value of the first 100 coefficients of the full-PG solution $\widehat{\mathbf{u}}_N^N$ to problem (2.36), with $N = 961$: $\mathcal{PS}$ approach with $L = 4$ (top); $\mathcal{SP}$ formulation with $R = 31$ (bottom).

2.4.2 A 2D advection-dominated example

After the assessment conducted on the 2D Poisson model problem, we evaluate the CORSING performances on the following 2D advection-dominated problem

$$\begin{cases} -\eta\Delta u + \mathbf{b}\cdot\nabla u = f & \text{in } \Omega = (0,1)^2, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{2.38}$$

where $\mathbf{b} = [1,1]^\intercal$, $0 < \eta \ll 1$ and the forcing term is computed so that the exact solution be

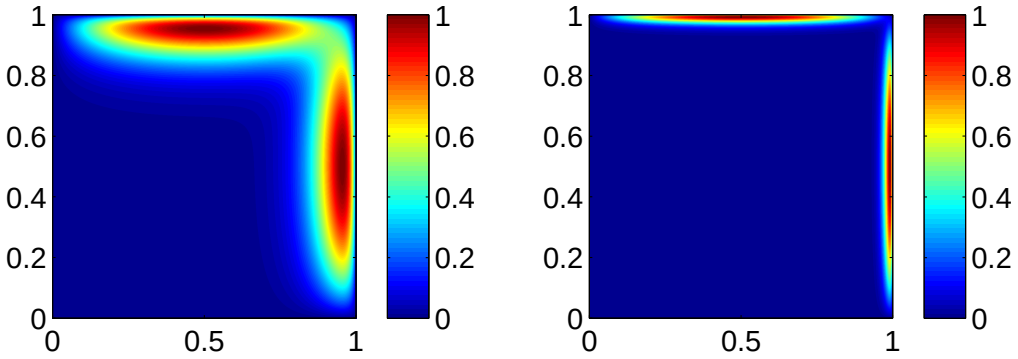
$$u_\eta^*(\mathbf{x}) = C_\eta(x_1 - x_1^2)(x_2 - x_2^2)(e^{x_1/\eta} + e^{x_2/\eta} - 2), \tag{2.39}$$

where $C_\eta > 0$ is chosen such that

$$\max_{\mathbf{x}\in\Omega} u_\eta^*(\mathbf{x}) = 1.$$

The function u_η^* exhibits two boundary layers along the edges $\{x_1 = 1\}$ and $\{x_2 = 1\}$ of Ω, that become thinner and thinner as η approaches the critical value zero, corresponding to a pure transport problem (see Figure 2.22).

Remark 2.4.2 (Condition number of the full-PG stiffness matrix). We numerically estimate the condition number associated to the full-PG discretization of problem (2.38), with $\eta = 0.1$. We consider $L = 1,2,3,4,5$, corresponding to $R = 3,7,15,31,63$ and a stiffness matrix of dimension $N = 9,49,225,961,3969$, respectively. In both the $\mathcal{PS}$ and $\mathcal{SP}$ cases, the condition number grows proportionally to $N^{1/2}$ (Figure 2.23), i.e., proportionally to $1/h$, where h is the mesh diameter associated with the last hierarchical level ($\ell = L$) of pyramids. This is remarkable, since employing the FE method, the condition number would grow faster, proportionally to $1/h^2$ (see [QV08, Section 6.3.2]). □



(a) $\eta = 0.05$ (b) $\eta = 0.01$

Figure 2.22: The exact solution u^*_η to (2.38) for $\eta = 0.05, 0.01$.

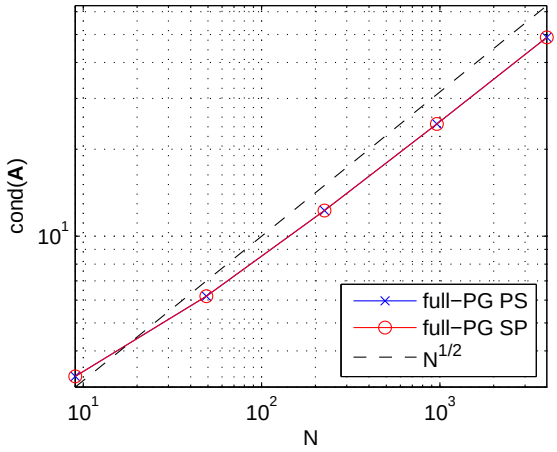


Figure 2.23: Condition number of the stiffness matrix associated with the full-PG $\mathcal{PS}$ and $\mathcal{SP}$ approaches applied to problem (2.38), with $\mu = 0.1$.

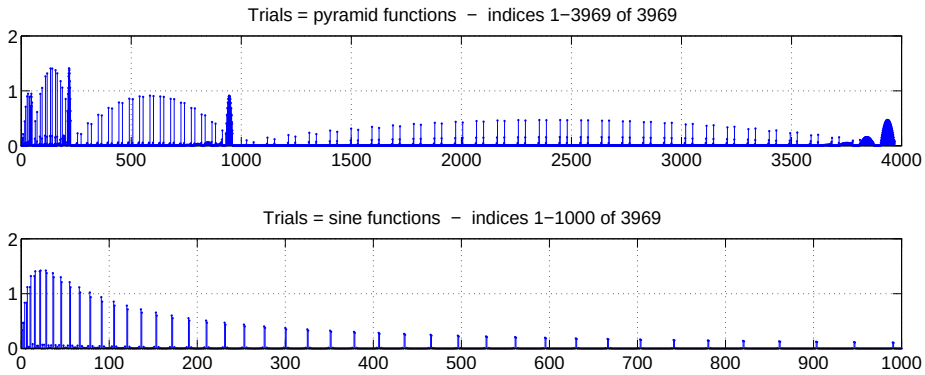


Figure 2.24: Absolute value of the coefficients of the full-PG solution $\widehat{\mathbf{u}}_N^N$ to problem (2.38), with $\eta = 0.05$ and $N = 3969$: $\mathcal{PS}$ approach with $L = 5$ (top); $\mathcal{SP}$ formulation with $R = 63$ (bottom).

2.4.3 CORSING performance

In this case, we choose the $\mathcal{SP}$ combination for trials and tests. Indeed, u_η^* is much sparser with respect to the sine functions than the pyramids. In particular, in Figure 2.24 we provide the full-PG solution coefficients in $\widehat{\mathbf{u}}_N^N$ with $\eta = 0.05$ and $N = 3969$ (corresponding to $R = 63$ and $L = 5$) for the $\mathcal{PS}$ case (top) and the $\mathcal{SP}$ case (bottom). We notice that the $\mathcal{SP}$ coefficients exhibit a faster decay than $\mathcal{PS}$ and that the most significant coefficients are essentially the first thousand. This situation leads us to adopt the R-CORSING $\mathcal{SP}$ strategy in order to reach higher TS rates and a substantial cost reduction with respect to the full-PG approach, as shown in Section 2.4.4.

In the first numerical test, we set $\eta = 0.05$ and $N = 3969$. The results are shown in Figure 2.25. For each value of TS = 80%, 85%, 87.5%, 50 random experiments have been performed using the OMP-BOX solver in the recovery phase, to guarantee faster performances for high TS values. The resulting ESP and the mean relative error with respect to the $L^2(\Omega)$-norm of the solutions in the first cluster are provided. Moreover, in order to show the robustness of the method even in the worst case scenario, for each value of TS we show the color plot of the solution characterized by the highest relative error in the first cluster. We can appreciate that, even for extremely high compression levels, the boundary layers are well captured and that the ESP values are very high (always greater than 0.92). We also notice that, as TS increases, a small noise appears where the solution is smooth. However, this does not spoil the results.

As a last numerical assessment, we select $\eta = 0.01$ and $R = 127$, corresponding to $L = 6$ and $N = 16129$. The values of TS are 85% and 90%, and, for each of them, 50 random experiments are performed. The results are organized similarly to those of the previous experiment and they are shown in Figure 2.26. We have an experimental confirmation that CORSING can be successfully applied to advection-dominated problems also in the two-dimensional case, also with

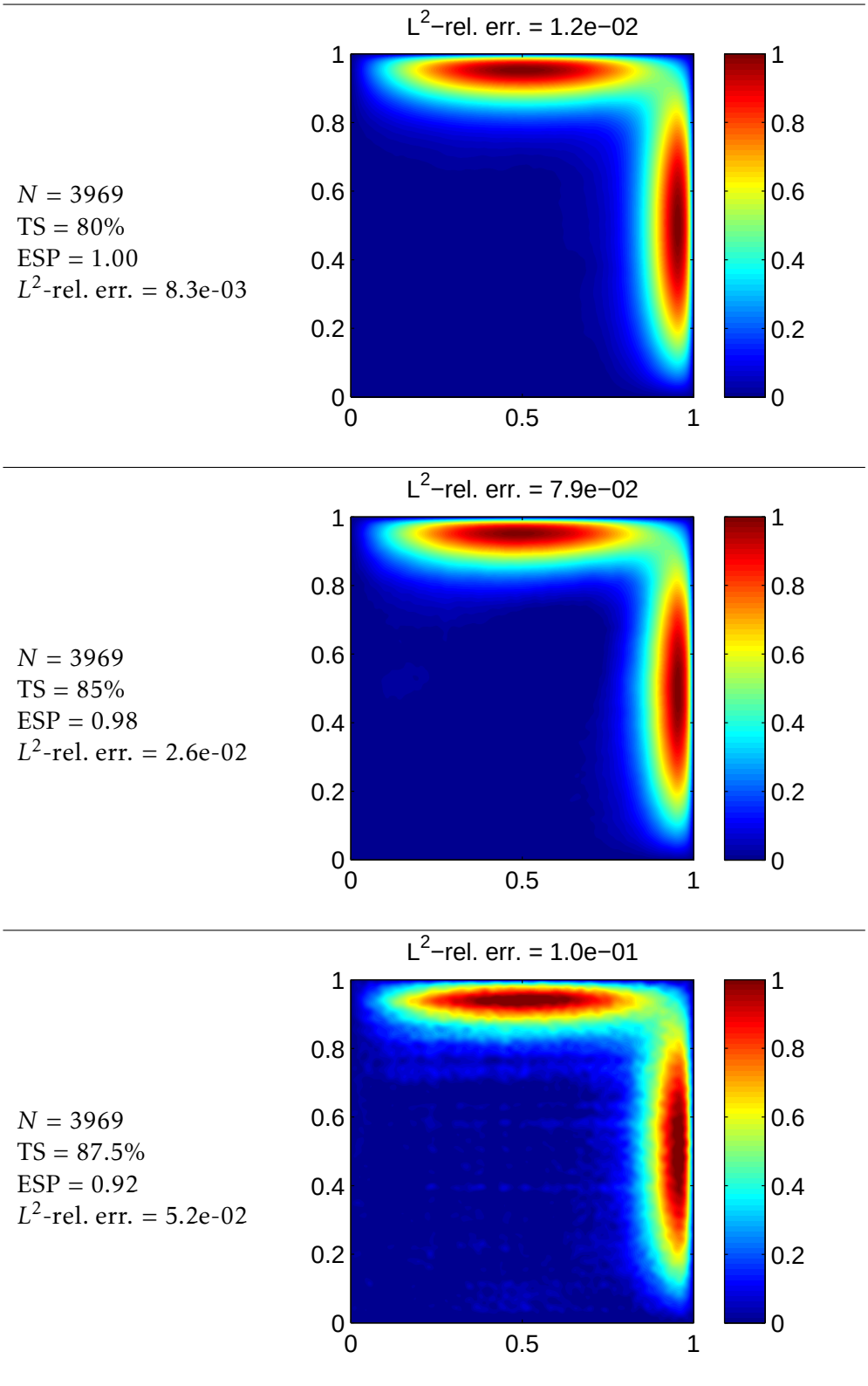


Figure 2.25: Assessment of R-CORSING $\mathcal{SP}$ on problem (2.38), with $\eta = 0.05$: worst solution in the first cluster (right).

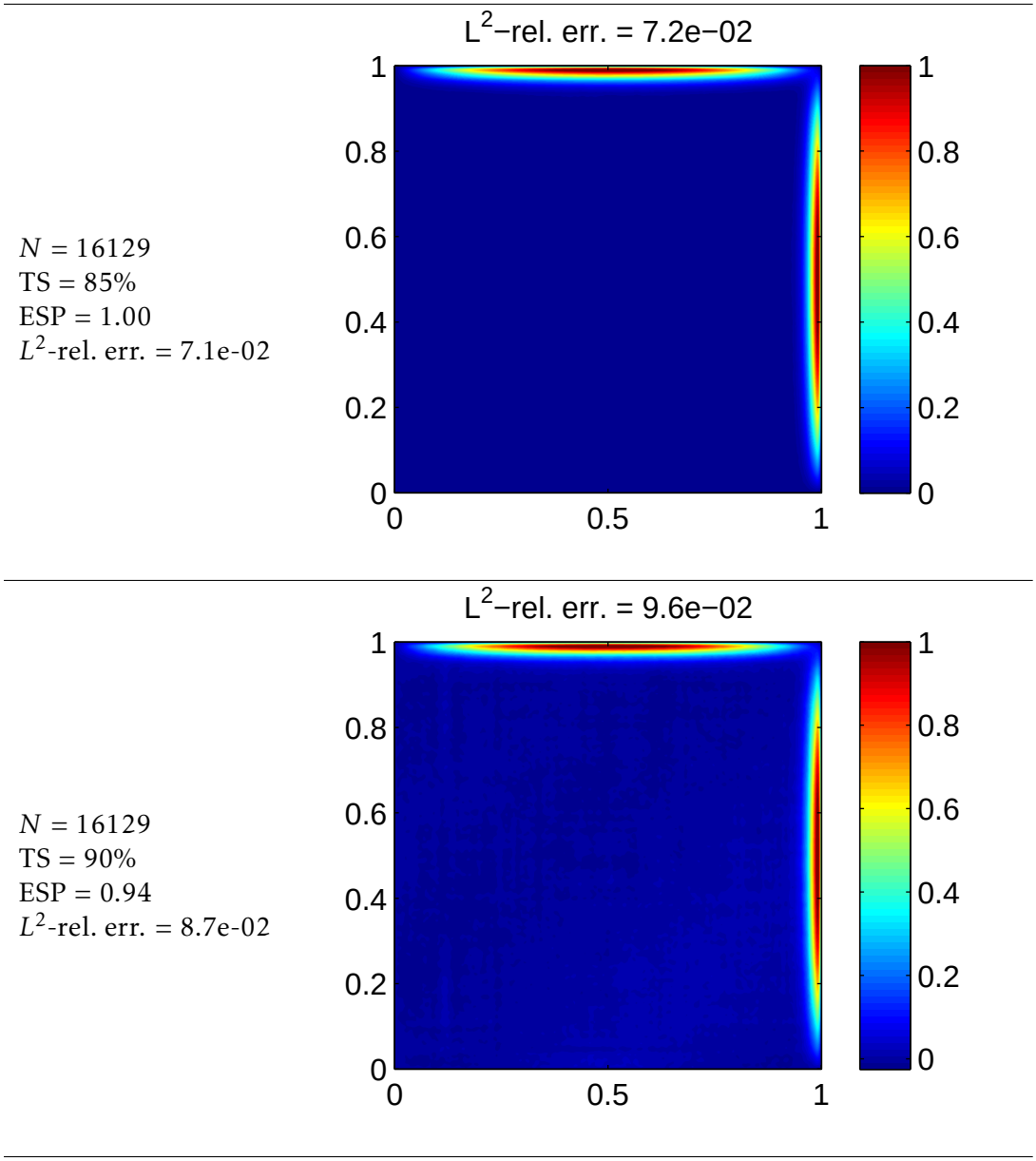


Figure 2.26: R-CORSING $\mathcal{SP}$ applied on problem (2.38), with $\eta = 0.01$: worst solution in the first cluster (right).

Problem	Bases	full-PG			CORSING			
		A	**f**	t_{rec}	TS	**A**	**f**	t_{rec}
(2.36) $N = 961$	$\mathcal{SP}$	4.5e+00	8.8e-03	3.7e-02	85%	6.1e-01	5.3e-03	1.7e-02
					90%	3.8e-01	4.5e-03	1.3e-02
					95%	2.1e-01	4.4e-03	1.0e-02
	$\mathcal{PS}$	4.4e+00	2.8e-03	3.1e-02	85%	6.0e-01	8.9e-04	2.4e-02
					90%	3.7e-01	8.5e-04	1.5e-02
					95%	1.9e-01	8.4e-04	9.5e-03
(2.38) $\eta = 0.05$ $N = 3969$	$\mathcal{SP}$	1.4e+02	2.9e-01	1.1e+00	80%	2.9e+01	1.2e-01	2.1e+00
					85%	2.3e+01	9.5e-02	1.1e+00
					87.5%	1.8e+01	8.4e-02	9.3e-01
(2.38) $\eta = 0.01$ $N = 16129$	$\mathcal{SP}$	2.5e+03	9.1e-01	7.1e+01	85%	3.8e+02	2.7e-01	8.1e+01
					90%	2.5e+02	2.0e-01	3.4e+01

Table 2.5: Comparison between full-PG and CORSING computing times in the 2D case.

very high TS rates. Also in this case, a small noise appears in the smooth region for TS = 90%, but the boundary layers are well captured even in the worst case scenario of this challenging problem.

2.4.4 Analysis of cost reduction with respect to the full-PG approach

The two-dimensional setting is better suited than the one-dimensional case to appreciate the cost reduction provided by the CORSING method with respect to the full-PG approach, both in the assembly and in the recovery phase. In particular, regarding the recovery phase, we restrict our attention to the OMP-BOX solver, since it exhibits faster performances than SPGL1 for high values of TS. The recovery phase of full-PG is instead performed using the MATLAB® command \ (backslash).

The results are shown in Table 2.5. All the reported times are expressed in seconds and they have been computed as mean values over 50 repeated experiments. The time to perform the symbolic computations to build **A** and **f** is not considered. In the recovery phase, the backslash command is a tough nut to crack, especially considering that OMP-BOX is not a built-in MATLAB® function. Nevertheless, the OMP algorithm is able to run four times faster for $N = 961$ and TS = 95% and two times faster for $N = 16129$ and TS = 90%. In the second case, the gain in recovery time is non-negligible as well. However, the speed-up is dramatic for the assembly computing times, especially if we compare those needed to assemble the stiffness matrix **A**. In fact, in all cases, the speed-up is around one order of magnitude. In particular, for $N = 16129$ the assembly phase of full-PG employs around 40 minutes, while that of CORSING with TS = 90%, only 4 minutes. Finally, there is also a memory burden associated with the full-PG approach, not reported in the table. For example, for $N = 16129$, **A** has 260144641 full entries, and its storage requires 1.94 GB, while, for CORSING

with TS=90%, the memory space required is only 198 MB.

The extensive numerical assessment carried out in this chapter shows the robustness and the reliability of the CORSING method applied to one- and two-dimensional ADR equations. Now, the next goal is to understand and formalize the empirical recipes just discovered in a rigorous way. This will be the intent of Chapter 3.

Chapter 3

A theoretical study of CORSING

The goal of this chapter is to set up a theoretical analysis of R-CORSING, (here simply referred to as CORSING) providing sufficient conditions for convergence, and formalizing the empirical recipes given in Chapter 2. With this aim, we introduce a novel variant of the classical inf-sup condition (see Section 2.1.1), where the infimum is considered among the sparse elements of the trial space and the supremum over a small test space. We refer to this condition as *Restricted Inf-Sup Property* (RISP), since it combines the inf-sup condition and the Restricted Isometry Property (RIP) of CS, discussed in Section 1.2.1. Another important tool of the analysis below is the concept of *local a-coherence*, a generalization of the *local coherence* (see Section 1.2.5) to bilinear forms on Hilbert spaces. In particular, we have been inspired by the results reviewed in Section 1.2.6, where an optimal recovery result for CS, with non-uniform random subsampling based on the local coherence, is proved for the Haar and Fourier discrete bases.

The great potential of the theory presented in this chapter is that the number of tests m grows linearly or, at most, quadratically with respect to the sparsity s of the CORSING solution and only logarithmically with respect to the total dimension N of the trial space, making the CORSING method appealing for problems where the total number of degrees of freedom N is huge, but one is interested in recovering only the most significant $s \ll N$ coefficients with respect to a suitable trial basis. For example, in the case of solutions exhibiting few features (details) arising in unknown regions of the domain, which, in principle, need a severe mesh refinement to be captured, thus making a FE direct numerical simulation very expensive.

A significant part of this chapter corresponds to the work in [BNMP15].

Outline of the chapter In Section 3.1, we formalize the CORSING procedure, defining all the input/output variables involved in the algorithm. The theoretical analysis based on the RISP is presented in Section 3.2, together with an

optimal RIP result. Then, the application of the theory to advection-diffusion-reaction equations is discussed in Section 3.3. In Section 3.4, we provide some additional numerical results.

3.1 Formalizing the CORSING procedure

In this section, after setting up the notation, we formalize the COmpRessed SolvING procedure, in short, CORSING, introduced in Chapter 2.

3.1.1 Notation

Consider two separable Hilbert spaces, $U = \text{span}\{\psi_j\}_{j\in\mathbb{N}}$ and $V = \text{span}\{\varphi_q\}_{q\in\mathbb{N}}$, generated by the bases $\{\psi_j\}_{j\in\mathbb{N}}$ and $\{\varphi_q\}_{q\in\mathbb{N}}$, respectively, and equipped with the inner products $(\cdot,\cdot)_U$ and $(\cdot,\cdot)_V$. Given two positive integers N and M, we define the finite dimensional truncations of U and V, which represent the *trial* and *test* space, respectively, as

$$U^N := \text{span}\{\psi_j\}_{j\in[N]} \quad \text{and} \quad V^M := \text{span}\{\varphi_q\}_{q\in[M]}.$$

The spaces U^N and V^M are associated with the full-PG discretization (see Section 2.1.2). Notice that, contrarily to the previous chapter, M and N can possibly be distinct (usually, $M \geq N$). We denote the span of the basis functions relative to a given subset of indices $\mathcal{S} \subseteq [N]$ as

$$U^N_{\mathcal{S}} := \text{span}\{\psi_j\}_{j\in\mathcal{S}}.$$

Given a positive integer $s \leq N$, we also define the set U^N_s of s-sparse functions of U^N with respect to the basis $\{\psi_j\}_{j\in[N]}$ as the set of all functions that are linear combinations of at most s basis functions, namely

$$U^N_s := \bigcup_{\mathcal{S}\subseteq[N];|\mathcal{S}|=s} U^N_{\mathcal{S}}.$$

We stress that U^N_s is not a vector space. Indeed, the sum of two s-sparse elements is in general $2s$-sparse. The sets $V^M_{\mathcal{T}}$ and V^M_m are defined analogously, for every $\mathcal{T} \subseteq [M]$ and $m \leq M$.

We denote by U^* and V^* the dual spaces of U and V, respectively.

In order to define the *reconstruction* and *decomposition* operators, we need $\{\psi_j\}_{j\in\mathbb{N}}$ and $\{\varphi_q\}_{q\in\mathbb{N}}$ to be *Riesz bases*.

Definition 3.1 (Riesz basis)**.** A sequence $\{\psi_j\}_{j\in\mathbb{N}} \subseteq U$ is a *Riesz basis* if it is complete in U and there exist two constants $0 < c_\psi \leq C_\psi < \infty$ such that

$$c_\psi \|\mathbf{u}\|_2^2 \leq \left\| \sum_{j\in\mathbb{N}} u_j \psi_j \right\|_U^2 \leq C_\psi \|\mathbf{u}\|_2^2, \quad \forall \mathbf{u} \in \ell^2. \tag{3.1}$$

In practice, condition (3.1) states that the U-norm of a function and the ℓ^2-norm of its coefficients with respect to the basis $\{\psi_j\}_{j\in\mathbb{N}}$ are equivalent. Analogous constants relative to the basis $\{\varphi_q\}_{q\in\mathbb{N}}$ are denoted c_φ and C_φ. For more details about Riesz bases, we refer the reader to [Chr02].

We can introduce now the reconstruction and decomposition operators. These allow us to switch between functions and the corresponding coefficients in the basis expansion.

Definition 3.2. The *reconstruction operator* $\Psi : \ell^2 \to U$ related to a Riesz basis $\{\psi_j\}_{j\in\mathbb{N}}$ of U associates the linear combination

$$u = \Psi\mathbf{u} := \sum_{j=1}^{\infty} u_j \psi_j,$$

with a sequence $\mathbf{u} = (u_j)_{j\in\mathbb{N}} \in \ell^2$. The *decomposition operator* $\Psi^* : U \to \ell^2$ applied to a given function $u \in U$ is defined componentwise as

$$(\Psi^* u)_k := (u, \psi_k^*)_U, \quad \forall k \in \mathbb{N},$$

where $\{\psi_k^*\}_{k\in\mathbb{N}}$ is the basis biorthogonal to $\{\psi_j\}_{j\in\mathbb{N}}$, namely, $(\psi_j, \psi_k^*)_U = \delta_{j,k}, \forall j,k \in \mathbb{N}$.

The reconstruction operator Φ and the decomposition operator Φ^* associated with the basis $\{\varphi_q\}_{q\in\mathbb{N}}$ of V are defined analogously.

Remark 3.1.1. We observe that $\Psi\Psi^* = Id_U$ and $\Psi^*\Psi = Id_{\ell^2}$. □

Throughout the chapter, we will focus on the weak problem (2.1), assuming the bilinear form $a : U\times V \to \mathbb{R}$ to fulfil the continuity property (2.3), the inf-sup property (2.4), and property (2.5), that ensure the validity of Theorem 2.2.

3.1.2 Main hypotheses

We will use three assumptions throughout the chapter.

Hypothesis 1 (Orthonormal tests). The test basis $\{\varphi_q\}_{q\in\mathbb{N}}$ is an orthonormal system of V.

Hypothesis 1 is not strictly necessary, but it makes the exposition simpler. Indeed, all the results shown in this chapter can be generalized assuming $\{\varphi_q\}_{q\in\mathbb{N}}$ to be a Riesz basis. Of course, the Riesz constants c_φ and C_φ should be appropriately tracked throughout the proofs.

We generalize the notion of local coherence (see Section 1.2.5) to bilinear forms defined over Hilbert spaces.

Definition 3.3 (Local a-coherence $\boldsymbol{\mu}^N$)**.** Given $N \in \mathbb{N} \cup \{\infty\}$, the real-valued sequence $\boldsymbol{\mu}^N$ defined as

$$\mu_q^N := \sup_{j \in [N]} |a(\psi_j, \varphi_q)|^2, \quad \forall q \in \mathbb{N},$$

is called *local a-coherence of* $\{\psi_j\}_{j\in[N]}$ *with respect to* $\{\varphi_q\}_{q\in\mathbb{N}}$.

The second hypothesis concerns the local a-coherence.

Hypothesis 2 (Summability of $\boldsymbol{\mu}^N$). The local a-coherence of $\{\psi_j\}_{j\in[N]}$ with respect to $\{\varphi_q\}_{q\in\mathbb{N}}$ fulfills the summability condition

$$\|\boldsymbol{\mu}^N\|_1 < +\infty,$$

or, equivalently, $\boldsymbol{\mu}^N \in \ell^1$.

Notice that Hypothesis 2 does not hinge on the ordering considered for the elements of the truncated trial basis $\{\psi_j\}_{j\in[N]}$.

The last hypothesis concerns an explicit upper bound to the local a-coherence.

Hypothesis 3 (Upper bound $\boldsymbol{\nu}^N$). For every $N \in \mathbb{N}$, we assume to have a computable componentwise upper bound $\boldsymbol{\nu}^N$ to the local a-coherence $\boldsymbol{\mu}^N$, i.e., a real-valued sequence such that

$$\mu_q^N \le \nu_q^N, \quad \forall q \in \mathbb{N}.$$

For every $M \in \mathbb{N}$, we define the vector $\boldsymbol{\nu}^{N,M} \in \mathbb{R}^M$ as the restriction of $\boldsymbol{\nu}^N$ to the first M components. Moreover, we require that

- the vector $\boldsymbol{\nu}^{N,M}/\|\boldsymbol{\nu}^{N,M}\|_1$ is efficiently computable for every $N, M \in \mathbb{N}$;
- there exists a real bivariate polynomial P such that

$$\|\boldsymbol{\nu}^{N,M}\|_1 \lesssim P(\log N, \log M).$$

The upper bound $\boldsymbol{\nu}^N$ needs not be sharp.

3.1.3 The CORSING procedure

The CORSING procedure is summarized in Algorithm 3.1. Let us now describe in more detail the input/output variables and the main steps of the method.

INPUT

- N: dimension of the trial space;
- $s \ll N$: number of trial coefficients to recover;
- upper bound $\boldsymbol{\nu}^N$ in Hypothesis 3 and four positive constants $\widehat{\gamma}$, $\widehat{C}$, $\overline{\gamma}$, and $\overline{C}$ used to select the dimension M of the test space and the m tests to perform.

OUTPUT

- $\widehat{u} \in U_s^N$: approximate s-sparse solution to (2.1).

1. Definition of M and m The test space dimension M and the number m of tests to perform are chosen as functions of N and s as

$$M = \widehat{C} s^{\widehat{\gamma}} N, \quad m = \overline{C} s^{\overline{\gamma}} \|\boldsymbol{\nu}^{N,M}\|_1 \log(N/s).$$

In Section 3.2, we prove the existence of suitable values for the constants $\widehat{\gamma}$, $\widehat{C}$, $\overline{\gamma}$ that ensure the CORSING algorithm to recover the best s-term approximation to u in expectation and in probability. In Section 3.2 we prove that $\overline{\gamma} = 1, 2$ are valid choices, and in Section 3.3 we perform a sensitivity analysis on the constants $\widehat{C}$ and $\overline{C}$ for some specific differential problems and with $\overline{\gamma} = 1, 2$. On the contrary, the value of $\widehat{\gamma}$ seems to depend on the trial and test bases considered.

2. Test selection In order to formalize the test selection procedure, we introduce a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ and consider $\tau_1, \dots, \tau_m$ as i.i.d. discrete random variables taking values in $[M]$, namely

$$\tau_i : \Omega \to [M], \quad \forall i \in [m].$$

Moreover, given a vector $\mathbf{p} = (p_q)_{q \in [M]} \in [0,1]^M$ such that $\|\mathbf{p}\|_1 = 1$, the probability law is defined as

$$\mathbb{P}\{\tau_i = q\} = p_q, \quad \forall q \in [M].$$

Throughout the chapter, the vector $\mathbf{p}$ will be assumed to be of the form

$$\mathbf{p} := \frac{\boldsymbol{\nu}^{N,M}}{\|\boldsymbol{\nu}^{N,M}\|_1}, \tag{3.2}$$

where the values for $\boldsymbol{\nu}^{N,M}$ are known from Hypothesis 3.

Notice that the independence of the indices $\tau_1, \dots, \tau_m$ is assumed in order to simplify the theoretical analysis. With this choice, we are admitting repetitions that, in principle, should be avoided. The case of indices without repetitions is discussed in Section 3.2.7

3. Assembly In this phase, we build the *stiffness matrix* $\mathbf{A} \in \mathbb{R}^{m \times N}$ and the *load vector* $\mathbf{f} \in \mathbb{R}^m$ associated with the PG discretization of (2.1), defined as

$$A_{ij} := a(\psi_j, \varphi_{\tau_i}), \quad f_i := \mathcal{F}(\varphi_{\tau_i}), \quad \forall j \in [N], \forall i \in [m]. \tag{3.3}$$

Moreover, the matrix $\mathbf{D} \in \mathbb{R}^{m \times m}$ is a diagonal preconditioner, depending on the vector $\mathbf{p}$ as

$$D_{ik} := \frac{\delta_{ik}}{\sqrt{m p_{\tau_i}}}, \quad \forall i \in [m]. \tag{3.4}$$

Algorithm 3.1

PROCEDURE $\widehat{u}$ = CORSING (N, s, $\boldsymbol{\nu}^N$, $\widehat{\gamma}$, $\widehat{C}$, $\overline{\gamma}$, $\overline{C}$)

1. **Definition of** M **and** m
   - $M \leftarrow \widehat{C} s^{\widehat{\gamma}} N$;
   - $m \leftarrow \overline{C} s^{\overline{\gamma}} \|\boldsymbol{\nu}^{N,M}\|_1 \log(N/s)$;
2. **Test selection**
   - $\mathbf{p} \leftarrow \boldsymbol{\nu}^{N,M} / \|\boldsymbol{\nu}^{N,M}\|_1$;
   - Draw $\tau_1, \ldots, \tau_m$ independently at random from $[M]$ according to $\mathbf{p}$;
3. **Assembly**
   - Build $\mathbf{A}$, $\mathbf{f}$ and $\mathbf{D}$, defined in (3.3) and (3.4), respectively;
4. **Recovery**
   - Find an approximate solution $\widehat{\mathbf{u}}$ to $\min_{\mathbf{v}\in\mathbb{R}^N} \|\mathbf{D}(\mathbf{A}\mathbf{v} - \mathbf{f})\|_2^2$, s.t. $\|\mathbf{v}\|_0 \le s$;
   - $\widehat{u} \leftarrow \Psi\widehat{\mathbf{u}}$.

4. Recovery The discrete CORSING solution $\widehat{\mathbf{u}}$ is defined as an approximate solution to

$$\min_{\mathbf{v}\in\mathbb{R}^N} \|\mathbf{D}(\mathbf{A}\mathbf{v} - \mathbf{f})\|_2^2, \quad \text{s.t.} \quad \|\mathbf{v}\|_0 \le s, \tag{3.5}$$

where $\|\cdot\|_0$ is the ℓ^0-*norm*, defined as in (1.1). Consequently, the CORSING solution is defined as $\widehat{u} := \Psi\widehat{\mathbf{u}}$. An equivalent functional formulation of (3.5) is

$$\min_{v\in U_s^N} \sum_{i=1}^{m} \frac{1}{m p_{\tau_i}} [a(v, \varphi_{\tau_i}) - \mathcal{F}(\varphi_{\tau_i})]^2. \tag{3.6}$$

We recall that problem (3.5) is NP-hard, thus computationally intractable. In practice, its solution is approximated employing the greedy algorithm Orthogonal Matching Pursuit (OMP), described in Section 1.2.3.

The reason for this choice is twofold. First, using OMP we can easily control the parameter s, i.e., the sparsity of the compressed solution $\widehat{u}$. Second, the time complexity of the OMP algorithm is easily estimated, namely $\mathcal{O}(smN)$ for basic implementations and $\mathcal{O}(sN \log N)$ using fast transforms. On the contrary, the complexity of ℓ^1-minimization (see Section 1.1.3) depends on the particular algorithm used to solve the corresponding Linear Programming and it is not

easily quantifiable. All the numerical experiments made in this chapter are performed using the OMP-BOX MATLAB® package, version 10 - see [RZE08, Rub09].

In Section 3.2.4, we will carry out a recovery error analysis in the ideal scenario where $\widehat{u}$ is supposed to solve (3.6) exactly. The case of $\widehat{u}$ computed via OMP is discussed in Section 3.2.6.

3.2 Theoretical analysis

3.2.1 Preliminary results

The main statistical tools employed in this chapter are Chernoff bounds for matrices, introduced by H. Chernoff during the early 50's in the scalar form [Che52], and generalized to the matrix setting by R. Ahlswede and A. Winter in 2003 [AW02]. These bounds have been recently generalized in 2012 by J. Tropp in [Tro12].

First, we present the main result employed in our analysis. The proof of the following theorem can be found in [Tro12, Corollary 5.2].

Theorem 3.4 (Matrix Chernoff bounds)**.** *Consider a finite sequence of i.i.d. random, symmetric $s \times s$ real matrices $\mathbf{X}^1, \dots, \mathbf{X}^m$ such that*

$$0 \leq \lambda_{\min}(\mathbf{X}^i) \text{ and } \lambda_{\max}(\mathbf{X}^i) \leq R \quad \text{almost surely}, \quad \forall i \in [m].$$

Define $\overline{\mathbf{X}} := \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} \mathbf{X}^i$, $E_{\min} := \lambda_{\min}(\mathbb{E}[\mathbf{X}^i])$ *and* $E_{\max} := \lambda_{\max}(\mathbb{E}[\mathbf{X}^i])$. *Then,*

$$\mathbb{P}\{\lambda_{\min}(\overline{\mathbf{X}}) \leq (1-\delta)E_{\min}\} \leq s \exp\left(-\frac{m\xi_\delta E_{\min}}{R}\right), \quad \forall \delta \in [0,1], \tag{3.7}$$

$$\mathbb{P}\{\lambda_{\max}(\overline{\mathbf{X}}) \geq (1+\delta)E_{\max}\} \leq s \exp\left(-\frac{m\widetilde{\xi}_\delta E_{\max}}{R}\right), \quad \forall \delta \geq 0,$$

with

$$\xi_\delta := (1-\delta)\log(1-\delta) + \delta, \quad \widetilde{\xi}_\delta := (1+\delta)\log(1+\delta) - \delta. \tag{3.8}$$

Notice that both constants $\xi_\delta, \widetilde{\xi}_\delta \sim \delta^2$ when $\delta \to 0$.

We conclude this section by recalling few results that will be repeatedly used in the next proofs.

Lemma 3.5. *If $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ are symmetric and $\mathbf{B}$ is also positive definite, it holds*

$$\lambda_{\min}(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}}) = \inf_{\mathbf{u}\in\mathbb{R}^d} \frac{\mathbf{u}^\intercal \mathbf{A}\mathbf{u}}{\mathbf{u}^\intercal \mathbf{B}\mathbf{u}}, \tag{3.9}$$

$$\lambda_{\max}(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}}) = \sup_{\mathbf{u}\in\mathbb{R}^d} \frac{\mathbf{u}^\intercal \mathbf{A}\mathbf{u}}{\mathbf{u}^\intercal \mathbf{B}\mathbf{u}}. \tag{3.10}$$

Lemma 3.6. *Consider a generic set X. The infimum and the supremum on X fulfil the following properties*

$$\sup_{x\in X} 1/f(x) = 1/\inf_{x\in X} f(x), \qquad \forall f : X \to (0,+\infty), \quad (3.11)$$

$$\sup_{x\in X} f(x)g(x) \le \sup_{x\in X} f(x) \sup_{x\in X} g(x), \qquad \forall f,g : X \to [0,+\infty), \quad (3.12)$$

$$\inf_{x\in X}(f(x)-g(x)) \ge \inf_{x\in X} f(x) - \sup_{x\in X} g(x), \qquad \forall f,g : X \to \mathbb{R}. \quad (3.13)$$

3.2.2 Non-uniform restricted inf-sup property

This section coincides with the core of the chapter, providing an analysis of the CORSING algorithm.

We fix a subset $\mathcal{S} := \{\sigma_1,\dots,\sigma_s\} \subseteq [N]$ of cardinality s.

We denote the space of vectors of $\mathbb{R}^N$ supported in $\mathcal{S}$ as $\mathbb{R}^N_{\mathcal{S}}$, namely

$$\mathbb{R}^N_{\mathcal{S}} := \{\mathbf{u} \in \mathbb{R}^N : u_j = 0,\ \forall j \notin \mathcal{S}\}.$$

Moreover, we introduce some further notation.

Definition 3.7 (Matrices $\mathbf{K}$, $\mathbf{K}_{\mathcal{S}}$ and $\mathbf{A}_{\mathcal{S}}$)**.** We define the matrix $\mathbf{K} \in \mathbb{R}^{N\times N}$ as

$$K_{jk} := (\psi_j, \psi_k)_U,$$

and its restriction $\mathbf{K}_{\mathcal{S}} \in \mathbb{R}^{s\times s}$ to $\mathcal{S}$ as

$$(K_{\mathcal{S}})_{jk} := (\psi_{\sigma_j}, \psi_{\sigma_k})_U.$$

Moreover, we denote by $\mathbf{A}_{\mathcal{S}} \in \mathbb{R}^{m\times s}$ the submatrix of $\mathbf{A}$ consisting only of the columns with indices in $\mathcal{S}$.

We observe that $\mathbf{K}$ is symmetric and positive definite (s.p.d.) and fulfills

$$\mathbf{u}^\intercal \mathbf{K}\mathbf{u} = \|\Psi\mathbf{u}\|_U^2, \quad \forall \mathbf{u} \in \mathbb{R}^N, \quad (3.14)$$

where the reconstruction operator in (3.14) is implicitly restricted from ℓ^2 to $\mathbb{R}^N$ (equivalently, the vector $\mathbf{u}$ is extended to ℓ^2 by adding zeros for $j > N$). The matrix $\mathbf{K}_{\mathcal{S}}$ is also s.p.d. and satisfies the relation

$$\mathbf{u}_{\mathcal{S}}^\intercal \mathbf{K}_{\mathcal{S}} \mathbf{u}_{\mathcal{S}} = \mathbf{u}^\intercal \mathbf{K}\mathbf{u}, \quad \forall \mathbf{u} \in \mathbb{R}^N_{\mathcal{S}},$$

where $\mathbf{u}_{\mathcal{S}} \in \mathbb{R}^s$ is the restriction of $\mathbf{u}$ to $\mathcal{S}$, namely $(u_{\mathcal{S}})_j = u_{\sigma_j}$, for every $j \in [s]$.

We introduce the Gram matrix $\mathbf{G}^\infty_{\mathcal{S}}$ relative to the restriction of $a(\cdot,\cdot)$ to $U^N_{\mathcal{S}} \times V^\infty$.

Definition 3.8 (Matrix $\mathbf{G}_{\mathcal{S}}^{\infty}$)**.** Define the matrix $\mathbf{G}_{\mathcal{S}}^{\infty} \in \mathbb{R}^{s\times s}$ such that

$$(G_{\mathcal{S}}^{\infty})_{jk} := \sum_{q=1}^{\infty} a(\psi_{\sigma_j}, \varphi_q) a(\psi_{\sigma_k}, \varphi_q), \quad \forall j,k \in [s],$$

where the series are well defined thanks to Hypothesis 2 and $(G_{\mathcal{S}}^{\infty})_{jk} \le \|\boldsymbol{\mu}^N\|_1$, for every $j,k \in [s]$.

The first lemma provides a relation between the inf-sup constant α associated with the bilinear form $a(\cdot,\cdot)$, corresponding to (2.4), and the Gram matrix $\mathbf{G}_{\mathcal{S}}^{\infty}$.

Lemma 3.9. *Suppose that the bilinear form $a(\cdot,\cdot)$ fulfills the inf-sup property* (2.4)*. Then, it holds*

$$\lambda_{\min}(\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}} \mathbf{G}_{\mathcal{S}}^{\infty} \mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}) \ge \alpha^2.$$

Proof. The following chain of inequalities holds

$$\begin{aligned}\alpha &\le \inf_{u\in U} \sup_{v\in V} \frac{a(u,v)}{\|u\|_U \|v\|_V} \le \inf_{u\in U_{\mathcal{S}}^N} \sup_{v\in V} \frac{a(u,v)}{\|u\|_U \|v\|_V} \\ &= \inf_{\mathbf{u}\in\mathbb{R}_{\mathcal{S}}^N} \sup_{\mathbf{v}\in\ell^2} \frac{1}{\|\mathbf{K}^{\frac{1}{2}}\mathbf{u}\|_2 \|\mathbf{v}\|_2} \sum_{q=1}^{\infty} a(\Psi\mathbf{u}, \varphi_q) v_q = \inf_{\mathbf{u}\in\mathbb{R}_{\mathcal{S}}^N} \frac{1}{\|\mathbf{K}^{\frac{1}{2}}\mathbf{u}\|_2} \left[\sum_{q=1}^{\infty} a(\Psi\mathbf{u}, \varphi_q)^2\right]^{\frac{1}{2}}.\end{aligned}$$

The first inequality is property (2.4), while the second inequality follows from taking the infimum over a subset of U. The first equality is obtained by expanding u and v with respect to the bases $\{\psi_j\}_{j\in\mathcal{S}}$ and $\{\varphi_q\}_{q\in\mathbb{N}}$, respectively; moreover, we use relations (3.14) and $\|\mathbf{v}\|_2 = \|v\|_V$ implied by Hypothesis 1. The last equality can be deduced by applying the definition of operator norm

$$\sup_{\mathbf{v}\in\ell^2} \frac{1}{\|\mathbf{v}\|_2} \sum_{q=1}^{\infty} a(\Psi\mathbf{u}, \varphi_q) v_q = \|(a(\Psi\mathbf{u}, \varphi_q))_{q\in\mathbb{N}}\|_{(\ell^2)^*}$$

and by identifying $(\ell^2)^*$ with ℓ^2. Now, since all the quantities involved in the

chain of inequalities are positive, we can square the terms

$$\begin{aligned}
\alpha^2 &\leq \inf_{\mathbf{u}\in\mathbb{R}^N_{\mathcal{S}}} \frac{1}{\mathbf{u}^\intercal \mathbf{K}\mathbf{u}} \sum_{q=1}^{\infty} a(\Psi\mathbf{u},\varphi_q)^2 = \inf_{\mathbf{u}\in\mathbb{R}^s} \frac{1}{\mathbf{u}^\intercal \mathbf{K}_{\mathcal{S}}\mathbf{u}} \sum_{q=1}^{\infty} \left[\sum_{j=1}^{s} u_j a(\psi_{\sigma_j},\varphi_q)\right]^2 \\
&= \inf_{\mathbf{u}\in\mathbb{R}^s} \frac{1}{\mathbf{u}^\intercal \mathbf{K}_{\mathcal{S}}\mathbf{u}} \sum_{q=1}^{\infty} \sum_{j=1}^{s}\sum_{k=1}^{s} u_j u_k a(\psi_{\sigma_j},\varphi_q) a(\psi_{\sigma_k},\varphi_q) \\
&= \inf_{\mathbf{u}\in\mathbb{R}^s} \frac{1}{\mathbf{u}^\intercal \mathbf{K}_{\mathcal{S}}\mathbf{u}} \sum_{j=1}^{s}\sum_{k=1}^{s} u_j u_k \sum_{q=1}^{\infty} a(\psi_{\sigma_j},\varphi_q) a(\psi_{\sigma_k},\varphi_q) \\
&= \inf_{\mathbf{u}\in\mathbb{R}^s} \frac{\mathbf{u}^\intercal \mathbf{G}^{\infty}_{\mathcal{S}}\mathbf{u}}{\mathbf{u}^\intercal \mathbf{K}_{\mathcal{S}}\mathbf{u}} = \lambda_{\min}(\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}} \mathbf{G}^{\infty}_{\mathcal{S}} \mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}).
\end{aligned}$$

We have expanded $\Psi\mathbf{u}$ and identified $\mathbf{u}$ with its restriction to $\mathcal{S}$. Then, we have exchanged the summations thanks to Hypothesis 2 and Fubini-Tonelli's theorem. Successively, we have used the definition of $\mathbf{G}^{\infty}_{\mathcal{S}}$ together with relation (3.9). □

The second lemma provides a recipe on how to choose the truncation level M on the tests, after selecting N and s.

Lemma 3.10. *Under the same hypotheses as in Lemma 3.9, we fix a real number $\widehat{\delta} \in [0,1)$. Then, if $M \in \mathbb{N}$ satisfies the truncation condition*

$$s \sum_{q>M} \mu_q^N \leq \alpha^2 \lambda_{\min}(\mathbf{K}_{\mathcal{S}})\widehat{\delta}, \tag{3.15}$$

the following inequality holds

$$\lambda_{\min}(\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}} \mathbf{G}^{M}_{\mathcal{S}} \mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}) \geq (1-\widehat{\delta})\alpha^2,$$

where $\mathbf{G}^{M}_{\mathcal{S}} \in \mathbb{R}^{s\times s}$ is the truncated version of $\mathbf{G}^{\infty}_{\mathcal{S}}$, namely

$$(G^{M}_{\mathcal{S}})_{jk} := \sum_{q=1}^{M} a(\psi_{\sigma_j},\varphi_q) a(\psi_{\sigma_k},\varphi_q). \tag{3.16}$$

Proof. First, consider the splitting $\mathbf{G}^{\infty}_{\mathcal{S}} = \mathbf{G}^{M}_{\mathcal{S}} + \mathbf{T}^{M}_{\mathcal{S}}$, where $\mathbf{T}^{M}_{\mathcal{S}}$ corresponds to the tail of the series identifying $\mathbf{G}^{\infty}_{\mathcal{S}}$,

$$(T^{M}_{\mathcal{S}})_{jk} = \sum_{q>M} a(\psi_{\sigma_j},\varphi_q) a(\psi_{\sigma_k},\varphi_q).$$

Now, notice that

$$\lambda_{\min}(\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}\mathbf{G}_{\mathcal{S}}^{M}\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}) = \lambda_{\min}(\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}(\mathbf{G}_{\mathcal{S}}^{\infty} - \mathbf{T}_{\mathcal{S}}^{M})\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}) \geq \lambda_{\min}(\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}\mathbf{G}_{\mathcal{S}}^{\infty}\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}) - \lambda_{\max}(\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}\mathbf{T}_{\mathcal{S}}^{M}\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}).$$

The inequality can be proved using Lemma 3.5 and exploiting property (3.13). Applying Lemma 3.9, we obtain

$$\lambda_{\min}(\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}\mathbf{G}_{\mathcal{S}}^{M}\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}) \geq \alpha^2(1 - \lambda_{\max}(\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}\mathbf{T}_{\mathcal{S}}^{M}\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}})/\alpha^2).$$

Thus, the thesis is proved if we bound the maximum eigenvalue of the tail as follows

$$\lambda_{\max}(\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}\mathbf{T}_{\mathcal{S}}^{M}\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}) \leq \widehat{\delta}\alpha^2.$$

For this purpose, we compute

$$\begin{aligned}
\lambda_{\max}(\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}\mathbf{T}_{\mathcal{S}}^{M}\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}) &= \sup_{\mathbf{u}\in\mathbb{R}^s} \frac{\mathbf{u}^\intercal\mathbf{T}_{\mathcal{S}}^{M}\mathbf{u}}{\mathbf{u}^\intercal\mathbf{K}_{\mathcal{S}}\mathbf{u}} \\
&= \sup_{\mathbf{u}\in\mathbb{R}^s} \frac{1}{\mathbf{u}^\intercal\mathbf{K}_{\mathcal{S}}\mathbf{u}} \sum_{j=1}^{s}\sum_{k=1}^{s} u_j u_k \sum_{q>M} a(\psi_{\sigma_j},\varphi_q)a(\psi_{\sigma_k},\varphi_q) \\
&= \sup_{\mathbf{u}\in\mathbb{R}^s} \frac{1}{\mathbf{u}^\intercal\mathbf{K}_{\mathcal{S}}\mathbf{u}} \sum_{q>M} \left[\sum_{j=1}^{s} u_j a(\psi_{\sigma_j},\varphi_q)\right]^2 \\
&\leq \sup_{\mathbf{u}\in\mathbb{R}^s} \frac{\mathbf{u}^\intercal\mathbf{u}}{\mathbf{u}^\intercal\mathbf{K}_{\mathcal{S}}\mathbf{u}} s \sum_{q>M} \mu_q^N = \frac{1}{\lambda_{\min}(\mathbf{K}_{\mathcal{S}})} s \sum_{q>M} \mu_q^N.
\end{aligned}$$

We start from definition (3.10). Then, by exploiting Hypothesis 2 and Fubini-Tonelli's theorem, combined with Cauchy-Schwarz inequality, the definition of $\boldsymbol{\mu}^N$, of (3.9) and of (3.11), we obtain the desired result under hypothesis (3.15). □

This lemma provides a sufficient condition on the truncation parameter M that ensures an arbitrarily small decrease of the inf-sup constant α by a factor $(1-\widehat{\delta})^{\frac{1}{2}}$. Moreover, a value M that fulfills (3.15) always exists thanks to Hypothesis 2. Relation (3.15) can be also interpreted as a sufficient condition for the space V^M to be δ-proximal for $U_{\mathcal{S}}^N$, with constant $\delta = \widehat{\delta}^{\frac{1}{2}}$ (see [DHSW12]).

Now, we prove the main result of this section.

Theorem 3.11 (Non-uniform RISP)**.** *Let the truncation condition in Lemma 3.10 hold. Then, for every* $0 < \varepsilon < 1$ *and* $\overline{\delta} \in [0,1)$*, provided that*

$$m \geq \widetilde{C}_{\mathcal{S}}\, s\|\boldsymbol{\nu}^{N,M}\|_1 \log(s/\varepsilon),$$

where $\widetilde{C}_{\mathcal{S}} := [\xi_{\delta}^{-}(1-\widehat{\delta})\alpha^2\lambda_{\min}(\mathbf{K}_{\mathcal{S}})]^{-1}$ *and* ξ_{δ}^{-} *is defined according to* (3.8)*, the following non-uniform RISP holds with probability greater than or equal to* $1-\varepsilon$

$$\inf_{\mathbf{u}\in\mathbb{R}^s}\sup_{\mathbf{v}\in\mathbb{R}^m}\frac{\mathbf{v}^\intercal\mathbf{D}\mathbf{A}_{\mathcal{S}}\mathbf{u}}{\|\mathbf{K}_{\mathcal{S}}^{\frac{1}{2}}\mathbf{u}\|_2\|\mathbf{v}\|_2} > \widetilde{\alpha} > 0, \tag{3.17}$$

where $\widetilde{\alpha} := (1-\widehat{\delta})^{\frac{1}{2}}(1-\overline{\delta})^{\frac{1}{2}}\alpha$ *and* $\mathbf{D}$ *is defined as in* (3.4).

Proof. The proof is organized as follows. First, we show that the inf-sup in (3.17) can be interpreted as the square root of the minimum eigenvalue of the sample mean of a sequence of certain i.i.d. random matrices $\mathbf{X}^{\tau_1},\dots,\mathbf{X}^{\tau_m}$. Then, we compute the expectation of $\mathbf{X}^{\tau_i}$ and show that the maximum eigenvalue of $\mathbf{X}^{\tau_i}$ is uniformly bounded. Finally, we apply the matrix Chernoff bound (3.7).

Let us discuss each step of the proof in detail. First, we compute

$$\begin{aligned}\inf_{\mathbf{u}\in\mathbb{R}^s}\sup_{\mathbf{v}\in\mathbb{R}^m}\frac{\mathbf{v}^\intercal\mathbf{D}\mathbf{A}_{\mathcal{S}}\mathbf{u}}{\|\mathbf{K}_{\mathcal{S}}^{\frac{1}{2}}\mathbf{u}\|_2\|\mathbf{v}\|_2} &= \inf_{\mathbf{u}\in\mathbb{R}^s}\frac{1}{\|\mathbf{K}_{\mathcal{S}}^{\frac{1}{2}}\mathbf{u}\|_2}\sup_{\mathbf{v}\in\mathbb{R}^m}\frac{\mathbf{v}^\intercal\mathbf{D}\mathbf{A}_{\mathcal{S}}\mathbf{u}}{\|\mathbf{v}\|_2}\\ &= \inf_{\mathbf{u}\in\mathbb{R}^s}\frac{\|\mathbf{D}\mathbf{A}_{\mathcal{S}}\mathbf{u}\|_2}{\|\mathbf{K}_{\mathcal{S}}^{\frac{1}{2}}\mathbf{u}\|_2} = [\lambda_{\min}(\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}\mathbf{A}_{\mathcal{S}}^\intercal\mathbf{D}^2\mathbf{A}_{\mathcal{S}}\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}})]^{\frac{1}{2}}.\end{aligned}$$

The second equality hinges on the definition of operator norm combined with the identification of $(\mathbb{R}^m)^*$ with $\mathbb{R}^m$ while the third one exploits (3.9).

Relying on the following relation,

$$[\mathbf{A}_{\mathcal{S}}^\intercal\mathbf{D}^2\mathbf{A}_{\mathcal{S}}]_{jk} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{p_{\tau_i}}a(\psi_{\sigma_j},\varphi_{\tau_i})a(\psi_{\sigma_k},\varphi_{\tau_i}),$$

we define the matrices $\mathbf{H}^{\tau_i}\in\mathbb{R}^{s\times s}$ with $H_{jk}^{\tau_i} := \frac{1}{p_{\tau_i}}a(\psi_{\sigma_j},\varphi_{\tau_i})a(\psi_{\sigma_k},\varphi_{\tau_i})$ and

$$\mathbf{X}^{\tau_i} := \mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}\mathbf{H}^{\tau_i}\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}},$$

so that

$$\overline{\mathbf{X}} := \frac{1}{m}\sum_{i=1}^{m}\mathbf{X}^{\tau_i} = \mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}\mathbf{A}_{\mathcal{S}}^\intercal\mathbf{D}^2\mathbf{A}_{\mathcal{S}}\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}.$$

Thus, it holds

$$\inf_{\mathbf{u}\in\mathbb{R}^s}\sup_{\mathbf{v}\in\mathbb{R}^m}\frac{\mathbf{v}^\intercal\mathbf{D}\mathbf{A}_{\mathcal{S}}\mathbf{u}}{\|\mathbf{K}_{\mathcal{S}}^{\frac{1}{2}}\mathbf{u}\|_2\|\mathbf{v}\|_2} = [\lambda_{\min}(\overline{\mathbf{X}})]^{\frac{1}{2}}. \tag{3.18}$$

With a view to the Chernoff bounds, we estimate $\mathbb{E}[\mathbf{X}^{\tau_i}]$ and the corresponding minimum eigenvalue. A direct computation yields

$$\mathbb{E}[H_{jk}^{\tau_i}] = \sum_{q=1}^{M}\mathbb{P}\{\tau_i = q\}H_{jk}^{q} = \sum_{q=1}^{M}p_q\frac{1}{p_q}a(\psi_{\sigma_j},\varphi_q)a(\psi_{\sigma_k},\varphi_q) = G_{jk}^{M}.$$

As a consequence, we have

$$\mathbb{E}[\mathbf{X}^{\tau_i}] = \mathbb{E}[\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}\mathbf{H}^{\tau_i}\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}] = \mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}\mathbb{E}[\mathbf{H}^{\tau_i}]\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}} = \mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}\mathbf{G}_{\mathcal{S}}^{M}\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}},$$

i.e., from Lemma 3.10,

$$\lambda_{\min}(\mathbb{E}[\mathbf{X}^{\tau_i}]) \geq (1-\widehat{\delta})\alpha^2. \tag{3.19}$$

Our aim is now to bound $\lambda_{\max}(\mathbf{X}^{\tau_i})$ from above. We have

$$\begin{aligned}
\lambda_{\max}(\mathbf{X}^{\tau_i}) &= \sup_{\mathbf{u}\in\mathbb{R}^s} \frac{\mathbf{u}^\intercal \mathbf{H}^{\tau_i}\mathbf{u}}{\mathbf{u}^\intercal \mathbf{K}_{\mathcal{S}}\mathbf{u}} \leq \sup_{\mathbf{u}\in\mathbb{R}^s} \frac{\mathbf{u}^\intercal\mathbf{u}}{\mathbf{u}^\intercal \mathbf{K}_{\mathcal{S}}\mathbf{u}} \sup_{\mathbf{u}\in\mathbb{R}^s} \frac{\mathbf{u}^\intercal \mathbf{H}^{\tau_i}\mathbf{u}}{\mathbf{u}^\intercal\mathbf{u}} \\
&= [\lambda_{\min}(\mathbf{K}_{\mathcal{S}})]^{-1} \sup_{\mathbf{u}\in\mathbb{R}^s} \frac{1}{\mathbf{u}^\intercal\mathbf{u}} \sum_{j=1}^{s}\sum_{k=1}^{s} u_j u_k \frac{1}{p_{\tau_i}} a(\psi_{\sigma_j},\varphi_{\tau_i})a(\psi_{\sigma_k},\varphi_{\tau_i}) \\
&= [\lambda_{\min}(\mathbf{K}_{\mathcal{S}})]^{-1} \frac{1}{p_{\tau_i}} \sup_{\mathbf{u}\in\mathbb{R}^s} \frac{1}{\mathbf{u}^\intercal\mathbf{u}} \left[\sum_{j=1}^{s} u_j a(\psi_{\sigma_j},\varphi_{\tau_i})\right]^2 \\
&\leq [\lambda_{\min}(\mathbf{K}_{\mathcal{S}})]^{-1} \frac{\|\boldsymbol{\nu}^{N,M}\|_1}{\nu^N_{\tau_i}} \sum_{j=1}^{s} a(\psi_{\sigma_j},\varphi_{\tau_i})^2 \\
&\leq [\lambda_{\min}(\mathbf{K}_{\mathcal{S}})]^{-1} s\|\boldsymbol{\nu}^{N,M}\|_1.
\end{aligned} \tag{3.20}$$

The first line follows from (3.10) and property (3.12). The equalities in the second and in the third line are algebraic manipulations. The fourth line exploits Cauchy-Schwarz inequality combined with definition (3.2) of $\mathbf{p}$, and the last one relies on Hypothesis 3.

Now, we compute the probability of failure of satisfying (3.17), i.e.,

$$\begin{aligned}
&\mathbb{P}\left\{\inf_{\mathbf{u}\in\mathbb{R}^s}\sup_{\mathbf{v}\in\mathbb{R}^m} \frac{\mathbf{v}^\intercal\mathbf{D}\mathbf{A}_{\mathcal{S}}\mathbf{u}}{\|\mathbf{K}_{\mathcal{S}}^{\frac{1}{2}}\mathbf{u}\|_2\|\mathbf{v}\|_2} \leq \widetilde{\alpha}\right\} = \mathbb{P}\left\{\lambda_{\min}(\overline{\mathbf{X}}) \leq (1-\overline{\delta})(1-\widehat{\delta})\alpha^2\right\} \\
&\leq \mathbb{P}\{\lambda_{\min}(\overline{\mathbf{X}}) \leq (1-\overline{\delta})\lambda_{\min}(\mathbb{E}[\mathbf{X}^{\tau_i}])\} \leq s\exp\left(-\frac{m\xi_{\overline{\delta}}\lambda_{\min}(\mathbb{E}[\mathbf{X}^{\tau_i}])}{s\|\boldsymbol{\nu}^{N,M}\|_1[\lambda_{\min}(\mathbf{K}_{\mathcal{S}})]^{-1}}\right) \\
&\leq s\exp\left(-\frac{m\xi_{\overline{\delta}}(1-\widehat{\delta})\alpha^2}{s\|\boldsymbol{\nu}^{N,M}\|_1[\lambda_{\min}(\mathbf{K}_{\mathcal{S}})]^{-1}}\right).
\end{aligned} \tag{3.21}$$

The first equality relies on (3.18) and on the definition of $\widetilde{\alpha}$. The first inequality in the second line hinges on (3.19), while the second inequality is the first matrix Chernoff bound (3.7), where the uniform estimate (3.20) has been employed. The final inequality follows from (3.19).

The thesis is finally proved on estimating that

$$s\exp\left(-\frac{m\xi_{\overline{\delta}}(1-\widehat{\delta})\alpha^2}{s\|\boldsymbol{\nu}^{N,M}\|_1[\lambda_{\min}(\mathbf{K}_{\mathcal{S}})]^{-1}}\right)\leq\varepsilon\Longleftrightarrow m\geq\widetilde{C}_{\mathcal{S}}\,s\|\boldsymbol{\nu}^{N,M}\|_1\log(s/\varepsilon),$$

with $\widetilde{C}_{\mathcal{S}}:=[\xi_{\overline{\delta}}(1-\widehat{\delta})\alpha^2\lambda_{\min}(\mathbf{K}_{\mathcal{S}})]^{-1}$.

□

3.2.3 Uniform restricted inf-sup property

We extend the results in the previous Section to the uniform case, i.e., we aim at proving the RISP over U_s^N, instead of $U_{\mathcal{S}}^N$, for a fixed subset $\mathcal{S}\subseteq[N]$ with $|\mathcal{S}|=s$. For this purpose, we use the non-uniform Theorem 3.11 and a union bound.

First, we recall the definition of the set Σ_s^N of s-sparse vectors of $\mathbb{R}^N$, namely

$$\Sigma_s^N:=\{\mathbf{x}\in\mathbb{R}^N:\|\mathbf{x}\|_0\leq s\}\equiv\bigcup_{\mathcal{S}\subseteq[N];|\mathcal{S}|=s}\mathbb{R}_{\mathcal{S}}^N.$$

The following theorem provides a sufficient condition for the uniform RISP to hold.

Theorem 3.12 (Uniform RISP)**.** *Given $\widehat{\delta}\in[0,1)$, choose $M\in\mathbb{N}$ such that the following truncation condition is fulfilled*

$$s\sum_{q>M}\mu_q^N\leq\alpha^2\kappa_s\widehat{\delta},\tag{3.22}$$

where

$$\kappa_s:=\min_{\mathcal{S}\subseteq[N];|\mathcal{S}|=s}\lambda_{\min}(\mathbf{K}_{\mathcal{S}}).\tag{3.23}$$

Then, for every $0<\varepsilon<1$ and $\overline{\delta}\in[0,1)$, provided

$$m\geq\widetilde{C}_s\,s\|\boldsymbol{\nu}^{N,M}\|_1[s\log(eN/s)+\log(s/\varepsilon)],\tag{3.24}$$

with

$$\widetilde{C}_s:=[\xi_{\overline{\delta}}(1-\widehat{\delta})\alpha^2\kappa_s]^{-1}\tag{3.25}$$

and $\xi_{\overline{\delta}}$ as in (3.8)*, the following uniform s-sparse RISP holds with probability greater than or equal to $1-\varepsilon$*

$$\inf_{\mathbf{u}\in\Sigma_s^N}\sup_{\mathbf{v}\in\mathbb{R}^m}\frac{\mathbf{v}^{\intercal}\mathbf{D}\mathbf{A}\mathbf{u}}{\|\mathbf{K}^{\frac{1}{2}}\mathbf{u}\|_2\|\mathbf{v}\|_2}>\widetilde{\alpha}>0,$$

where $\widetilde{\alpha}:=(1-\widehat{\delta})^{\frac{1}{2}}(1-\overline{\delta})^{\frac{1}{2}}\alpha$.

Proof. First, we define the event where the RISP holds non-uniformly over a single subset $\mathcal{S} \subseteq [N]$ with $|\mathcal{S}| = s$:

$$\Omega_{\mathcal{S}} := \left\{ \omega \in \Omega : \inf_{\mathbf{u} \in \mathbb{R}^s} \sup_{\mathbf{v} \in \mathbb{R}^m} \frac{\mathbf{v}^\intercal \mathbf{D}(\omega) \mathbf{A}_{\mathcal{S}}(\omega) \mathbf{u}}{\|\mathbf{K}_{\mathcal{S}}^{\frac{1}{2}} \mathbf{u}\|_2 \|\mathbf{v}\|_2} > \widetilde{\alpha} \right\},$$

where the dependence of $\mathbf{A}_{\mathcal{S}}$ and $\mathbf{D}$ on ω has been highlighted. Analogously, we define the event where the RISP holds uniformly

$$\Omega_s := \left\{ \omega \in \Omega : \inf_{\mathbf{u} \in \Sigma_s^N} \sup_{\mathbf{v} \in \mathbb{R}^m} \frac{\mathbf{v}^\intercal \mathbf{D}(\omega) \mathbf{A}(\omega) \mathbf{u}}{\|\mathbf{K}^{\frac{1}{2}} \mathbf{u}\|_2 \|\mathbf{v}\|_2} > \widetilde{\alpha} \right\}. \tag{3.26}$$

In particular, the following relation holds

$$\Omega_s = \bigcap_{\mathcal{S} \subseteq [N]; |\mathcal{S}| = s} \Omega_{\mathcal{S}},$$

and, thanks to the subadditivity of $\mathbb{P}$ and De Morgan's laws, we have

$$\mathbb{P}(\Omega_s^c) = \mathbb{P}\left(\left(\bigcap \Omega_{\mathcal{S}} \right)^c \right) = \mathbb{P}\left(\bigcup \Omega_{\mathcal{S}}^c \right) \leq \sum_{\mathcal{S} \subseteq [N]; |\mathcal{S}| = s} \mathbb{P}(\Omega_{\mathcal{S}}^c), \tag{3.27}$$

where the superscript c denotes the complement of a set. Now, the non-uniform inequality (3.21) and the definition (3.23) of κ_s yield the following uniform upper bound

$$\mathbb{P}(\Omega_{\mathcal{S}}^c) \leq s \exp\left(-\frac{m \xi_{\overline{\delta}} (1 - \widehat{\delta}) \alpha^2}{s \|\boldsymbol{\nu}^{N,M}\|_1 [\lambda_{\min}(\mathbf{K}_{\mathcal{S}})]^{-1}} \right) \leq s \exp\left(-\frac{m \xi_{\overline{\delta}} (1 - \widehat{\delta}) \alpha^2}{s \|\boldsymbol{\nu}^{N,M}\|_1 \kappa_s^{-1}} \right). \tag{3.28}$$

Moreover, Stirling's formula furnishes the following upper bound

$$|\{\mathcal{S} \subseteq [N] : |\mathcal{S}| = s\}| = \binom{N}{s} = \frac{N!}{s!(N-s)!} \leq \frac{N^s}{s!} \leq \left(\frac{eN}{s} \right)^s. \tag{3.29}$$

Combining (3.27), (3.28) and (3.29), we finally obtain the uniform estimate

$$\mathbb{P}(\Omega_s^c) \leq \left(\frac{eN}{s} \right)^s s \exp\left(-\frac{m \xi_{\overline{\delta}} (1 - \widehat{\delta}) \alpha^2}{s \|\boldsymbol{\nu}^{N,M}\|_1 \kappa_s^{-1}} \right). \tag{3.30}$$

Simple algebraic manipulations show that the right hand-side of (3.30) is less than or equal to ε if and only if relation (3.24) holds. □

Remark 3.2.1. A lower bound for the quantity κ_s defined in (3.23) is provided by the Riesz constant c_ψ, defined in (3.1). Indeed, Lemma 3.5 yields

$$\lambda_{\min}(\mathbf{K}_{\mathcal{S}}) = \min_{\mathbf{u}\in\mathbb{R}^N_{\mathcal{S}}} \frac{\mathbf{u}^\intercal \mathbf{K}\mathbf{u}}{\mathbf{u}^\intercal \mathbf{u}} = \min_{\mathbf{u}\in\mathbb{R}^N_{\mathcal{S}}} \frac{\|\Psi\mathbf{u}\|_2^2}{\|\mathbf{u}\|_2^2} \geq c_\psi, \quad \forall \mathcal{S} \subseteq [N],$$

where the representation operator Ψ is implicitly restricted from ℓ^2 to $\mathbb{R}^N$. As a consequence, Theorem 3.12 can be restated in a weaker form, replacing κ_s with c_ψ. □

We note that the sufficient condition (3.24) is, in general, too pessimistic. Indeed, in the classical literature on CS (see Section 1.2), the optimal asymptotically dependence of m on s is linear (up to logarithmic factors). This lack of optimality is due to the union bound, that is a very rough estimate. It is possible to achieve the optimal behavior by using more advanced tools, such as those described in Section 1.2.7. This will be investigated in Section 3.2.5.

3.2.4 Recovery error analysis under the RISP

In this section, we deal with the analysis of the recovery error associated with the CORSING procedure. The CORSING solution $\widehat{u}$ is supposed to solve the minimization problem (3.6) exactly. Although this is an ideal scenario, since (3.6) is NP-hard, it may be useful to understand such a situation. Moreover, this theoretical analysis highlights the fundamental role played by the RISP. The case of $\widehat{u}$ approximated via OMP (or ℓ^1-minimization) is discussed in Section 3.2.6.

The recovery error is computed with respect to the trial norm $\|\cdot\|_U$ and corresponds to the quantity $\|\widehat{u} - u\|_U$. Notice that this error is a random variable, depending on the extracted indices $\tau_1, \dots, \tau_m$. Our aim is to compare the recovery error with the best s-term approximation error of the exact solution u in U^N, i.e., the quantity $\|u^s - u\|_U$, where

$$u^s := \arg\min_{w\in U^N_s} \|w - u\|_U. \tag{3.31}$$

Due to the s-sparsity constraint in the recovery procedure (3.5), u^s is the best result that CORSING can ideally provide.[1]

For this purpose, we show that the uniform $2s$-sparse RISP implies a recovery result, depending on a random preconditioned residual (Lemma 3.13), whose second moment is controlled by the square of the best s-term approximation error (Lemma 3.14). Afterwards, in Theorem 3.16, we prove that the best s-term approximation error dominates the first moment of the error associated with a truncated version of the CORSING solution and, finally, we provide a recovery error estimate that holds with high probability in Theorem 3.17.

[1] The quantity in (3.31) is actually a minimum and not an infimum, since the function $w \mapsto \|w - u\|_U$ is convex and U^N_s is a finite union of linear subspaces.

In the following, a key quantity is the preconditioned random residual

$$\mathcal{R}(w) := \left[\frac{1}{m} \sum_{i=1}^{m} \frac{1}{p_{\tau_i}} [a(w, \varphi_{\tau_i}) - \mathcal{F}(\varphi_{\tau_i})]^2 \right]^{\frac{1}{2}}, \quad \forall w \in U. \tag{3.32}$$

Now, we prove the two lemmas.

Lemma 3.13. *If the uniform 2s-sparse RISP*

$$\inf_{\mathbf{u} \in \Sigma_{2s}^N} \sup_{\mathbf{v} \in \mathbb{R}^m} \frac{\mathbf{v}^\intercal \mathbf{D} \mathbf{A} \mathbf{u}}{\|\mathbf{K}^{\frac{1}{2}} \mathbf{u}\|_2 \|\mathbf{v}\|_2} > \widetilde{\alpha} > 0, \tag{3.33}$$

holds, then the CORSING procedure computes a solution $\widehat{u}$ such that

$$\|\widehat{u} - u^s\|_U < \frac{2}{\widetilde{\alpha}} \mathcal{R}(u^s).$$

Proof. Define $\widehat{\mathbf{u}} := \Psi^* \widehat{u}$ and $\mathbf{u}^s := \Psi^* u^s$. Then, casting (3.26) in Ω_{2s}, since $\widehat{\mathbf{u}} - \mathbf{u}^s$ is at most $2s$-sparse and thanks to the RISP property (3.33), and the definition of operator norm, we have

$$\|\widehat{u} - u^s\|_U = \|\mathbf{K}^{\frac{1}{2}} (\widehat{\mathbf{u}} - \mathbf{u}^s)\|_2 < \frac{1}{\widetilde{\alpha}} \sup_{\mathbf{v} \in \mathbb{R}^m} \frac{\mathbf{v}^\intercal \mathbf{D} \mathbf{A} (\widehat{\mathbf{u}} - \mathbf{u}^s)}{\|\mathbf{v}\|_2} = \frac{1}{\widetilde{\alpha}} \|\mathbf{D} \mathbf{A} (\widehat{\mathbf{u}} - \mathbf{u}^s)\|_2.$$

Moreover, the last norm can be bounded as

$$\begin{aligned}
\|\mathbf{D} \mathbf{A} (\widehat{\mathbf{u}} - \mathbf{u}^s)\|_2^2 &= \frac{1}{m} \sum_{i=1}^{m} \frac{1}{p_{\tau_i}} a(\widehat{u} - u^s, \varphi_{\tau_i})^2 \\
&= \frac{1}{m} \sum_{i=1}^{m} \frac{1}{p_{\tau_i}} [a(\widehat{u}, \varphi_{\tau_i}) - \mathcal{F}(\varphi_{\tau_i}) - a(u^s, \varphi_{\tau_i}) + \mathcal{F}(\varphi_{\tau_i})]^2 \\
&\leq \frac{2}{m} \sum_{i=1}^{m} \frac{1}{p_{\tau_i}} \{ [a(\widehat{u}, \varphi_{\tau_i}) - \mathcal{F}(\varphi_{\tau_i})]^2 + [a(u^s, \varphi_{\tau_i}) - \mathcal{F}(\varphi_{\tau_i})]^2 \} \\
&\leq \frac{4}{m} \sum_{i=1}^{m} \frac{1}{p_{\tau_i}} [a(u^s, \varphi_{\tau_i}) - \mathcal{F}(\varphi_{\tau_i})]^2 = 4 \mathcal{R}(u^s)^2,
\end{aligned}$$

where the last inequality exploits the optimality of $\widehat{u}$. □

Lemma 3.14. *The following upper bound holds*

$$\mathbb{E}[\mathcal{R}(u^s)^2] \leq \beta^2 \|u^s - u\|_U^2, \tag{3.34}$$

where β is the continuity constant of $a(\cdot, \cdot)$ defined in (2.3).

Proof. Thanks to (2.1), the residual (3.32) becomes

$$\mathcal{R}(u^s)^2 = \frac{1}{m}\sum_{i=1}^{m} p_{\tau_i}^{-1} a(u^s - u, \varphi_{\tau_i})^2,$$

Thus, in expectation, we obtain

$$\mathbb{E}[\mathcal{R}(u^s)^2] = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[p_{\tau_i}^{-1} a(u^s - u, \varphi_{\tau_i})^2]. \tag{3.35}$$

Each term in the last summation can be bounded as

$$\mathbb{E}[p_{\tau_i}^{-1} a(u^s - u, \varphi_{\tau_i})^2] = \sum_{q=1}^{M} p_q^{-1} a(u^s - u, \varphi_q)^2 p_q \leq \sum_{q=1}^{\infty} a(u^s - u, \varphi_q)^2. \tag{3.36}$$

Now, exploiting Hypothesis 1, we have

$$\begin{aligned} \|a(u^s - u, \cdot)\|_{V^*} &= \sup_{v \in V} \frac{|a(u^s - u, v)|}{\|v\|_V} \\ &= \sup_{\mathbf{v} \in \ell^2} \frac{|\sum_{q=1}^{\infty} v_q a(u^s - u, \varphi_q)|}{\|\mathbf{v}\|_2} = \left[\sum_{q=1}^{\infty} a(u^s - u, \varphi_q)^2\right]^{\frac{1}{2}}. \end{aligned}$$

Plugging this equality and (3.36) in (3.35), and thanks to (2.3), we have

$$\mathbb{E}[\mathcal{R}(u^s)^2] \leq \|a(u^s - u, \cdot)\|_{V^*}^2 \leq \beta^2 \|u^s - u\|_U^2.$$

□

If an upper bound of the form $\|u\|_U \leq \mathcal{K}$ is known, a near-optimal recovery result holds in expectation for a truncation of the CORSING solution. This truncation is obtained through the operator $\mathcal{T}_{\mathcal{K}} : U \to U$ defined as

$$\mathcal{T}_{\mathcal{K}} w := \begin{cases} w & \text{if } \|w\|_U \leq \mathcal{K}, \\ \mathcal{K} w / \|w\|_U & \text{if } \|w\|_U > \mathcal{K}, \end{cases} \qquad \forall w \in U. \tag{3.37}$$

Using (2.1) and (2.4), a possible choice of $\mathcal{K}$ is $\|\mathcal{F}\|_{V^*}/\alpha$.

Then, we have the following lemma whose proof is straightforward.

Lemma 3.15. *$\mathcal{T}_{\mathcal{K}}$ is 1-Lipschitz, with respect to $\|\cdot\|_U$, for every $\mathcal{K} > 0$.*

Employing an argument similar to that used in [CDL13, CCM+15], we show an upper bound to the error associated with the truncated CORSING solution.

Theorem 3.16 (Error estimate in expectation)**.** *Let $\mathcal{K} > 0$ be such that $\|u\|_U \leq \mathcal{K}$. Given $\widehat{\delta} \in [0,1)$, choose $M \in \mathbb{N}$ such that the truncation condition* (3.22) *is fulfilled and fix $\overline{\delta} \in [0,1)$. Then, for every $0 < \varepsilon < 1$, provided*

$$m \geq 2\,\widetilde{C}_{2s}\, s \|\boldsymbol{\nu}^{N,M}\|_1 [2s \log(eN/(2s)) + \log(2s/\varepsilon)], \tag{3.38}$$

with $\widetilde{C}_{2s}$ defined analogously to (3.25) *and $\widetilde{\alpha} = (1-\widehat{\delta})^{\frac{1}{2}}(1-\overline{\delta})^{\frac{1}{2}}\alpha$, the truncated CORSING solution $\mathcal{T}_{\mathcal{K}}\widehat{u}$ fulfills*

$$\mathbb{E}[\|\mathcal{T}_{\mathcal{K}}\widehat{u} - u\|_U] < \left(1 + \frac{2\beta}{\widetilde{\alpha}}\right)\|u^s - u\|_U + 2\mathcal{K}\varepsilon,$$

where β is the continuity constant of $a(\cdot,\cdot)$ defined in (2.3)*.*

Proof. First, recalling the definition (3.26) of the event Ω_s, and considering the partitioning $\Omega = \Omega_{2s} \cup \Omega_{2s}^c$, we have the splitting

$$\mathbb{E}[\|\mathcal{T}_{\mathcal{K}}\widehat{u} - u\|_U] = \int_{\Omega_{2s}} \|\mathcal{T}_{\mathcal{K}}(\widehat{u} - u)\|_U \,\mathrm{d}\,\mathbb{P} + \int_{\Omega_{2s}^c} \|\mathcal{T}_{\mathcal{K}}\widehat{u} - u\|_U \,\mathrm{d}\,\mathbb{P}.$$

Then, the second term is easily bounded as

$$\int_{\Omega_{2s}^c} \|\mathcal{T}_{\mathcal{K}}\widehat{u} - u\|_U \,\mathrm{d}\,\mathbb{P} \leq 2\mathcal{K}\varepsilon.$$

Indeed, thanks to the adopted choice of m, Theorem 3.12 guarantees $\mathbb{P}(\Omega_{2s}^c) \leq \varepsilon$. Moreover, $\|\mathcal{T}_{\mathcal{K}}\widehat{u} - u\|_U \leq 2\mathcal{K}$, since both $\|\mathcal{T}_{\mathcal{K}}\widehat{u}\|_U$ and $\|u\|_U$ are less than or equal to $\mathcal{K}$.

Now, employing Lemma 3.15 and the triangle inequality, we have

$$\int_{\Omega_{2s}} \|\mathcal{T}_{\mathcal{K}}(\widehat{u} - u)\|_U \,\mathrm{d}\,\mathbb{P} \leq \int_{\Omega_{2s}} \|\widehat{u} - u\|_U \,\mathrm{d}\,\mathbb{P} \leq \int_{\Omega_{2s}} \|\widehat{u} - u^s\|_U \,\mathrm{d}\,\mathbb{P} + \int_{\Omega_{2s}} \|u^s - u\|_U \,\mathrm{d}\,\mathbb{P}.$$

The second integral on the right-hand side is less than or equal to the best s-term approximation error $\|u^s - u\|_U$. In order to bound the first integral, we apply Lemmas 3.13 and 3.14, obtaining

$$\int_{\Omega_{2s}} \|\widehat{u} - u^s\|_U \,\mathrm{d}\,\mathbb{P} < \frac{2}{\widetilde{\alpha}} \int_{\Omega_{2s}} \mathcal{R}(u^s) \,\mathrm{d}\,\mathbb{P} \leq \frac{2}{\widetilde{\alpha}} \mathbb{E}[\mathcal{R}(u^s)] \leq \frac{2\beta}{\widetilde{\alpha}} \|u^s - u\|_U,$$

where the last relation follows on applying Jensen's inequality to (3.34). Notice that Lemma 3.13 can be employed since the $2s$-sparse RISP holds on the restricted domain Ω_{2s}. Combining all the inequalities yields the thesis. □

Finally, we provide a recovery estimate in probability. This is asymptotically optimal, but the constant grows like the inverse of the square root of the probability of failure.

Theorem 3.17 (Error estimate in probability)**.** *Given* $\widehat{\delta} \in [0,1)$*, choose* $M \in \mathbb{N}$ *such that the truncation condition* (3.22) *is fulfilled. Then, for every* $0 < \varepsilon < 1$ *and* $\overline{\delta} \in [0,1)$*, provided*

$$m \geq 2\widetilde{C}_{2s}\, s \|\boldsymbol{\nu}^{N,M}\|_1 [2s \log(eN/(2s)) + \log(2s/\varepsilon)],$$

with $\widetilde{C}_{2s}$ *defined analogously to* (3.25)*, with probability greater than or equal to* $1 - 2\varepsilon$*, the* CORSING *procedure computes a solution* $\widehat{u}$ *such that*

$$\|\widehat{u} - u\|_U < \left(1 + \frac{2\beta}{\widetilde{\alpha}\sqrt{\varepsilon}}\right) \|u^s - u\|_U,$$

where $\widetilde{\alpha} := (1-\widehat{\delta})^{\frac{1}{2}}(1-\overline{\delta})^{\frac{1}{2}}\alpha$ *and* β *is the continuity constant of* $a(\cdot,\cdot)$ *defined in* (2.3)*.*

Proof. Define $e_s := \|u^s - u\|_U$ and the random variables $Z := \|\widehat{u} - u\|_U$ and $Z_s := \|\widehat{u} - u^s\|_U$. Moreover, consider the quantity

$$b_s := \left(1 + \frac{2\beta}{\widetilde{\alpha}\sqrt{\varepsilon}}\right) e_s. \tag{3.39}$$

The goal is to show that $\mathbb{P}\{Z \geq b_s\} \leq 2\varepsilon$. The triangle inequality implies $Z \leq Z_s + e_s$. Thus,

$$\mathbb{P}\{Z \geq b_s\} \leq \mathbb{P}\{Z_s \geq b_s - e_s\}.$$

Moreover, defining the event Ω_{2s} according to (3.26) and denoting by I_A the indicator function of a generic set A, we have

$$\begin{aligned}\mathbb{P}\{Z_s \geq b_s - e_s\} = \mathbb{E}[I_{\{Z_s \geq b_s - e_s\}}] &= \int_{\Omega_{2s}} I_{\{Z_s \geq b_s - e_s\}}\, \mathrm{d}\,\mathbb{P} + \int_{\Omega_{2s}^c} I_{\{Z_s \geq b_s - e_s\}}\, \mathrm{d}\,\mathbb{P} \\ &\leq \int_{\Omega_{2s}} I_{\{Z_s \geq b_s - e_s\}}\, \mathrm{d}\,\mathbb{P} + \mathbb{P}\{\Omega_{2s}^c\}.\end{aligned}$$

Theorem 3.12 implies $\mathbb{P}\{\Omega_{2s}^c\} \leq \varepsilon$. Moreover, employing Lemmas 3.13 and 3.14, we can bound the first integral as

$$\begin{aligned}\int_{\Omega_{2s}} I_{\{Z_s \geq b_s - e_s\}}\, \mathrm{d}\,\mathbb{P} &\leq \int_{\Omega_{2s}} I_{\{(2/\widetilde{\alpha})\mathcal{R}(u^s) > b_s - e_s\}}\, \mathrm{d}\,\mathbb{P} \\ &< \mathbb{E}\left[\frac{4\mathcal{R}(u^s)^2}{\widetilde{\alpha}^2 (b_s - e_s)^2}\right] \leq \frac{4\beta^2 e_s^2}{\widetilde{\alpha}^2 (b_s - e_s)^2} = \varepsilon,\end{aligned}$$

where the last equality follows from (3.39).

□

We conclude this section with a useful corollary dealing with a particular truncation condition. In practice, this corollary provides sufficient conditions for Theorem 3.16 to hold. We will apply this result to some examples in Section 3.3.

Corollary 3.18. *Suppose that there exist two positive constants $\widehat{K}$ and $\widehat{\gamma}$ such that*

$$\sum_{q>M} \mu_q^N \leq \widehat{K}\left(\frac{N}{M}\right)^{1/\widehat{\gamma}}, \quad \forall M \in \mathbb{N}. \tag{3.40}$$

Then, for every $\varepsilon \in (0, 2^{-1/3}]$ and for $s \leq 2N/e$ there exist two positive constants $\widehat{C}$ and $\overline{C}$ such that, for

$$M \geq \widehat{C} s^{\widehat{\gamma}} N \quad \textit{and} \quad m \geq \overline{C} s \|\boldsymbol{\nu}^{N,M}\|_1 [s \log(N/s) + \log(s/\varepsilon)], \tag{3.41}$$

the CORSING solution $\widehat{u}$ fulfills

$$\mathbb{E}[\|\mathcal{T}_{\mathcal{K}}\widehat{u} - u\|_U] < \left(1 + \frac{4\beta}{\alpha}\right)\|u^s - u\|_U + 2\mathcal{K}\varepsilon,$$

for every $\mathcal{K} > 0$ such that $\|u\|_U \leq \mathcal{K}$, with $\mathcal{T}_{\mathcal{K}}$ defined as in (3.37) *and where α and β are defined by* (2.4) *and* (2.3)*, respectively. In particular, two possible upper bounds for the constants $\widehat{C}$ and $\overline{C}$ are*

$$\widehat{C} \leq \left(\frac{2\widehat{K}}{\kappa_s \alpha^2}\right)^{\widehat{\gamma}} \quad \textit{and} \quad \overline{C} \leq \frac{105}{\alpha^2},$$

respectively, with κ_s defined in (3.23).

Proof. The idea is to choose $\overline{\delta} = \widehat{\delta} = 1/2$ and, as anticipated, to apply Theorem 3.16. First, notice that assumption (3.40) is consistent with Hypothesis 2, on passing to the limit for $M \to +\infty$. In view of Theorem 3.16, we show that the second inequality in (3.41) implies (3.38) with a suitable choice of $\overline{C}$. Moreover, the truncation condition (3.22), on which Theorem 3.16 relies on, is implied by

$$s\widehat{K}\left(\frac{N}{M}\right)^{1/\widehat{\gamma}} \leq \frac{\alpha^2 \kappa_s}{2},$$

which, in turn, is equivalent to

$$M \geq \left(\frac{2\widehat{K}}{\kappa_s \alpha^2}\right)^{\widehat{\gamma}} s^{\widehat{\gamma}} N.$$

Moreover, thanks to the assumptions on ε and s, we have

$$\begin{aligned} &\varepsilon \leq 2^{-1/3} \Longrightarrow \log(2s/\varepsilon) \leq 4\log(s/\varepsilon), \\ &s \leq 2N/e \Longrightarrow \log(eN/(2s)) \leq 2\log(N/s). \end{aligned}$$

Thus, recalling the right-hand side of (3.38), we have

$$2\,\widetilde{C}_{2s}\,s\|\boldsymbol{\nu}^{N,M}\|_1[2s\log(eN/(2s))+\log(2s/\varepsilon)] \\ \leq 8\,\widetilde{C}_{2s}\,s\|\boldsymbol{\nu}^{N,M}\|_1[s\log(N/s)+\log(s/\varepsilon)],$$

where $\widetilde{C}_{2s}$ is defined analogously to (3.25). In particular, if $\overline{C}$ in (3.41) is chosen such that

$$\overline{C} \leq 8\,\widetilde{C}_{2s} = \frac{32}{(1-\log 2)\alpha^2} \leq \frac{105}{\alpha^2},$$

then (3.38) holds. Moreover, relation $\widetilde{\alpha} = (1-\widehat{\delta})^{\frac{1}{2}}(1-\overline{\delta})^{\frac{1}{2}}\alpha$ yields $\widetilde{\alpha} = \frac{1}{2}\alpha$, so that the quantity $2\beta/\widetilde{\alpha}$ in Theorem 3.16 can be replaced by $4\beta/\alpha$. □

We conclude this section with some technical clarifications.

Remark 3.2.2. The assumptions $\varepsilon \leq 2^{-1/3} \approx 0.79$ and $s \leq 2N/e \approx 0.74N$ made in Corollary 3.18 are quite weak and they are chosen in such a way that the upper bounds to $\widehat{C}$ and $\overline{C}$ are easy to derive. Of course, more restrictive hypotheses on ε and s would give sharper upper bounds for the asymptotic constants. Moreover, the parameters $\widehat{\delta}$ and $\overline{\delta}$ could be chosen differently from $\overline{\delta} = \widehat{\delta} = 1/2$ and this would lead to different values for the constant in the recovery error estimate. □

Remark 3.2.3. If $\varepsilon \geq s^{s+1}/N^s$, then $s\log(N/s)+\log(s/\varepsilon) \leq 2s\log(N/s)$ and the term $\log(s/\varepsilon)$ disappears from the inequality on m by doubling the constant $\overline{C}$, giving the trend

$$m \geq \overline{C}\|\boldsymbol{\nu}^{N,M}\|_1 s^2\log(N/s),$$

claimed in Algorithm 3.1. This assumption on ε is not restrictive, since $s \ll N$ guarantees $s^{s+1}/N^s \ll 1$. □

Remark 3.2.4. A result analogous to Corollary 3.18 holds in probability by resorting to Theorem 3.17 instead of Theorem 3.16 in the proof. □

Remark 3.2.5. Throughout all this chapter, the reconstruction and decomposition operators Ψ and Ψ^* (see Definition 3.2) are restricted to $\mathbb{R}^N \subseteq \ell^2$ and $U^N \subseteq U$, respectively. Therefore, only the operators

$$\Psi|_{\mathbb{R}^N} : \mathbb{R}^N \to U^N \quad \text{and} \quad \Psi^*|_{U^N} : U^N \to \mathbb{R}^N$$

need to be well-defined and, consequently, the assumption that $\{\psi_j\}_{j\in\mathbb{N}}$ be a Riesz basis can be weakened. It is sufficient to suppose $\{\psi_j\}_{j\in[N]}$ to be a Riesz basis. Recalling relation (3.14) and Lemma 3.5, the Riesz constants can be explicitly computed as $c_\psi = \lambda_{\min}(\mathbf{K})$ and $C_\psi = \lambda_{\max}(\mathbf{K})$. □

Remark 3.2.6. Finally, we add a technical clarification with respect to the CS framework. All the recovery results shown in this section are *nonuniform* in the sense of *instance optimality* (see [FR13, Chapter 11]), since, whenever we state a result in probability, we implicitly fix the operator $\mathcal{F}$, and thus the exact solution u. □

3.2.5 Restricted Isometry Property

We present an argument to prove that a linear dependence between m and s (up to logarithmic factors) is sufficient to guarantee the RIP (and, thus, the RISP) with high probability. The principal tool employed here is Theorem 1.21.

Theorem 3.19. *Let $s, N \in \mathbb{N}$, with $s < N$. Fix $\widehat{\delta} \in (0,1)$ and suppose the truncation condition* (3.15) *to be fulfilled with $\mathcal{S} = [N]$.*

Then, for every

$$\delta \in \Big(1 - (1-\widehat{\delta})\frac{c_\psi \alpha^2}{C_\psi \beta^2}, 1\Big), \tag{3.42}$$

there exists a universal constant C such that, provided

$$m \geq \widetilde{C}_{N,M} s \log^3(s)\log(N),$$

and $s \geq \widetilde{C}_{N,M}\log(N)$, where

$$\widetilde{C}_{N,M} = C \max\{\|\boldsymbol{\nu}^{N,M}\|_1, C_\psi\beta^2\} C_\psi^{-1}\beta^{-2}\left(\delta - 1 + (1-\widehat{\delta})\frac{c_\psi\alpha^2}{C_\psi\beta^2}\right)^{-2},$$

it holds

$$\mathbb{P}\{C_\psi^{-1/2}\beta^{-1}\mathbf{DA} \in RIP(\delta, s)\} \geq 1 - N^{-\log^3(s)}.$$

Proof. This theorem is a direct application of Theorem 1.21, where the matrix $\mathbf{B}$ is the stiffness matrix associated with the $M \times N$ linear system of the full-PG discretization, namely,

$$B_{qj} := a(\psi_j, \varphi_q), \quad \forall j \in [N], \forall q \in [M].$$

A direct computation immediately shows that

$$\mathbf{B}^\intercal\mathbf{B} = \mathbf{G}^M_{[N]},$$

where $\mathbf{G}^M_{[N]}$ is defined according to (3.16), with $\mathcal{S} = [N]$. Then, we have

$$\begin{aligned}\lambda_{\min}(\mathbf{G}^M_{[N]}) &= \min_{\mathbf{u}\in\mathbb{R}^N} \frac{\mathbf{u}^\intercal\mathbf{G}^M_{[N]}\mathbf{u}}{\mathbf{u}^\intercal\mathbf{u}} \geq c_\psi \min_{\mathbf{u}\in\mathbb{R}^N}\frac{\mathbf{u}^\intercal\mathbf{G}^M_{[N]}\mathbf{u}^\intercal}{\mathbf{u}^\intercal\mathbf{K}\mathbf{u}} \\ &= \lambda_{\min}(\mathbf{K}^{-\frac{1}{2}}\mathbf{G}^M_{[N]}\mathbf{K}^{-\frac{1}{2}}) \geq (1-\widehat{\delta})c_\psi\alpha^2,\end{aligned}$$

where the equalities are due to Lemma 3.5, the second relation is implied by (3.1) and (3.14), and the last one follows combining Lemma 3.10 with the estimate $\lambda_{\min}(\mathbf{K}) \geq c_\psi$ (see also Remark 3.2.1).

Moreover, exploiting the continuity (2.3) of $a(\cdot,\cdot)$, relation (3.1), and employing Lemma 3.5, we obtain

$$\lambda_{\max}(\mathbf{G}^M_{[N]}) = \sup_{\mathbf{u}\in\mathbb{R}^N} \frac{\mathbf{u}^\intercal \mathbf{G}^M_{[N]}\mathbf{u}}{\mathbf{u}^\intercal\mathbf{u}} = \sup_{\mathbf{u}\in\mathbb{R}^N} \frac{1}{\|\mathbf{u}\|_2^2} \sum_{j\in[N]}\sum_{k\in[N]} u_j u_k \sum_{q\in[M]} a(\psi_j,\varphi_q)a(\psi_k,\varphi_q)$$
$$\leq \sup_{u\in U^N} \frac{C_\psi}{\|u\|_U^2} \sum_{q\in[M]} a(u,\varphi_q)^2 \leq C_\psi \sup_{u\in U^N} \frac{\|a(u,\cdot)\|_{V^*}^2}{\|u\|_U^2} \leq C_\psi \beta^2.$$

The thesis is now implied by Theorem 1.21, with $r = (1-\widehat{\delta})c_\psi\alpha^2$ and $R = C_\psi\beta^2$. □

Theorem 3.19 has several important consequences. First, we observe that $C_\psi^{-1/2}\beta^{-1}\mathbf{DA} \in \mathrm{RIP}(\delta, s)$ is equivalent to

$$(1-\delta)\|\mathbf{u}\|_2^2 \leq \|C_\psi^{-1/2}\beta^{-1}\mathbf{DAu}\|_2^2 \leq (1+\delta)\|\mathbf{u}\|_2^2, \quad \forall \mathbf{u}\in\Sigma_s^N,$$

that, in turn, implies

$$C_\psi^{1/2}\beta(1-\delta)^{\frac{1}{2}} \leq \inf_{\mathbf{u}\in\Sigma_s^N} \frac{\|\mathbf{DAu}\|_2}{\|\mathbf{u}\|_2} = \inf_{\mathbf{u}\in\Sigma_s^N}\sup_{\mathbf{v}\in\mathbb{R}^m} \frac{\mathbf{v}^\intercal\mathbf{DAu}}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2},$$

i.e., the s-sparse RISP with constant $\widehat{\alpha} = \beta(1-\delta)^{1/2}$. Notice that, thanks to (3.42), it holds $0 < \widehat{\alpha} < (1-\widehat{\delta})^{1/2}c_\psi\alpha$.

3.2.6 Recovery error analysis under the RIP

The RIP result stated in Theorem 3.19 implies important consequences for the recovery error analysis of CORSING when problem (3.5) is approximated via OMP. Employing Theorem 1.13, we prove a result analogous to Lemma 3.13, where the RISP is replaced by the RIP.

Lemma 3.20. *There exists $K\in\mathbb{N}$, $C>0$ and $\delta\in(0,1)$ such that, for every $s\in\mathbb{N}$, if*

$$C_\psi^{-1/2}\beta^{-1}\mathbf{DA} \in \mathit{RIP}(\delta,(K+1)s),$$

the OMP algorithm computes in Ks iterations a CORSING solution $\widehat{u}$ that fulfills

$$\|\widehat{u}-u^s\|_U \leq \sqrt{\frac{2(1+C)}{\beta^2(1-\delta)}}\mathcal{R}(u^s),$$

where $\mathcal{R}(u^s)$ is defined as in (3.32)*. The constants K, C and δ are the same as in Theorem 1.13.*

Proof. Define $\widehat{\mathbf{u}} := \Psi^* \widehat{u}$ and $\mathbf{u}^s := \Psi^* u^s$ and consider the constants K, C, and δ as in Theorem 1.13. Then, using the RIP$(\delta, (K+1)s)$, the fact that $\widehat{\mathbf{u}} - \mathbf{u}^s$ is $(K+1)s$-sparse, and employing relation (3.1), we estimate

$$\|\widehat{u} - u^s\|_U^2 \leq C_\psi \|\widehat{\mathbf{u}} - \mathbf{u}^s\|_2^2 \leq \frac{C_\psi}{C_\psi \beta^2 (1-\delta)} \|\mathbf{D}\mathbf{A}(\widehat{\mathbf{u}} - \mathbf{u}^s)\|_2^2.$$

Then, analogously to the proof of Lemma 3.13, we estimate

$$\begin{aligned}\|\mathbf{D}\mathbf{A}(\widehat{\mathbf{u}} - \mathbf{u}^s)\|_2^2 &\leq \frac{2}{m} \sum_{i=1}^m \frac{1}{p_{\tau_i}} \{[a(\widehat{u}, \varphi_{\tau_i}) - \mathcal{F}(\varphi_{\tau_i})]^2 + [a(u^s, \varphi_{\tau_i}) - \mathcal{F}(\varphi_{\tau_i})]^2\} \\ &\leq (1+C) \frac{2}{m} \sum_{i=1}^m \frac{1}{p_{\tau_i}} [a(u^s, \varphi_{\tau_i}) - \mathcal{F}(\varphi_{\tau_i})]^2 = 2(1+C)\mathcal{R}(u^s)^2,\end{aligned}$$

where the second inequality is implied by Theorem 1.13 and the last equality relies on (3.32). □

Remark 3.2.7. Let $s = Ks'$, with $s' \in \mathbb{N}$. Then, Lemma 3.20 admits an equivalent formulation where $\widehat{u}$ is s-sparse and the RIP$(\delta, (K+1)s)$ is replaced by the RIP$(\delta, s + s/K)$. □

Following the same roadmap as in Section 3.2.4, Theorem 3.19 and Lemma 3.20 can be combined with Lemma 3.14 to prove estimates in expectation and in probability. In particular, by letting $\widehat{\delta} \to 0$ in (3.42), we obtain an extra condition on the constant δ in Lemma 3.20, i.e.,

$$\delta > 1 - \frac{c_\psi \alpha^2}{C_\psi \beta^2}. \tag{3.43}$$

As discussed in [CDD15, Section 1], the constants K, C and δ of Lemma 3.20 are coupled. Therefore, a proper choice of K and C is able to guarantee (3.43). However, in order to produce quantitative recovery results, the relation linking the three constants should be made explicit.

Finally, thanks to Theorem 3.19, we can apply all the recovery results of CS based on the RIP. For example, resorting to Proposition 1.6, we can recover the best s-term approximation of the exact solution $\mathbf{u}$ using the (P_1) sparse optimization program[2]

$$\widehat{\mathbf{u}} := \arg\min_{\mathbf{w} \in \mathbb{R}^N} \|\mathbf{w}\|_1 \quad \text{s.t. } \beta^{-1} C_\psi^{-1/2} \mathbf{D}\mathbf{A}\mathbf{w} = \beta^{-1} C_\psi^{-1/2} \mathbf{D}\mathbf{f},$$

[2] The exact equality constraint corresponds to an idealistic scenario, but it is useful to understand the situation qualitatively. In general, one should apply recovery results for the $(\mathrm{P}_1^\varepsilon)$ problem

$$\widehat{\mathbf{u}} := \arg\min_{\mathbf{w} \in \mathbb{R}^N} \|\mathbf{w}\|_1 \quad \text{s.t. } \|\beta^{-1} C_\psi^{-1/2} \mathbf{D}(\mathbf{A}\mathbf{w} - \mathbf{f})\|_2 \leq \varepsilon,$$

such as, e.g., [FR13, Theorem 6.12].

up to an error

$$\|\widehat{\mathbf{u}} - \mathbf{u}\|_2 \lesssim s^{-\frac{1}{2}} \sigma_s(\mathbf{u})_1.$$

Then, the crucial point is to characterize the ℓ^1-norm of $\mathbf{u} = \Psi^* u$ at a function level. This can be done, e.g., when $\{\psi_j\}_{j\in\mathbb{N}}$ is a wavelet family thanks to the *Besov spaces* (see [DeV98, Section 7.3]).

3.2.7 Avoiding repetitions during the test selection

We conclude the theoretical analysis by showing how the test selection step in Algorithm 3.1 can be slightly optimized by avoiding repetitions of the indices τ_i, i.e., performing the random drawings *without* replacement, as we did in the previous chapter (see Algorithm 2.1). Indeed, if the presence of repeated test indices is avoided, the resulting stiffness matrix $\mathbf{A}$ is guaranteed to have rank m, i.e., the amount of information contained in the CORSING discretization $\mathbf{Au} = \mathbf{f}$ is optimal.

Due to the assumption of independence of $\tau_1, \dots, \tau_m$, the theoretical results of this chapter apply to the case of a test selection *with* replacement. Indeed, in the case *without* replacement, $\tau_1, \dots, \tau_m$ become dependent. In [Tro11], the Chernoff bounds are generalized to the case *without* replacement and with a selection made according to a *uniform* probability density. In our setting, the selection is based on a *non-uniform* probability density $\mathbf{p}$ and, unfortunately, the Chernoff bounds have not been generalized to this case, so far.

Nevertheless, this problem can be overcome by suitably changing the preconditioner $\mathbf{D}$. The test selection procedure can be performed as follows:

1. keep drawing indices τ_i from $[M]$ according to $\mathbf{p}$ independently at random until m of them are distinct;
2. the result will be a set $\mathcal{T} := \{\tau_1, \dots, \tau_{m'}\}$ of m' possibly repeated indices and a subset $\widetilde{\mathcal{T}} := \{\widetilde{\tau}_1, \dots, \widetilde{\tau}_m\}$ of m distinct indices, with $m \le m'$;
3. define the numbers of repetitions $r_i := |\{k \in [m'] : \tau_k = \widetilde{\tau}_i\}|$, for every $i \in [m]$.

The CORSING procedure can be applied using only the collection of non-repeated indices $\widetilde{\tau}_1, \dots, \widetilde{\tau}_m$. The resulting stiffness matrix $\widetilde{\mathbf{A}} \in \mathbb{R}^{m\times N}$ and load vector $\widetilde{\mathbf{f}} \in \mathbb{R}^m$ are defined analogously to the standard case

$$\widetilde{A}_{ij} := a(\psi_j, \varphi_{\widetilde{\tau}_i}), \quad \widetilde{f}_i = \mathcal{F}(\varphi_{\widetilde{\tau}_i}), \quad \forall i \in [m], \forall j \in [N],$$

whereas the definition of the diagonal preconditioner $\widetilde{\mathbf{D}} \in \mathbb{R}^{m\times m}$ is modified as follows

$$\widetilde{D}_{ik} := \delta_{ik} \sqrt{\frac{r_i}{m' p_{\widetilde{\tau}_i}}}, \quad \forall i, k \in [m].$$

We show that the non-uniform RISP still holds when the test are not repeated, proving a result analogous to Theorem 3.11.

Theorem 3.21. *Given $\widehat{\delta} \in [0,1)$, choose $M \in \mathbb{N}$ such that the local a-coherence $\boldsymbol{\mu}^N$ fulfills the truncation condition* (3.15)*. Then, for every $\varepsilon > 0$ and $\overline{\delta} \in [0,1)$, provided*

$$m \geq \widetilde{C}_{\mathcal{S}}\, s \|\boldsymbol{\nu}^{N,M}\|_1 \log(s/\varepsilon),$$

where $\widetilde{C}_{\mathcal{S}} := [\xi_{\overline{\delta}}(1-\widehat{\delta})\alpha^2 \lambda_{\min}(\mathbf{K}_{\mathcal{S}})]^{-1}$ and $\xi_{\overline{\delta}}$ is defined according to (3.8)*, the following non-uniform RISP holds with probability greater than or equal to $1-\varepsilon$*

$$\inf_{\mathbf{u}\in\mathbb{R}^s} \sup_{\mathbf{v}\in\mathbb{R}^m} \frac{\mathbf{v}^\intercal \widetilde{\mathbf{D}}\widetilde{\mathbf{A}}_{\mathcal{S}}\mathbf{u}}{\|\mathbf{K}_{\mathcal{S}}^{\frac{1}{2}}\mathbf{u}\|_2 \|\mathbf{v}\|_2} > \widetilde{\alpha} > 0,$$

where $\widetilde{\alpha} := (1-\widehat{\delta})^{\frac{1}{2}}(1-\overline{\delta})^{\frac{1}{2}}\alpha$.

Proof. The proof is identical to that of Theorem 3.11, thus we report only the different parts. First, notice that

$$\inf_{\mathbf{u}\in\mathbb{R}^s} \sup_{\mathbf{v}\in\mathbb{R}^m} \frac{\mathbf{v}^\intercal \widetilde{\mathbf{D}}\widetilde{\mathbf{A}}_{\mathcal{S}}\mathbf{u}}{\|\mathbf{K}_{\mathcal{S}}^{\frac{1}{2}}\mathbf{u}\|_2 \|\mathbf{v}\|_2} = \lambda_{\min}(\mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}} \widetilde{\mathbf{A}}_{\mathcal{S}}^\intercal \widetilde{\mathbf{D}}^2 \widetilde{\mathbf{A}}_{\mathcal{S}} \mathbf{K}_{\mathcal{S}}^{-\frac{1}{2}}).$$

Moreover, it turns out that

$$\widetilde{\mathbf{A}}_{\mathcal{S}}^\intercal \widetilde{\mathbf{D}}^2 \widetilde{\mathbf{A}}_{\mathcal{S}} = \mathbf{A}_{\mathcal{S}}^\intercal \mathbf{D}^2 \mathbf{A}_{\mathcal{S}},$$

where $\mathbf{A} \in \mathbb{R}^{m'\times N}$ and $\mathbf{D} \in \mathbb{R}^{m'\times m'}$ are the stiffness matrix and the preconditioner relative to the indices $\tau_1, \dots, \tau_{m'}$ considered with repetitions. Thus, we have

$$\inf_{\mathbf{u}\in\mathbb{R}^s} \sup_{\mathbf{v}\in\mathbb{R}^m} \frac{\mathbf{v}^\intercal \widetilde{\mathbf{D}}\widetilde{\mathbf{A}}_{\mathcal{S}}\mathbf{u}}{\|\mathbf{K}_{\mathcal{S}}^{\frac{1}{2}}\mathbf{u}\|_2 \|\mathbf{v}\|_2} = \inf_{\mathbf{u}\in\mathbb{R}^s} \sup_{\mathbf{v}\in\mathbb{R}^{m'}} \frac{\mathbf{v}^\intercal \mathbf{D}\mathbf{A}_{\mathcal{S}}\mathbf{u}}{\|\mathbf{K}_{\mathcal{S}}^{\frac{1}{2}}\mathbf{u}\|_2 \|\mathbf{v}\|_2}.$$

Consequently, the probability of failure is

$$\begin{aligned}
\mathbb{P}\left\{ \inf_{\mathbf{u}\in\mathbb{R}^s} \sup_{\mathbf{v}\in\mathbb{R}^m} \frac{\mathbf{v}^\intercal \widetilde{\mathbf{D}}\widetilde{\mathbf{A}}_{\mathcal{S}}\mathbf{u}}{\|\mathbf{K}_{\mathcal{S}}^{\frac{1}{2}}\mathbf{u}\|_2 \|\mathbf{v}\|_2} \leq \widetilde{\alpha} \right\} &= \mathbb{P}\left\{ \inf_{\mathbf{u}\in\mathbb{R}^s} \sup_{\mathbf{v}\in\mathbb{R}^{m'}} \frac{\mathbf{v}^\intercal \mathbf{D}\mathbf{A}_{\mathcal{S}}\mathbf{u}}{\|\mathbf{K}_{\mathcal{S}}^{\frac{1}{2}}\mathbf{u}\|_2 \|\mathbf{v}\|_2} \leq \widetilde{\alpha} \right\} \\
&= \mathbb{P}\left\{ \lambda_{\min}(\overline{\mathbf{X}}) \leq (1-\overline{\delta})(1-\widehat{\delta})\alpha^2 \right\} \\
&\leq \mathbb{P}\{\lambda_{\min}(\overline{\mathbf{X}}) \leq (1-\overline{\delta})\lambda_{\min}(\mathbb{E}[\mathbf{X}^{\tau_i}])\}.
\end{aligned}$$

Conditioning the probability given the events $\{m' = k\}$, for $k \in \mathbb{N}$, and employing the matrix Chernoff bounds (Theorem 3.4) with the probabilities $\mathbb{P}\{\cdot \,|m' = k\}$, we obtain

$$\begin{aligned}
\mathbb{P}\left\{\frac{\lambda_{\min}(\overline{\mathbf{X}})}{\lambda_{\min}(\mathbb{E}[\mathbf{X}^{\tau_i}])} \leq (1-\overline{\delta})\right\} &= \sum_{k\geq m} \mathbb{P}\left\{\frac{\lambda_{\min}(\overline{\mathbf{X}})}{\lambda_{\min}(\mathbb{E}[\mathbf{X}^{\tau_i}])} \leq (1-\overline{\delta})\middle| m' = k\right\}\mathbb{P}\{m' = k\} \\
&\leq \sum_{k\geq m} s\exp\left(-\frac{k\xi_{\overline{\delta}}\lambda_{\min}(\mathbb{E}[\mathbf{X}^{\tau_i}])}{s\|\boldsymbol{\nu}^{N,M}\|_1[\lambda_{\min}(\mathbf{K}_{\mathcal{S}})]}\right)\mathbb{P}\{m' = k\} \\
&\leq s\exp\left(-\frac{m\xi_{\overline{\delta}}\lambda_{\min}(\mathbb{E}[\mathbf{X}^{\tau_i}])}{s\|\boldsymbol{\nu}^{N,M}\|_1[\lambda_{\min}(\mathbf{K}_{\mathcal{S}})]}\right)\sum_{k\geq m}\mathbb{P}\{m' = k\} \\
&\leq s\exp\left(-\frac{m\xi_{\overline{\delta}}(1-\widehat{\delta})\alpha^2}{s\|\boldsymbol{\nu}^{N,M}\|_1[\lambda_{\min}(\mathbf{K}_{\mathcal{S}})]^{-1}}\right).
\end{aligned}$$

□

The argument to prove the RISP in the uniform case remains unchanged.

Finally, avoiding repetitions, we are able to provide a meaningful functional interpretation of the RISP. Indeed, we can write the uniform RISP as

$$\inf_{\mathbf{u}\in\Sigma_s^N}\sup_{\mathbf{v}\in\mathbb{R}^m}\frac{\mathbf{v}^\intercal\widetilde{\mathbf{D}}\widetilde{\mathbf{A}}\mathbf{u}}{\|\mathbf{K}^{\frac{1}{2}}\mathbf{u}\|_2\|\mathbf{v}\|_2} = \inf_{\mathbf{u}\in\Sigma_s^N}\sup_{\mathbf{v}\in\mathbb{R}^m}\frac{\mathbf{v}^\intercal\widetilde{\mathbf{A}}\mathbf{u}}{\|\mathbf{K}^{\frac{1}{2}}\mathbf{u}\|_2\|\widetilde{\mathbf{D}}^{-1}\mathbf{v}\|_2} = \inf_{u\in U_s^N}\sup_{v\in V_{\widetilde{\mathcal{T}}}^M}\frac{a(u,v)}{\|u\|_U|||v|||},$$

where the norm

$$|||v|||^2 := \|\widetilde{\mathbf{D}}^{-1}\Phi^{-1}v\|_2^2 = \sum_{i=1}^m \frac{r_i}{m' p_{\tau_i}} v_{\tau_i}^2,$$

is a weighted ℓ^2-norm depending on the representation of v with respect to the basis $\{\varphi_{\widetilde{\tau}_i}\}_{i\in[m]}$. The operator Φ^{-1} is implicitly restricted from $V_{\widetilde{\mathcal{T}}}^M$ to $\mathbb{R}^m$. This provides a meaningful interpretation of the discrete RISP in a functional setting. We underline that in the case with repetitions, this remark does not hold, since the norm $|||\cdot|||$ would not be well defined.

In the next section of this chapter, we will exploit the recovery results based on the RISP, and apply them to the ADR equation.

3.3 Application to advection-diffusion-reaction equations

In this section, we apply the general theory presented in Section 3.2 to advection-diffusion-reaction (ADR) equations.

We adopt Corollary 3.18 as the main tool. In particular, we provide estimates for α, β, κ_s, $\widehat{K}$, $\widehat{\gamma}$, $\boldsymbol{\nu}^N$ and $\|\boldsymbol{\nu}^{N,M}\|_1$, and then deduce suitable hypotheses

on m and M such that the CORSING method recovers the best s-term approximation u^s to u. All the recovery results of the section are given in expectation, but they can be easily converted in probability (see Remark 3.2.4). Finally, as in Section 3.2.4, we assume $\widehat{u}$ to solve (3.6) exactly.

Let us first fix the notation. Consider $\Omega = (0,1)$, $U = V = H_0^1(\Omega)$ and

$$(u,v)_U = (u,v)_V = \int_\Omega u'(x)v'(x)\,\mathrm{d}x,$$

resulting in $\|\cdot\|_U = \|\cdot\|_V = |\cdot|_{H^1(\Omega)}$, the $H^1(\Omega)$-seminorm. Moreover, consider the bases $\mathcal{S}^R$ and $\mathcal{H}^L$ defined in Section 2.3 and the corresponding CORSING $\mathcal{HS}$ and $\mathcal{SH}$ strategies.

In both cases, $\mathcal{HS}$ and $\mathcal{SH}$, we observe that Hypothesis 1 is fulfilled and that $\mathbf{K} = \mathbf{I}$. Thus, in particular, from (3.23), $\kappa_s = 1$.

As reference problem, we consider the one-dimensional ADR equation over Ω, with Dirichlet boundary conditions

$$\begin{cases} -u'' + bu' + \rho u = f & \text{in } \Omega \\ u(0) = u(1) = 0, \end{cases} \tag{3.44}$$

with $b, \rho \in \mathbb{R}$ and $f : (0,1) \to \mathbb{R}$, corresponding to the weak problem

$$\text{find } u \in H_0^1(\Omega): \quad (u',v') + b(u',v) + \rho(u,v) = (f,v), \quad \forall v \in H_0^1(\Omega), \tag{3.45}$$

where $(\cdot,\cdot)$ denotes the standard inner product in $L^2(\Omega)$.

3.3.1 The 1D Poisson equation ($\mathcal{HS}$).

First, we deal with the Poisson equation, corresponding to (3.44) with $b = \rho = 0$, whose weak formulation is

$$\text{find } u \in H_0^1(\Omega): \quad a_\Delta(u,v) = (f,v), \quad \forall v \in H_0^1(\Omega). \tag{3.46}$$

where $a_\Delta(u,v) := (u',v')$. In such a case, we denote the local a-coherence by $\boldsymbol{\mu}_\Delta^N$. The inf-sup and continuity constants of $a_\Delta(\cdot,\cdot)$ are $\alpha = \beta = 1$.

We can prove the following result for the CORSING $\mathcal{HS}$ procedure applied to (3.46).

Proposition 3.22. *Fix a maximum hierarchical level $L \in \mathbb{N}$, corresponding to $N = 2^{L+1} - 1$. Then, for every $\varepsilon \in (0, 2^{-1/3}]$ and $s \le 2N/e$, provided that*

$$M \ge \widehat{C} sN, \quad m \ge \overline{C} s \log M [s \log(N/s) + \log(s/\varepsilon)],$$

for suitable constants $\overline{C}$ and $\widehat{C}$, and chosen the upper bound $\boldsymbol{\nu}^N$ as

$$\nu_q^N := \frac{8}{\pi q}, \quad \forall q \in \mathbb{N},$$

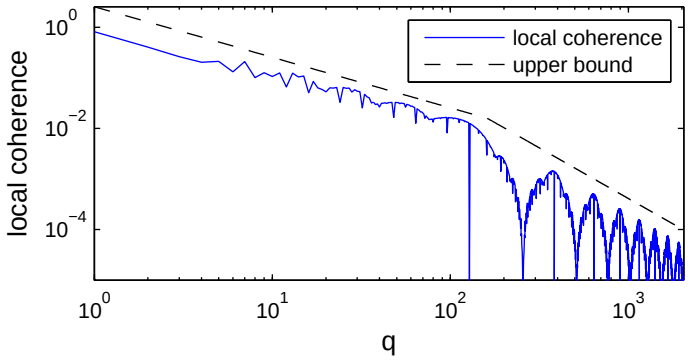


Figure 3.1: Sharpness of the upper bound (3.48) with $N = 127$ and $M = 2047$.

the CORSING $\mathcal{HS}$ solution to (3.46) *fulfills*

$$\mathbb{E}[|\mathcal{T}_{\mathcal{K}}\widehat{u} - u|_{H^1(\Omega)}] < 5|u^s - u|_{H^1(\Omega)} + 2\mathcal{K}\varepsilon,$$

for every $\mathcal{K} > 0$ such that $|u|_{H^1(\Omega)} \leq \mathcal{K}$, with $\mathcal{T}_{\mathcal{K}}$ defined as in (3.37)*. In particular, two possible upper bounds for $\widehat{C}$ and $\overline{C}$ are*

$$\widehat{C} \leq \frac{80}{3\pi^2} \approx 2.70 \quad \textit{and} \quad \overline{C} \leq \frac{840}{\pi}\left(1 + \frac{1}{\log 3}\right) \approx 511.$$

Proof. An explicit computation yields the exact stiffness matrix entries (the dependence of ℓ and k on j is omitted)

$$a_\Delta(\mathcal{H}_{\ell,k}, \mathcal{S}_q) = \frac{4\sqrt{2}}{\pi}\frac{2^{\frac{\ell}{2}}}{q}\sin\left(\frac{\pi q}{2^\ell}\left(k + \frac{1}{2}\right)\right)\sin^2\left(\frac{\pi}{4}\frac{q}{2^\ell}\right). \tag{3.47}$$

Using Definition 3.3, employing the inequalities $\sin^2(x) \leq 1$ on the first sine and $\sin^4(x) \leq \min\{1, |x|\}$ on the second sine, for every $x \in \mathbb{R}$, we have

$$|a_\Delta(\mathcal{H}_{\ell,k}, \mathcal{S}_q)|^2 \leq \frac{32}{\pi^2}\frac{2^\ell}{q^2}\sin^4\left(\frac{\pi}{4}\frac{q}{2^\ell}\right) \leq \min\left\{\frac{32}{\pi^2}\frac{2^\ell}{q^2}, \frac{8}{\pi q}\right\},$$

and, thus, we obtain the upper bound

$$\mu^N_{\Delta,q} \leq \min\left\{\frac{32}{\pi^2}\frac{2^L}{q^2}, \frac{8}{\pi q}\right\}. \tag{3.48}$$

Figure 3.1 shows that this bound is sharp. Considering the first argument of

the minimum in (3.48), on noticing that $2^L = (N+1)/2$, we obtain

$$\sum_{q>M} \mu^N_{\Delta,q} \le \frac{32}{\pi^2}\frac{N+1}{2}\sum_{q>M}\frac{1}{q^2} \le \frac{16}{\pi^2}(N+1)\left[\frac{1}{(M+1)^2} + \int_{M+1}^{\infty}\frac{1}{q^2}\,\mathrm{d}q\right]$$
$$= \frac{16}{\pi^2}\frac{N+1}{M+1}\left[\frac{1}{M+1}+1\right] \le \frac{20}{\pi^2}\frac{N+1}{M+1} \le \frac{80}{3\pi^2}\frac{N}{M}.$$

The fourth and fifth relations hinge on the assumption $L \ge 1$, that implies $N \ge 3$. Consequently, assuming $M \ge N$ we have also $M \ge 3$. This implies $1/(M+1) \le 1/4$ (fourth relation) and $(N+1)/(M+1) \le 4N/(3M)$ (fifth relation). Thus, in view of Corollary 3.18, we can pick

$$\widehat{K} = \frac{80}{3\pi^2} \quad \text{and} \quad \widehat{\gamma} = 1.$$

Now, to bound $\|\boldsymbol{\nu}^{N,M}\|_1$, which is required by Corollary 3.18, we deal with the second argument of the minimum in (3.48) and set

$$\nu^N_q := \frac{8}{\pi q}.$$

This choice leads to the estimate

$$\|\boldsymbol{\nu}^{N,M}\|_1 = \frac{8}{\pi}\sum_{q=1}^{M}\frac{1}{q} \le \frac{8}{\pi}\left[1+\int_1^M \frac{1}{q}\,\mathrm{d}q\right] = \frac{8}{\pi}(1+\log M) \le \frac{8}{\pi}\left(1+\frac{1}{\log 3}\right)\log M, \tag{3.49}$$

since $M \ge 3$. Thus, combining the lower bound for m and M in Corollary 3.18 with (3.49), we conclude the proof. □

Remark 3.3.1. The upper bound $\sin^4(x) \le \min\{1, |x|\}$ can be improved as $\sin^4(x) \le \min\{1, 0.68|x|\}$. This change leads to rescaling the value of $\overline{C}$ by a factor 0.68, i.e., $\overline{C} \approx 347$. □

Remark 3.3.2. The choice $\nu^N_q = 8/(\pi q)$ is suboptimal. If we choose the sharper upper bound

$$\nu^N_q = \min\left\{\frac{32}{\pi^2}\frac{2^L}{q^2}, \frac{8}{\pi q}\right\},$$

the term $\log M$ in the lower bound to m can be replaced by $\log N$. Indeed, in this case

$$\|\boldsymbol{\nu}^{N,M}\|_1 \lesssim \sum_{q=1}^{N}\frac{1}{q} + N\sum_{q=N+1}^{M}\frac{1}{q^2} \lesssim \log N + N\left(\frac{1}{N}-\frac{1}{M}\right) \lesssim \log N + 1 - \frac{1}{s} \lesssim \log N.$$

□

3.3.2 The 1D ADR equation ($\mathcal{HS}$)

We consider problem (3.44) and state the following result.

Proposition 3.23. *Fix a maximum hierarchical level $L \in \mathbb{N}$, corresponding to $N = 2^{L+1} - 1$. Then, for every $\varepsilon \in (0, 2^{-1/3}]$ and $s \leq 2N/e$, provided that*

$$M \gtrsim sN, \quad \frac{|b|}{M} \lesssim 1, \quad \frac{|\rho|}{M^2} \lesssim 1,$$

$$m \gtrsim s(\log M + |b|^2 + |\rho|^2)[s\log(N/s) + \log(s/\varepsilon)],$$

and chosen the upper bound $\boldsymbol{\nu}^N$ such that

$$\nu_q^N \sim \frac{1}{q} + \frac{|b|^2}{q^3} + \frac{|\rho|^2}{q^5}, \quad \forall q \in \mathbb{N},$$

the CORSING $\mathcal{HS}$ solution to (3.45)*, with $\rho > -2$, fulfills*

$$\mathbb{E}[|\mathcal{I}_{\mathcal{K}}\widehat{u} - u|_{H^1(\Omega)}] < \left(1 + \frac{4 + 2\sqrt{2}|b| + 2|\rho|}{1 + \min(0, \rho/2)}\right)|u^s - u|_{H^1(\Omega)} + 2\mathcal{K}\varepsilon,$$

for every $\mathcal{K} > 0$ such that $|u|_{H^1(\Omega)} \leq \mathcal{K}$, with $\mathcal{I}_{\mathcal{K}}$ defined as in (3.37)*.*

Proof. The argument is the same as in Proposition 3.22, thus we will just highlight the different parts. The precise values of the asymptotic constants will not be tracked during the proof.

First, a straightforward computation gives

$$a(\mathcal{H}_{\ell,k}, \mathcal{S}_q) = \frac{4\sqrt{2}}{\pi}\frac{2^{\frac{\ell}{2}}}{q}\sin^2\left(\frac{\pi}{4}\frac{q}{2^\ell}\right)\left[\left(1 + \frac{\rho}{(\pi q)^2}\right)\sin\left(\frac{\pi q}{2^\ell}(k + \tfrac{1}{2})\right) - \frac{b}{\pi q}\cos\left(\frac{\pi q}{2^\ell}(k + \tfrac{1}{2})\right)\right].$$

Hence, using the same upper bounds as in Proposition 3.22, we obtain

$$|a(\mathcal{H}_{\ell,k}, \mathcal{S}_q)|^2 \lesssim \min\left\{\frac{2^\ell}{q^2}, \frac{1}{q}\right\}\left(1 + \frac{|b|^2}{q^2} + \frac{|\rho|^2}{q^4}\right),$$

and, consequently,

$$\mu_q^N \lesssim \min\left\{\frac{N}{q^2}, \frac{1}{q}\right\}\left(1 + \frac{|b|^2}{q^2} + \frac{|\rho|^2}{q^4}\right). \tag{3.50}$$

Considering the first argument of the minimum in (3.50), yields

$$\sum_{q>M} \mu_q^N \lesssim N\left[\sum_{q>M}\frac{1}{q^2} + |b|^2\sum_{q>M}\frac{1}{q^4} + |\rho|^2\sum_{q>M}\frac{1}{q^6}\right]$$
$$\lesssim N\left[\frac{1}{M} + \frac{|b|^2}{M^3} + \frac{|\rho|^2}{M^5}\right] \lesssim \frac{N}{M}.$$

The second inequality hinges on estimates of the sums by suitable integrals, whereas the third one is implied by the hypotheses $|b|/M \lesssim 1$ and $|\rho|/M^2 \lesssim 1$.

Now, considering the second argument of the minimum in (3.50), we have the upper bound

$$\nu_q^N \sim \frac{1}{q} + \frac{|b|^2}{q^3} + \frac{|\rho|^2}{q^5}, \quad \forall q \in \mathbb{N},$$

and, consequently, the ℓ^1-norm of its truncation fulfills

$$\|\boldsymbol{\nu}^{N,M}\|_1 \sim \sum_{q=1}^{M}\frac{1}{q} + \sum_{q=1}^{M}\frac{|b|^2}{q^3} + \sum_{q=1}^{M}\frac{|\rho|^2}{q^5} \lesssim \log M + |b|^2 + |\rho|^2.$$

Finally, we notice that (2.4) and (2.3) hold with

$$\alpha = 1 + \min\left(0, \frac{\rho}{2}\right), \quad \beta = 1 + \frac{|b|}{\sqrt{2}} + \frac{|\rho|}{2},$$

thanks to the Poincaré inequality

$$\sqrt{2}\|v\|_{L^2(\Omega)} \le |v|_{H^1(\Omega)}, \quad \forall v \in H_0^1(\Omega).$$

The thesis is now a direct consequence of Corollary 3.18. □

3.3.3 The 1D Poisson equation ($\mathcal{SH}$)

We prove a recovery result for the CORSING $\mathcal{SH}$ method applied to the Poisson problem (3.46).

Proposition 3.24. *For every $\varepsilon \in (0, 2^{-1/3}]$ and $s \le 2N/e$, there exist two positive constants $\overline{C}$ and $\widehat{C}$ such that, provided*

$$M \ge \widehat{C}\sqrt{s}N, \quad m \ge \overline{C}s\log(M)[s\log(N/s) + \log(s/\varepsilon)],$$

with M of the form $M = 2^{L+1} - 1$ for some $L \in \mathbb{N}$, and chosen the upper bound $\boldsymbol{\nu}^N$ as

$$\nu_q^N = \frac{1}{2^{\ell(q)-1}}, \quad \forall q \in \mathbb{N},$$

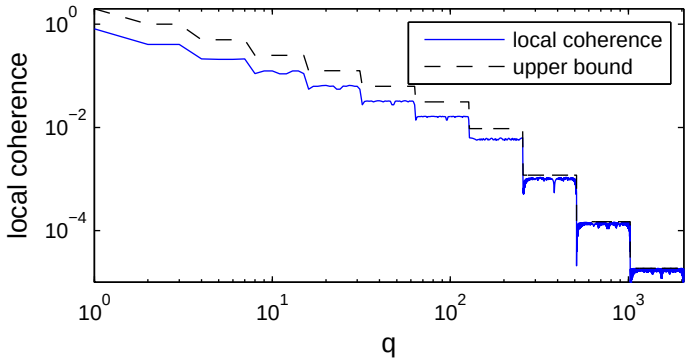


Figure 3.2: Sharpness of the upper bound (3.51) with $N = 127$ and $M = 2047$.

the CORSING $\mathcal{SH}$ solution to (3.46) *fulfills*

$$\mathbb{E}[|\mathcal{T}_{\mathcal{K}}\widehat{u} - u|_{H^1(\Omega)}] \leq 5|u^s - u|_{H^1(\Omega)} + 2\mathcal{K}\varepsilon,$$

for every $\mathcal{K} > 0$ such that $|u|_{H^1(\Omega)} \leq \mathcal{K}$, with $\mathcal{T}_{\mathcal{K}}$ defined as in (3.37) *and where α and β are defined by* (2.4) *and* (2.3)*, respectively. In particular, two possible upper bounds for $\widehat{C}$ and $\overline{C}$ are*

$$\widehat{C} \leq \frac{\pi}{\sqrt{3}} \approx 1.81 \quad \textit{and} \quad \overline{C} \leq \frac{210\log_2(e)\log(4)}{\log(3)} \approx 382.$$

Proof. The proof is analogous to that of Proposition 3.22. We highlight only the main differences. First, notice that

$$a_\Delta(\mathcal{S}_j, \mathcal{H}_{\ell(q),k(q)}) = a_\Delta(\mathcal{H}_{\ell(q),k(q)}, \mathcal{S}_j).$$

Moving from (3.47) and employing the inequality $\sin^4(x) \leq \min\{|x|^4, |x|^2\}$, for every $x \in \mathbb{R}$, we obtain

$$\mu^N_{\Delta,q} \leq \min\left\{\frac{\pi^2}{8}\frac{N^2}{2^{3\ell(q)}}, \frac{1}{2^{\ell(q)-1}}\right\}. \tag{3.51}$$

Figure 3.2 shows the sharpness of this bound.

Considering the first argument of the minimum in (3.51), and since $M = 2^{L+1} - 1$, we have that

$$\sum_{q>M} \mu^N_{\Delta,q} \leq \frac{\pi^2}{8}N^2 \sum_{\ell>L}\sum_{k=0}^{2^\ell-1}\frac{1}{2^{3\ell}} = \frac{\pi^2}{8}N^2\sum_{\ell>L}\frac{1}{2^{2\ell}} = \frac{\pi^2}{8}\frac{N^2}{2^{2(L+1)}}\sum_{\ell\geq 0}\frac{1}{2^{2\ell}} \leq \frac{\pi^2}{6}\left(\frac{N}{M}\right)^2,$$

where the change of variable $q \mapsto (\ell, k)$ has been used. Thus, if follows that

$$\widehat{K} = \frac{\pi^2}{6} \quad \text{and} \quad \widehat{\gamma} = \frac{1}{2}.$$

Now, by considering the second argument of the minimum in (3.51), we select

$$\nu_q^N := \frac{1}{2^{\ell-1}}$$

and conclude the proof by computing

$$\begin{aligned}\|\boldsymbol{\nu}^{N,M}\|_1 &= \sum_{\ell=0}^{L}\sum_{k=0}^{2^\ell-1}\frac{1}{2^{\ell-1}} = 2(L+1) = 2\log_2(e)\log(M+1)\\ &\le 2\log_2(e)\frac{\log(M+1)}{\log(M)}\log(M) \le \frac{2\log_2(e)\log(4)}{\log(3)}\log(M),\end{aligned}$$

since $M \ge 3$, thanks to $L \ge 1$. □

Remark 3.3.3. The choice of $\mathbf{p}$ prompted by Proposition 3.24 (i.e., $p_q \sim 2^{-\ell(q)}$) coincides with that in [BMP15], in the R-CORSING $\mathcal{SH}$ case, for the corresponding parameter $\mathbf{w}$, tuned via a trial-and-error procedure. □

3.3.4 The 1D ADR equation ($\mathcal{SH}$)

Considerations analogous to those made in the $\mathcal{HS}$ case hold in the advective/reactive case. It suffices to notice that

$$(u',v') + b(u',v) + \rho(u,v) = (v',u') - b(v',u) + \rho(v,u), \quad \forall u,v \in H_0^1(\Omega),$$

and then apply the same arguments as in the $\mathcal{HS}$ case.

3.3.5 The 1D diffusion equation ($\mathcal{HS}$)

Consider now the diffusion equation with a nonconstant diffusion coefficient and homogeneous Dirichlet boundary conditions over the domain $\Omega = (0,1)$

$$\begin{cases} -(\eta u')' = f, & \text{in } \Omega \\ u(0) = u(1) = 0, \end{cases} \tag{3.52}$$

where $\eta : \overline{\Omega} \to \mathbb{R}$ is the diffusion term, such that $\eta \in L^\infty(\Omega)$. Moreover, assume that there exists $\eta_{\min} > 0$ such that $\eta(x) \ge \eta_{\min}$ for almost every $x \in \Omega$. The resulting weak formulation is

$$\text{find } u \in H_0^1(\Omega): \quad (\eta u', v') = (f, v), \quad \forall v \in H_0^1(\Omega), \tag{3.53}$$

with associated bilinear form

$$a(u,v) = (\eta u', v').$$

In order to apply the arguments used for the Poisson equation, we expand η with respect to the cosine basis $\{\mathcal{C}_r\}_{r\in\mathbb{N}_0}$,

$$\eta = \sum_{r=0}^{\infty} \eta_r \mathcal{C}_r. \tag{3.54}$$

where $\mathcal{C}_r(x) := \sqrt{2}\cos(\pi r x)$, for every $r \in \mathbb{N}$, and $\mathcal{C}_0 \equiv 1$. This is allowed since $\eta \in L^\infty(\Omega) \subseteq L^2(\Omega)$ and $\{\mathcal{C}_r\}_{r\in\mathbb{N}_0}$ is a complete orthonormal system of $L^2(\Omega)$, equipped with the standard inner product.

Then, the elements of the stiffness matrix associated with (3.52) can be explicitly computed in terms of those of the (infinite) stiffness matrix associated with the Poisson problem (3.46)

$$\begin{aligned} a(\mathcal{H}_{\ell,k}, \mathcal{S}_q) &= (\sum_{r=0}^{\infty} \eta_r \mathcal{C}_r \mathcal{H}'_{\ell,k}, \mathcal{S}'_q) = \sum_{r=0}^{\infty} \eta_r (\mathcal{H}'_{\ell,k}, \mathcal{C}_r \mathcal{C}_q) \\ &= \sum_{r=0}^{\infty} \eta_r \zeta_r (\mathcal{H}'_{\ell,k}, \mathcal{C}_{q+r} + \mathcal{C}_{|r-q|}) \\ &= \sum_{r=0}^{\infty} \eta_r \zeta_r [a_\Delta(\mathcal{H}_{\ell,k}, \mathcal{S}_{q+r}) + a_\Delta(\mathcal{H}_{\ell,k}, \mathcal{S}_{|r-q|})], \end{aligned} \tag{3.55}$$

where we employed the trigonometric relations

$$\mathcal{S}'_q = \mathcal{C}_q, \quad \mathcal{C}_r \mathcal{C}_q = \zeta_r (\mathcal{C}_{r+q} + \mathcal{C}_{|r-q|}), \quad \forall q \in \mathbb{N},\ \forall r \in \mathbb{N}_0,$$

and

$$\zeta_r := \begin{cases} 1/2 & \text{if } r = 0, \\ 1/\sqrt{2} & \text{if } r \neq 0. \end{cases}$$

Formula (3.55) links the bilinear form $a(\cdot,\cdot)$, associated with the pure diffusive case, with the bilinear form $a_\Delta(\cdot,\cdot)$ of the Poisson equation.

Before applying the results of Section 3.2 to problem (3.53), we need a technical lemma about piecewise $\mathcal{C}^1$ odd periodic functions and their sine series expansion. Although elementary, we provide the proof for the sake of completeness. In the following, we refer to the one-dimensional torus as $\mathbb{T} := \mathbb{R}/2\mathbb{Z}$.

Lemma 3.25. *Consider a 2-periodic odd function $g : \mathbb{T} \to \mathbb{R}$ and suppose that there exists a finite set of points $\mathcal{P} \subseteq \mathbb{T}$ such that*

- $g \in \mathcal{C}^1(\mathbb{T} \setminus \mathcal{P})$;
- $\sup_{x\in\mathbb{T}\setminus\mathcal{P}} |g^{(k)}(x)| < \infty$ *for* $k = 0, 1$.

Then, the following asymptotic estimate holds

$$|(g, \sin(\pi r \cdot))_{L^2(\mathbb{T})}| \lesssim 1/r, \quad \forall r \in \mathbb{N}.$$

Proof. First, suppose $\mathcal{P} = \{\overline{x}\}$, with $\overline{x} = 1$. Define

$$f_0(x) := \sin(\pi x/2), \quad f_1(x) := -2/\pi \cos(\pi x/2), \quad \forall x \in [-1, 1)$$

and extend them periodically over $\mathbb{T}$. Notice that $f_0(1^-) = -f_0(1^+) = 1$ and $f_0'(1^-) = f_0'(1^+) = 0$. Moreover, $f_1(1^-) = f_1(1^+) = 0$ and $f_1'(1^-) = -f_1'(1^+) = -1$.

Then, define

$$\widetilde{g} := g + \frac{1}{2} \sum_{k=0,1} (g^{(k)}(1^+) - g^{(k)}(1^-)) f_k.$$

It is not difficult to verify that $\widetilde{g} \in \mathcal{C}^1(\mathbb{T})$; this, in turn, implies

$$|(\widetilde{g}, \sin(\pi r \cdot))_{L^2(\mathbb{T})}| \lesssim 1/r, \quad \forall r \in \mathbb{N}.$$

Moreover, by direct computation, we have

$$|(f_k, \sin(\pi r \cdot))_{L^2(\mathbb{T})}| \lesssim 1/r, \quad \forall r \in \mathbb{N}, \quad \forall k = 0, 1.$$

These two facts imply the thesis.

If $\overline{x} \neq 1$, the same argument can be applied to $g(\cdot - 1 + \overline{x})$. When $|\mathcal{P}| > 1$, it is necessary to remove every point of discontinuity using different translates of f_0 and f_1. □

Remark 3.3.4. In the previous lemma, g is not necessarily continuous on $\mathbb{T}$. □

The following proposition assesses the performances of the CORSING $\mathcal{HS}$ procedure applied to problem (3.53). For the sake of simplicity, we will not keep track of the constants $\widehat{C}$ and $\overline{C}$ during the proof.

Proposition 3.26. *Let $\Omega = (0, 1)$ and $\eta \in L^\infty(\Omega)$ be such that*

- *there exists $\eta_{\min} > 0$ so that $\eta(x) \geq \eta_{\min}$, for almost every $x \in \Omega$;*
- *there exists a finite set $\mathcal{P} \subseteq \overline{\Omega}$ such that $\eta \in \mathcal{C}^2(\Omega \setminus \mathcal{P})$;*
- $\sup_{x \in \Omega \setminus \mathcal{P}} |\eta^{(k)}(x)| < \infty$, *for $k = 1, 2$.*

Fix a maximum hierarchical level $L \in \mathbb{N}$ and put $N = 2^{L+1} - 1$. Then, there exists $C > 0$ such that

$$\nu_q^N = \frac{C}{q}, \quad \forall q \in \mathbb{N},$$

and there exist two positive constants $\widehat{C}$ *and* $\overline{C}$ *such that, provided*

$$M \geq \widehat{C} s N, \quad m \geq \overline{C} s \log M [s \log(N/s) + \log(s/\varepsilon)],$$

the CORSING $\mathcal{HS}$ *solution* $\widehat{u}$ *to* (3.53) *fulfills*

$$\mathbb{E}[|\mathcal{I}_{\mathcal{K}} \widehat{u} - u|_{H^1(\Omega)}] \leq \left(1 + \frac{4\|\eta\|_{L^\infty}}{\eta_{\min}}\right) |u^s - u|_{H^1(\Omega)} + 2\mathcal{K}\varepsilon,$$

for every $\mathcal{K} > 0$ *such that* $|u|_{H^1(\Omega)} \leq \mathcal{K}$, *with* $\mathcal{I}_{\mathcal{K}}$ *defined as in* (3.37) *and where* α *and* β *are defined by* (2.4) *and* (2.3), *respectively.*

Proof. First, notice that the inf-sup constant associated with $a(\cdot,\cdot)$ fulfills $\alpha \geq \eta_{\min}$, indeed

$$a(u,v) \geq \eta_{\min} a_\Delta(u,v), \quad \forall u, v \in H_0^1(\Omega),$$

whereas the continuity constant satisfies $\beta \leq \|\eta\|_{L^\infty}$. Moreover, recall that $\kappa_s = 1$.

Now, consider the diffusion term η. Extend it evenly from $[0,1]$ to $[-1,1]$ and then, in turn, extend the resulting function periodically from $[-1,1]$ to the torus $\mathbb{T} = \mathbb{R}/2\mathbb{R}$. Denote the resulting function $\widetilde{\eta} : \mathbb{T} \to \mathbb{R}$. Thanks to the regularity hypothesis made on η, the function $\widetilde{\eta}'$ fulfills the hypotheses of Lemma 3.25. Henceforth, we have the following asymptotic estimate for the coefficients η_r, defined by the expansion (3.54)

$$|\eta_r| \sim |(\widetilde{\eta}, \cos(\pi r \cdot))_{L^2(\mathbb{T})}| \sim \frac{|(\widetilde{\eta}', \sin(\pi r \cdot))_{L^2(\mathbb{T})}|}{(r+1)} \lesssim \frac{1}{(r+1)^2}, \quad \forall r \in \mathbb{N}_0. \tag{3.56}$$

Now, exploiting relation (3.55), we obtain an upper bound to $\boldsymbol{\mu}^N$, depending on $\boldsymbol{\mu}_\Delta^N$, i.e., the local a-coherence associated with (3.46), $\mathcal{HS}$ case (the dependence of (ℓ, k) on j is omitted)

$$\begin{aligned}
\mu_q^N &= \sup_{j \in [N]} |a(\mathcal{H}_{\ell,k}, \mathcal{S}_q)|^2 \\
&= \sup_{j \in [N]} \left[\sum_{r=0}^{\infty} \eta_r \zeta_r (a_\Delta(\mathcal{H}_{\ell,k}, \mathcal{S}_{q+r}) + a_\Delta(\mathcal{H}_{\ell,k}, \mathcal{S}_{|r-q|})) \right]^2 \\
&= \sup_{j \in [N]} \left[\sum_{r=0}^{\infty} \eta_r \zeta_r \frac{(r+1)^\vartheta}{(r+1)^\vartheta} (a_\Delta(\mathcal{H}_{\ell,k}, \mathcal{S}_{q+r}) + a_\Delta(\mathcal{H}_{\ell,k}, \mathcal{S}_{|r-q|})) \right]^2 \\
&\leq \frac{1}{2} \left[\sum_{r=1}^{\infty} \frac{1}{r^{2\vartheta}} \right] \sup_{j \in [N]} \sum_{r=0}^{\infty} \eta_r^2 (r+1)^{2\vartheta} [a_\Delta(\mathcal{H}_{\ell,k}, \mathcal{S}_{q+r}) + a_\Delta(\mathcal{H}_{\ell,k}, \mathcal{S}_{|r-q|})]^2 \\
&\leq \left[\sum_{r=1}^{\infty} \frac{1}{r^{2\vartheta}} \right] \sum_{r=0}^{\infty} \eta_r^2 (r+1)^{2\vartheta} (\mu_{\Delta,q+r}^N + \mu_{\Delta,|r-q|}^N).
\end{aligned}$$

The fourth relation hinges on Cauchy-Schwarz inequality, combined with relation $\zeta_r^2 \leq 1/2$, for every $r \in \mathbb{N}_0$. In the last inequality, the estimate $(x+y)^2 \leq 2(x^2+y^2)$, for every $x, y \in \mathbb{R}$, has been employed. We notice that the series corresponding to the first factor in the last right-hand side is convergent if and only if $\vartheta > 1/2$. In particular, exploiting (3.56) and choosing $\vartheta = 1$, we have

$$\sum_{q>M} \mu_q^N \lesssim \underbrace{\sum_{q>M} \sum_{r=0}^{\infty} \frac{\mu_{\Delta,q+r}^N}{(r+1)^2}}_{=:S_1} + \underbrace{\sum_{q>M} \sum_{r=0}^{\infty} \frac{\mu_{\Delta,|q-r|}^N}{(r+1)^2}}_{=:S_2}. \tag{3.57}$$

Recalling that (3.48) implies

$$\mu_{\Delta,q}^N \lesssim \frac{N}{q^2}, \quad \forall q \in \mathbb{N}, \tag{3.58}$$

and plugging (3.58) into (3.57), we obtain

$$S_1 \lesssim N \sum_{r=0}^{\infty} \frac{1}{(r+1)^2} \sum_{q>M} \frac{1}{(r+q)^2} \lesssim N \sum_{r=0}^{\infty} \frac{1}{(r+1)^2(M+r)} \leq \frac{N}{M} \sum_{r=0}^{\infty} \frac{1}{(r+1)^2} \lesssim \frac{N}{M}.$$

Moreover, exploiting again (3.57), (3.58) and the fact that $\mu_{\Delta,0}^N = 0$, we estimate

$$S_2 \lesssim N \Bigg[\underbrace{\sum_{q>M} \sum_{r=0}^{q-1} \frac{(r+1)^{-2}}{(q-r)^2}}_{=:S_{21}} + \underbrace{\sum_{q>M} \sum_{r=q+1}^{\infty} \frac{(r+1)^{-2}}{(q-r)^2}}_{=:S_{22}} \Bigg].$$

We deal with the term S_{21} as follows:

$$\begin{aligned} S_{21} &= \sum_{r=0}^{\infty} (r+1)^{-2} \sum_{q \geq \max(r,M)+1} \frac{1}{(q-r)^2} \lesssim \sum_{r=0}^{\infty} \frac{(r+1)^{-2}}{\max(r,M)+1-r} \\ &= \sum_{r=0}^{M} \frac{(r+1)^{-2}}{M+1-r} + \sum_{r>M} (r+1)^{-2}. \end{aligned}$$

Now, consider the function $f(r) := (r+1)^{-2}(M+1-r)^{-1}$. We note that $f : [0, M] \to \mathbb{R}$ has a global minimum in $r = (2M+1)/3 \in (0, M]$. Henceforth, from elementary geometric considerations, we have

$$\sum_{r=0}^{M} f(r) \leq f(0) + f(M) + \int_0^M f(r)\,\mathrm{d}r, \quad \forall M \in \mathbb{N}.$$

Via direct computation, it turns out that

$$\int_0^M f(r)\,\mathrm{d}r = \frac{2}{M+2} - \frac{1}{M+1} + \frac{2\log(M+1)}{(M+2)^2},$$

henceforth, $S_{21}(M) \lesssim 1/M$. Analogous arguments show that $S_{22}(M) \lesssim 1/M$.

To summarize, we have the truncation condition

$$\sum_{q>M} \mu_q^N \lesssim \frac{N}{M},$$

that yields the existence of $\widehat{C} > 0$ and $\widehat{\gamma} = 1$.

Exploiting the local a-coherence upper bound

$$\mu_{\Delta,q}^N \lesssim 1/q, \quad \forall q \in \mathbb{N},$$

derived from (3.48), and plugging it into (3.57), implies

$$\mu_q^N \lesssim \underbrace{\sum_{r=0}^{\infty} \frac{(r+1)^{-2}}{q+r}}_{T_1} + \underbrace{\sum_{r=0}^{q-1} \frac{(r+1)^{-2}}{q-r}}_{T_2} + \underbrace{\sum_{r>q} \frac{(r+1)^{-2}}{r-q}}_{T_3}.$$

First, we have

$$T_1 \leq \frac{1}{q} \sum_{r=0}^{\infty} \frac{1}{(r+1)^2} \lesssim \frac{1}{q}.$$

Moreover, applying the same argument used to bound S_{21}, we obtain

$$T_2 \leq \frac{1}{q} + \frac{1}{q} + \int_0^{q-1} \frac{(r+1)^{-2}}{q-r}\,\mathrm{d}r = \frac{2}{q} + \frac{2q\log q + q^2 - 1}{q(1+q)^2} \lesssim \frac{1}{q}.$$

Finally,

$$T_3 \leq \frac{1}{q} \sum_{r>q} \frac{(r+1)^{-1}}{(r-q)} \leq \frac{1}{q} \sum_{r>q} \frac{1}{(r-q)^2} \lesssim \frac{1}{q}.$$

Henceforth, we have $\nu_q^N \lesssim 1/q$ and, in particular, $\|\boldsymbol{\nu}^{N,M}\|_1 \lesssim \log M$.

The application of Corollary 3.18 concludes the proof. □

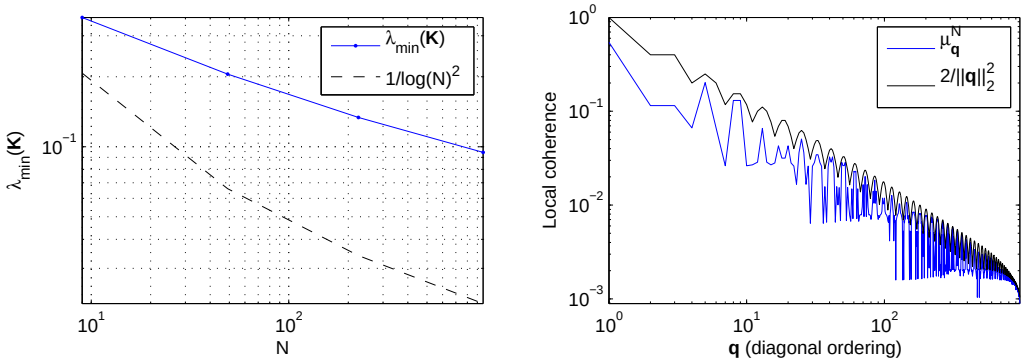


Figure 3.3: CORSING $\mathcal{PS}$ for 2D the Poisson equation. Numerical validation of inequality (3.59), (a); numerical validation of the upper bound $\boldsymbol{\mu}^N \leq \boldsymbol{\nu}^{N,M}$ in (3.60), (b).

3.3.6 The 2D Poisson equation ($\mathcal{PS}$)

Considering the extension to the two-dimensional case, we recall the results presented in Section 2.4, where CORSING is applied to the 2D ADR equation with constant coefficients, employing the hierarchical pyramids $\mathcal{P}^L$ and the tensor product of sine functions $\mathcal{S}^R$, yielding the approaches CORSING $\mathcal{PS}$ and $\mathcal{SP}$.

Due to the lack of orthogonality of the hierarchical pyramids in $\mathcal{P}^L$, they can only be used as trial functions. Indeed, in view of the theoretical setting of this work, Hypothesis 1 does not hold in the $\mathcal{SP}$ case. Hence, we focus on the CORSING $\mathcal{PS}$ approach.[3]

First, recalling Remarks 3.2.1 and 3.2.5, we estimate $\kappa_s \geq \lambda_{\min}(\mathbf{K})$ and employ the inequality

$$\lambda_{\min}(\mathbf{K}) \geq \log(N)^{-2}, \tag{3.59}$$

shown in [Yse86] (see also the numerical validation in Figure 3.3, (a)), in order to show that the term κ_s^{-1} grows at most logarithmically in N.

A less trivial task is to provide a sharp upper bound $\boldsymbol{\nu}^{N,M}$ to $\boldsymbol{\mu}^N$, due to the involved explicit expressions of the stiffness matrix entries. However, we can numerically check that the following upper bound holds quite sharply

$$\mu_{\mathbf{q}}^N \lesssim \frac{C}{\|\mathbf{q}\|_2^2} =: \nu_{\mathbf{q}}^N, \tag{3.60}$$

for a suitable value of $C > 0$. In Figure 3.3, (b) we numerically validate (3.60) with $N = M = 961$ and $C = 2$. This provides a practical recipe to implement

[3] As already noticed in Section 3.1.2, Hypothesis 1 can be weakened by assuming the test functions to form a Riesz basis.

CORSING $\mathcal{PS}$. In particular, when truncating the test basis at level R, i.e., considering $\mathcal{S}^R$, we have $M = R^2$ and

$$\begin{aligned}\|\boldsymbol{\nu}^{N,M}\|_1 &= \sum_{q_1=1}^{R}\sum_{q_2=1}^{R} \nu_{\mathbf{q}}^N \sim \sum_{q_1=1}^{R}\sum_{q_2=1}^{R} \frac{1}{\|\mathbf{q}\|_2^2} \\ &\leq \sum_{q_1=1}^{R} \frac{1}{1+q_2^2} + \sum_{q_2=1}^{R} \frac{1}{q_1^2+1} + \int_1^R \int_1^R \frac{1}{q_1^2+q_2^2}\,\mathrm{d}q_1\,\mathrm{d}q_2 \\ &\lesssim 1 + \int_0^{\pi/2}\int_1^R \frac{1}{r^2} r\,\mathrm{d}r\,\mathrm{d}\vartheta = 1 + \frac{\pi}{2}\log(R) \sim \log(M).\end{aligned}$$

In the first inequality, we estimate the double series with the integral plus the boundary terms, thanks to the fact that $\mathbf{q} \to 1/\|\mathbf{q}\|_2^2$ is decreasing with respect to q_1 and q_2; moreover, we employ the change of variable to polar coordinates (r, ϑ) and enlarge the integration domain in the second inequality. In particular, this implies the validity of Hypothesis 3 when the upper bound $\boldsymbol{\nu}^{N,M}$ is chosen as in (3.60).

The considerations made here allow for a practical use of CORSING $\mathcal{PS}$ to the Poisson problem, with the drawing probability

$$p_{\mathbf{q}} \sim \frac{1}{\|\mathbf{q}\|_2^2}, \quad \forall \mathbf{q} \in [R]^2$$

on the test space. Analogous numerical checks can be made in the general ADR case. Nevertheless, a formal application of the theory to the 2D CORSING $\mathcal{PS}$ case needs the analytical derivation of $\boldsymbol{\nu}^{N,M}$ and the verification of the truncation condition (3.22). These are still open issues.

A different possibility to deal with the multi-dimensional case, when the domain is of the form $\Omega = [0,1]^d$, with $d > 1$, is to generalize the $\mathcal{HS}$ formulation by tensorization of both the trial and the test functions. This option is discussed in Section 4.2

3.4 Further numerical experiments

We validate the above theoretical results by both a qualitative and a quantitative analysis.

3.4.1 Sensitivity analysis of the RISP constant

We investigate the sensitivity of $\widetilde{\alpha}$ to the constant $\overline{C}$ on the Poisson problem (3.46), in the setting $\mathcal{HS}$. We fix the hierarchical level to $L = 14$, corresponding

to $N = 32767$. We consider the values $s = 1, 2, 3, 4, 5$ and choose $M = sN$, while selecting m according to one of the following rules

$$\begin{aligned} &\text{Rule 1:} \quad m = \lceil \overline{C} s^2 \log M \log(N/s) \rceil, \\ &\text{Rule 2:} \quad m = \lceil \overline{C} s \log M \log(N/s) \rceil, \\ &\text{Rule 3:} \quad m = \lceil \overline{C} s \log(N/s) \rceil. \end{aligned} \tag{3.61}$$

Rule 1 is the one derived in this chapter, corresponding to $\overline{\gamma} = 2$. Rule 2 is associated with $\overline{\gamma} = 1$, and Rule 3 is the asymptotically optimal lower bound that a general sparse recovery procedure requires to be stable (see [FR13, Proposition 10.7]). For each choice of M and m, we repeat the following experiment 50 times: first, extract $\tau_1, \dots, \tau_m \in [M]$ i.i.d. with probability $p_q \sim 1/q$ and build the corresponding matrices $\mathbf{D}$ and $\mathbf{A}$; then, generate 1000 random subsets $\mathcal{S}_1, \dots, \mathcal{S}_{1000} \subseteq [N]$ of cardinality s and compute the non-uniform RISP constant $\widetilde{\alpha}_{\mathcal{S}_k}$ for every $k \in [1000]$, corresponding to the minimum singular value of $\mathbf{DA}$, using the `svd` command; finally, approximate the uniform RISP constant as

$$\widetilde{\alpha} \approx \min_{k \in [1000]} \widetilde{\alpha}_{\mathcal{S}_k}.$$

We consider the three trends in (3.61) and $\overline{C} = 2$ or 5. The corresponding six boxplots relative to the 50 different values of $\widetilde{\alpha}$, computed for each s, are shown in Figure 3.4, where the crosses represent the outliers.

For Rule 1 and 2, $\widetilde{\alpha}$ shows a similar behavior since both trends are approaching the value of the inf-sup constant, $\alpha = 1$, when s grows. We notice that the values computed for Rule 1 are more concentrated around the mean, implying that $\overline{\gamma} = 2$ is too conservative. For Rule 3, $\widetilde{\alpha}$ exhibits the lowest values, though the corresponding boxplots are quite aligned and have similar size, especially for $\overline{C} = 5$, where $\widetilde{\alpha}$ seems to stabilize around the value $\alpha/2$. For $\overline{C} = 2$, $\widetilde{\alpha}$ approaches the value $\alpha/4$, even though the presence of too many outliers suggests that the RISP is not being satisfied for a reasonable value of ε. However, since Rule 3 is quite satisfactory, especially for $\overline{C} = 5$, the quantity $\log M$ does not seem to be really necessary in Rule 2. Moreover, Rule 1 is penalized by both the $\log M$ term and the extra s factor.

3.4.2 CORSING validation

We test CORSING $\mathcal{HS}$ on the one-dimensional Poisson equation (3.46), choosing the forcing term so that the exact solution be

$$u(x) := \widetilde{u}_{0.2,0.7,1000}(x) + 0.3 \cdot \widetilde{u}_{0.4,0.4005,2000}(x), \quad \forall x \in [0,1] \tag{3.62}$$

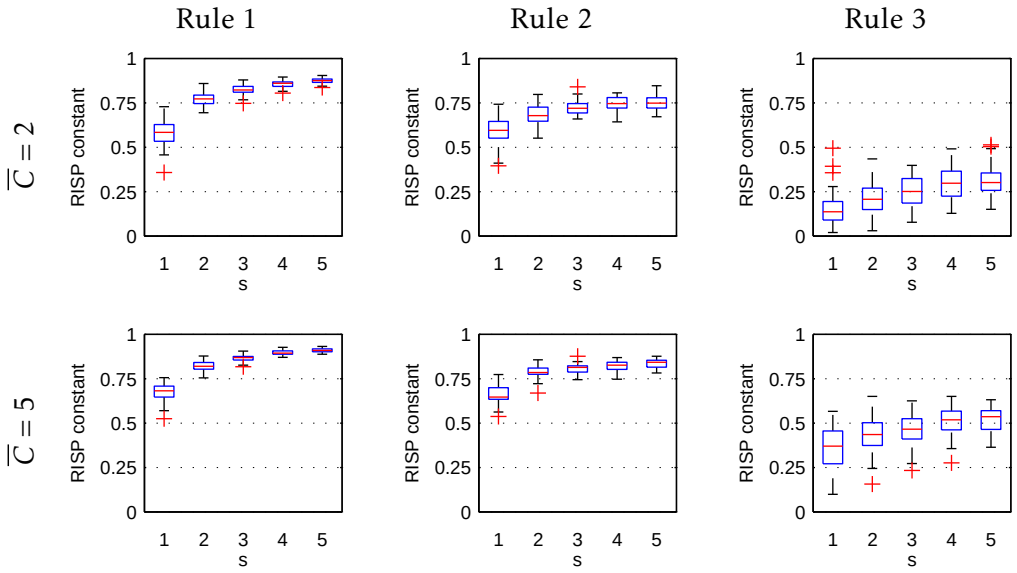


Figure 3.4: Sensitivity analysis of the RISP constant, with $M = sN$ and m defined according to (3.61).

with

$$\begin{aligned}
\widetilde{u}_{x_1,x_2,t}(x) &:= \overline{u}_{x_1,x_2,t}(x) - e_{x_1,x_2,t}(x),\\
e_{x_1,x_2,t}(x) &:= x\,\overline{u}_{x_1,x_2,t}(1) + (1-x)\,\overline{u}_{x_1,x_2,t}(0),\\
\overline{u}_{x_1,x_2,t}(x) &:= \arctan(t(x-x_1)) - \arctan(t(x-x_2)),
\end{aligned}$$

for every $x \in [0,1]$, $0 \le x_1 < x_2 \le 1$ and $t \in \mathbb{R}$. This particular solution is designed so as to exhibit two boundary layers at $x = 0.2$ and $x = 0.7$, and a small spike-shaped detail at $x = 0.4$ (see Figure 3.5).

The hierarchical multiscale basis is particularly suited to capture these sharp features. We fix $L = 12$, corresponding to $N = 8191$, $s = 50$, $M = sN$ and $m = 1200$.

In Figure 3.5, we compare u (dashed line) and $\widehat{u}$ (solid line). The exact solution is well recovered. Both boundary layers are correctly captured and also the spike-shaped feature is successfully detected. More quantitatively, the best 50-term relative error is $|u - u_{50}|_{H^1}/|u|_{H^1} \approx 0.092$ and the relative error of the CORSING solution is $|u - \widehat{u}|_{H^1}/|u|_{H^1} \approx 0.111$. Thus, via CORSING, we loose only the 21% of the best possible accuracy.

Figures 3.6 and 3.7 highlight that CORSING is able to find the most important coefficients of $\mathbf{u}$. In particular, in Figure 3.6, the coefficients of $\mathbf{u}$ and $\widehat{\mathbf{u}}$ are plotted according to the lexicographic ordering, whereas in Figure 3.7 they are shown in two dimensions: level ℓ is the vertical axis, and each level is divided horizontally into 2^ℓ parts, corresponding to $k = 0, \dots, 2^\ell - 1$, (left to right).

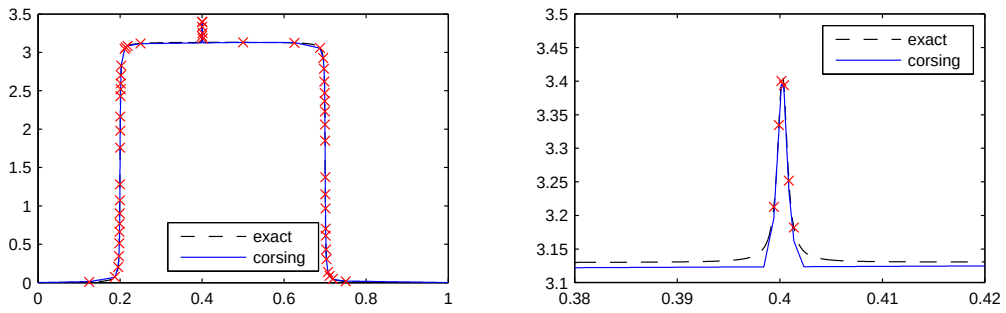


Figure 3.5: Left: comparison between u defined in (3.62) (dashed line) and $\widehat{u}$ (solid line). Right: a zoom in on the spike-shaped detail of u. Crosses correspond to the selected trial functions.

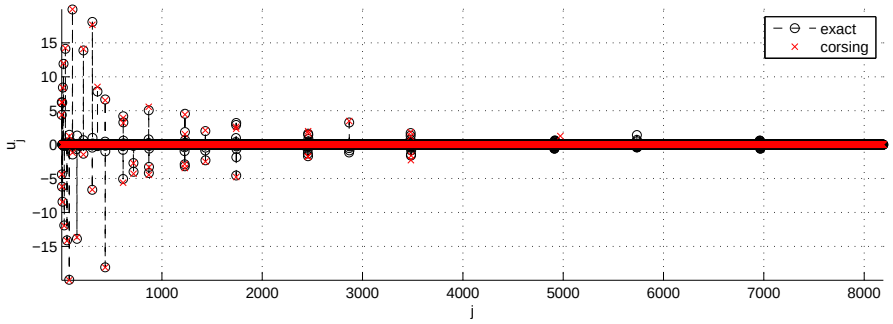


Figure 3.6: Comparison between $\mathbf{u}$ (circles) and $\widehat{\mathbf{u}}$ (crosses).

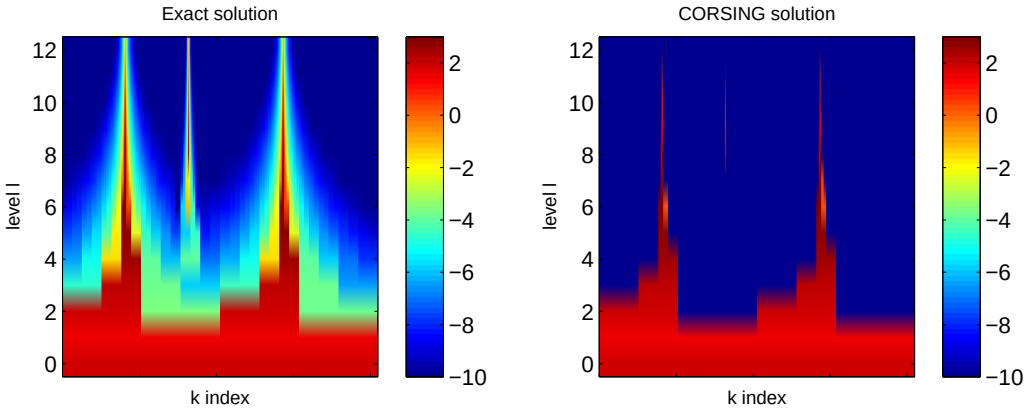


Figure 3.7: 2D color plot of $|u_{\ell,k}|$ and $|\widehat{u}_{\ell,k}|$ in logarithmic scale.

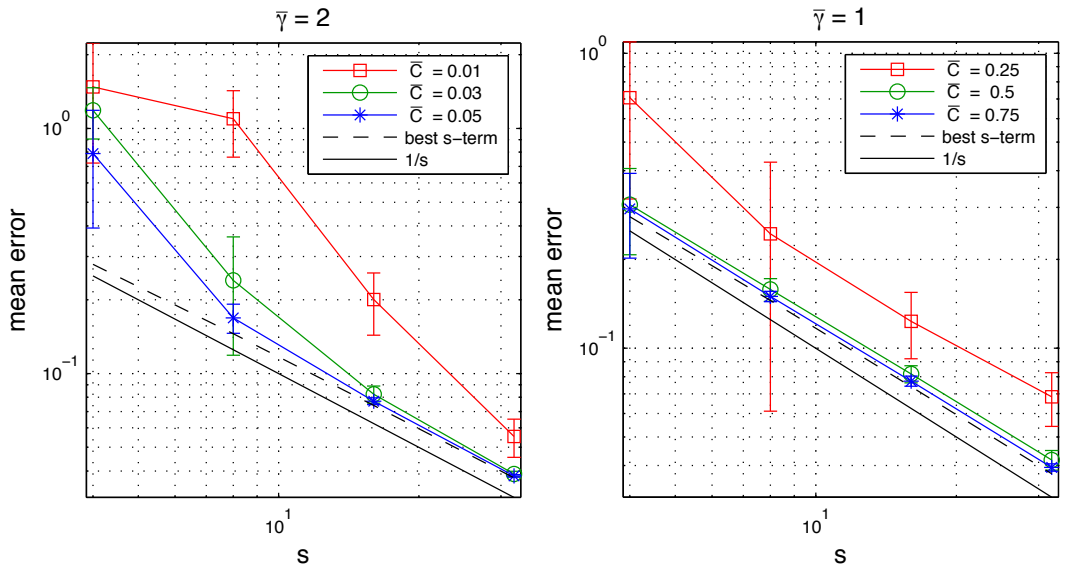


Figure 3.8: Convergence analysis: mean error ± standard deviation and best s-term approximation error. Case $\overline{\gamma} = 2$ (right) and $\overline{\gamma} = 1$ (left).

The color plots refer to $|u_{\ell,k}|$ (left) and $|\widehat{u}_{\ell,k}|$ (right), in logarithmic scale. It is remarkable the capability of CORSING in detecting the localized features of the solution (see the isolated vertical line in Figure 3.7 (right)).

3.4.3 Convergence analysis

We now perform a convergence analysis of CORSING $\mathcal{HS}$ applied to (3.46), showing that the mean error shares the same trend as the best s-term approximation error, as predicted by the theoretical results. In particular, the forcing term f is chosen such that the exact solution be

$$u(x) := C^*(1-x)(\exp(100x)-1),$$

where C^* is chosen such that $|u|_{H^1} = 1$. We take $L = 11$, corresponding to $N = 4095$. For $s = 4, 8, 16, 32$, we define $M = sN$ and $m = \lceil \overline{C} s^{\overline{\gamma}} \log M \log(N/s) \rceil$ for $\overline{\gamma} = 1, 2$, and for different values of $\overline{C}$. For every combination of $\overline{\gamma}$ and $\overline{C}$, we run 100 CORSING experiments and show the mean error obtained ± the standard deviation, computed using the unbiased estimator. In the case $\overline{\gamma} = 1$, we select $\overline{C} = 0.25, 0.5, 0.75$, whereas for $\overline{\gamma} = 2$, we consider $\overline{C} = 0.01, 0.03, 0.05$. The values of $\overline{C}$ are smaller for $\overline{\gamma} = 2$, in order to ensure that $m < N$ for every s.

The results are shown in Figure 3.8. The mean error reaches the best s-term approximation rate, which is proportional to $1/s$.

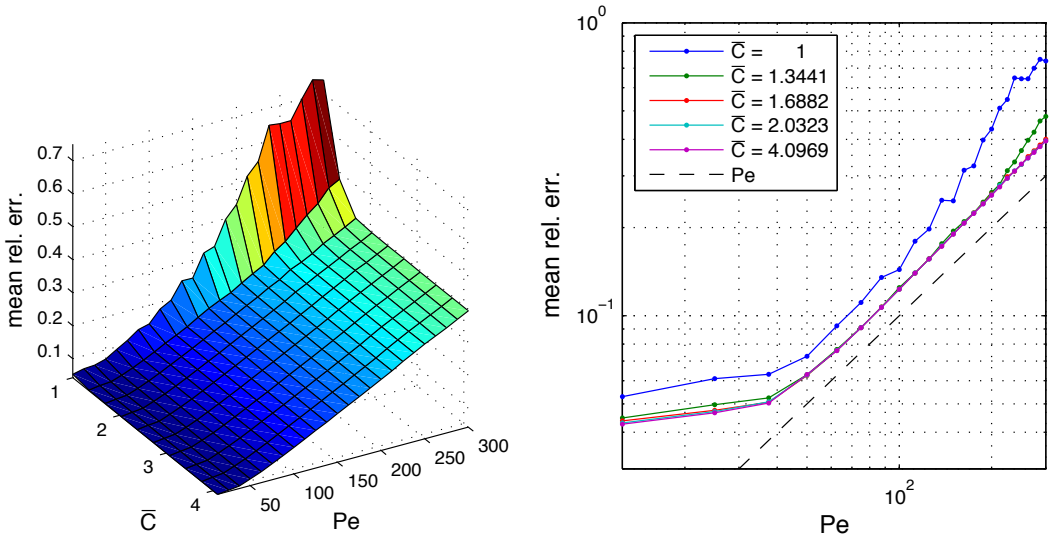


Figure 3.9: Surface plot of the mean relative error as a function of $\mathbb{P}\mathrm{e}$ and $\overline{C}$ (left) and logarithmic plot of the mean error as a function of $\mathbb{P}\mathrm{e}$, for different values of $\overline{C}$ (right).

3.4.4 Sensitivity analysis with respect to the Péclet number

In order to understand how the Péclet number influences the link between m and s, we consider the one-dimensional advection-diffusion problem (2.33). We focus on the CORSING $\mathcal{HS}$ case, and carry out a sensitivity analysis with respect to the constant $\overline{C}$ and to the Péclet number $\mathbb{P}\mathrm{e} = b/(2\eta)$, assuming $\overline{\gamma} = 1$.

Fix $L = 8$, corresponding to $N = 511$, $s = 20$, $M = N$, and define

$$\overline{C}_{\max} := \frac{N}{s\log(N)} \approx 4.10.$$

Then, for each $b = 25, 50, 75 \ldots, 575, 600$ and for each value of $\overline{C}$ in a grid of 10 equispaced points on $[1, \overline{C}_{\max}]$, we fix $m = \lceil \overline{C} s \log N \rceil$ and compute the empirical mean of the $H^1(\Omega)$-relative error over 200 runs. The repetitions of the test indices are avoided as described in Section 3.2.7, and the local coherence upper bound is chosen as $\nu_q^N \sim 1/q$.

In Figure 3.9 (left), we visualize the whole set of numerical experiments as a surface plot. The mean relative error is plotted as a function of $\mathbb{P}\mathrm{e}$ and $\overline{C}$. For values of $\overline{C}$ and $\mathbb{P}\mathrm{e}$ sufficiently large, the mean relative error grows linearly with respect to $\mathbb{P}\mathrm{e}$ and remains constant with respect to $\overline{C}$. In Figure 3.10 (right), the mean relative error is plotted as a function of $\mathbb{P}\mathrm{e}$ on logarithmic scale, considering the first four values of $\overline{C}$ and $\overline{C} = \overline{C}_{\max}$. The trend is linear for $\overline{C} \gtrsim 1.6882$ and $\mathbb{P}\mathrm{e} \gtrsim 125/2 = 62.5$. For smaller values of $\mathbb{P}\mathrm{e}$, the mean error grows sublinearly in $\mathbb{P}\mathrm{e}$.

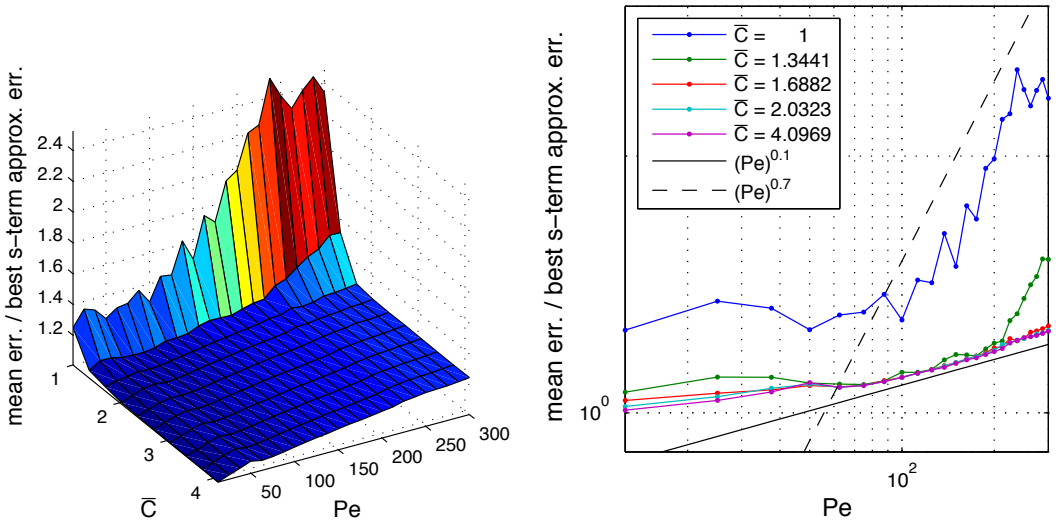


Figure 3.10: Surface plot of the ratio between the mean error and the best s-term approximation error as a function of $\mathbb{P}\mathrm{e}$ and $\overline{C}$ (left) and logarithmic plot of the same ratio as a function of $\mathbb{P}\mathrm{e}$, for different values of $\overline{C}$ (right).

In Figure 3.10 (left), we plot the ratio between the mean error and the best s-term approximation error as a function of $\mathbb{P}\mathrm{e}$ and $\overline{C}$. For $\overline{C}$ sufficiently large, the trend is constant with respect to $\overline{C}$, whereas the dependence on $\mathbb{P}\mathrm{e}$ is less clear. Analogously to Figure 3.9 (right), the same data are plotted in a logarithmic scale with respect to $\mathbb{P}\mathrm{e}$ for some fixed values of $\overline{C}$ in Figure 3.10 (right). For $\mathbb{P}\mathrm{e}$ sufficiently large, the ratio between the mean error and the best s-term approximation error with respect to $\mathbb{P}\mathrm{e}$ grows between the algebraic trends $(\mathbb{P}\mathrm{e})^{0.1}$ and $(\mathbb{P}\mathrm{e})^{0.7}$. For smaller values of $\mathbb{P}\mathrm{e}$, the growth is more moderate.

We conclude by comparing these numerical results with Proposition 3.23. Neglecting, for simplicity, the contributions of the truncation operator $\mathcal{T}_{\mathcal{K}}$ and of the constant ε, the proposition applied to problem (2.33) states that, provided that $m \gtrsim |\mathbb{P}\mathrm{e}|^2 s^2 \log(N)$, the CORSING error fulfills

$$\mathbb{E}[|\widehat{u} - u|_{H^1(\Omega)}] \leq C_{\mathbb{P}\mathrm{e}} |u^s - u|_{H^1(\Omega)}, \tag{3.63}$$

with $C_{\mathbb{P}\mathrm{e}} \sim |\mathbb{P}\mathrm{e}|$. However, considering that the ratio between the mean error and the best s-term approximation error is constant as a function of $\overline{C}$ (for $\overline{C}$ sufficiently large) and recalling Figure 3.10 (right), these asymptotic trends seem to be rather pessimistic. In practice, we can choose $m \gtrsim |\mathbb{P}\mathrm{e}| s \log N$, which is equivalent to restrict the surface in Figure 3.10 (left) to a direction $\overline{C} = \vartheta \mathbb{P}\mathrm{e}$, with $\vartheta > 0$. Then, the asymptotic constant in (3.63) fulfills

$$C_{\mathbb{P}\mathrm{e}} \sim |\mathbb{P}\mathrm{e}|^{\chi}, \quad \text{with } \chi \in [0.1, 0.7],$$

for $\mathbb{P}e$ sufficiently large. Notice that χ tends to 0.1 when ϑ grows larger and larger. This is promising, considering that we are not applying any stabilization technique.

Chapter 4

Further applications of CORSING

In the previous chapters, we applied the CORSING strategy to one- and two-dimensional ADR equations. Now, we deal with more challenging settings.

First, we assess the performances of CORSING on the *Stokes problem*, employing the trial and test bases presented in Chapter 2. In particular, we assess CORSING $\mathcal{SP}$. Then, we deal with ADR equations in dimension $d > 2$, by resorting to *tensorization* of the one-dimensional CORSING $\mathcal{HS}$ technique, providing a novel extension, the CORSING $\mathcal{QS}$. We furnish a two- and three-dimensional validation of CORSING $\mathcal{QS}$ and explain how to implement this strategy in a higher-dimensional scenario, by employing the theoretical concepts introduced in Chapter 3.

Analogously to Chapter 2, the goal of this chapter is to propose a numerical validation of CORSING on problems more interesting with a view to practical applications, while referring to a future work for a rigorous formalization.

Outline of the chapter In Section 4.1 we deal with the Stokes problem. Then, we generalize the CORSING $\mathcal{HS}$ method to ADR equations in dimension $d > 2$ in Section 4.2, employing tensorization.

4.1 The Stokes problem

In this section, we focus on the Stokes problem. In Section 4.1.1, we introduce the strong and the weak formulation of Stokes equations and present a test case with an analytical solution. Then, we derive a PG discretization (Section 4.1.2) and assess the performance of the full-PG approach, with particular attention to the stability of the method (Section 4.1.3). Finally, we apply the CORSING $\mathcal{SP}$ to the proposed test case in Section 4.1.4.

4.1.1 Problem setting

Let $\Omega = (0,1)^2$ and consider the *Stokes problem*

$$\begin{cases} -\Delta \mathbf{u} + \nabla p = \mathbf{f} & \text{in } \Omega, \\ \operatorname{div} \mathbf{u} = 0 & \text{in } \Omega, \\ \mathbf{u} = \mathbf{0} & \text{on } \partial\Omega. \end{cases} \tag{4.1}$$

These equations model the stationary flow of an incompressible liquid with unitary viscosity, subject to a body force $\mathbf{f} : \Omega \to \mathbb{R}^2$. The unknowns $\mathbf{u} : \Omega \to \mathbb{R}^2$ and $p : \Omega \to \mathbb{R}$ model the *velocity* and the *pressure* of the liquid, respectively. The first equation represents the *momentum equation*, whereas the second one is the *continuity equation*. The boundary condition employed is referred to as *no-slip* condition.

Weak formulation Problem (4.1) admits the following weak formulation

$$\text{find } (\mathbf{u}, p) \in [H_0^1(\Omega)]^2 \times L^2(\Omega) : \quad \begin{cases} a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) = \int_\Omega \mathbf{f} \cdot \mathbf{v} & \forall \mathbf{v} \in [H_0^1(\Omega)]^2 \\ b(\mathbf{u}, q) = 0 & \forall q \in L^2(\Omega), \end{cases} \tag{4.2}$$

with $a : [H_0^1(\Omega)]^2 \times [H_0^1(\Omega)]^2 \to \mathbb{R}$ and $b : [H_0^1(\Omega)]^2 \times L^2(\Omega) \to \mathbb{R}$ bilinear forms defined as

$$a(\mathbf{u}, \mathbf{v}) := \int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v} = \sum_{k=1}^{2} \int_\Omega \nabla u_k \cdot \nabla v_k, \quad \forall \mathbf{u}, \mathbf{v} \in [H_0^1(\Omega)]^2,$$

and

$$b(\mathbf{u}, p) := -\int_\Omega p \operatorname{div} \mathbf{u}, \quad \forall \mathbf{u} \in [H_0^1(\Omega)]^2, \ \forall p \in L^2(\Omega).$$

If the forcing term fulfills $\mathbf{f} \in [L^2(\Omega)]^2$, then problem (4.2) admits a unique solution $(\mathbf{u}, p) \in [H_0^1(\Omega)]^2 \times (L^2(\Omega)/\mathbb{R})$. In particular the pressure is unique up to an additive constant (see [BF91, Section IV.2]).

The Stokes equations are an example of a *saddle-point* problem, since they are equivalent to finding a saddle-point $(\mathbf{u}, p)$ to the functional

$$\mathcal{L}(\mathbf{v}, q) := \frac{1}{2} a(\mathbf{v}, \mathbf{v}) + b(\mathbf{v}, q) - (\mathbf{f}, \mathbf{v}).$$

The Stokes problem has been extensively studied in the last decades. For further details, we refer the reader to [BF91] and [Tem01].

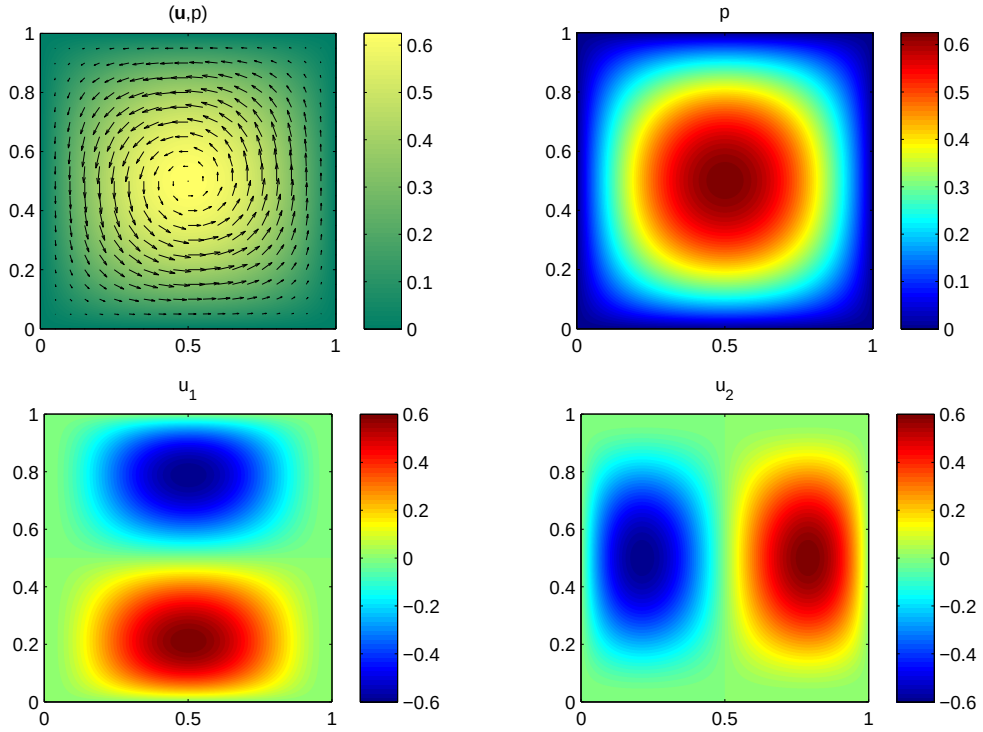


Figure 4.1: Exact solution to the Stokes problem (4.1), defined by (4.3)-(4.4).

The test case In this section, in order to assess the reliability of the CORSING method applied to the Stokes problem, we consider a test case with an analytic solution. The example is a slight modification of a test case in [Li09]. We define the following analytic expressions for the velocity

$$\mathbf{u}(\mathbf{x}) = \begin{bmatrix} 100x_1^2(x_1-1)^2x_2(x_2-1)(2x_2-1) \\ -100x_1(x_1-1)(2x_1-1)x_2^2(x_2-1)^2 \end{bmatrix}, \quad \forall \mathbf{x} \in \Omega, \tag{4.3}$$

and for the pressure

$$p(\mathbf{x}) = 10x_1(x_1-1)x_2(x_2-1), \quad \forall \mathbf{x} \in \Omega. \tag{4.4}$$

The exact solution is shown in Figure 4.1. The forcing term in (4.1) is computed as $\mathbf{f} = -\Delta\mathbf{u} + \nabla p$. We choose an exact pressure vanishing at the boundary, in order to employ the bases $\mathcal{P}^L$ and $\mathcal{S}^R$, defined in Section 2.4. As a "thought experiment", we can imagine the solution $(\mathbf{u}, p)$ defined by (4.3)-(4.4) as a metal in liquid state, subject to a magnetic forcing term $\mathbf{f}$, that makes the liquid rotate in a stationary way, free from any container.

4.1.2 Petrov-Galerkin discretization

In order to discretize the weak problem (4.2), we consider two finite dimensional subspaces $U^N, V^M \subseteq H_0^1(\Omega) \subseteq L^2(\Omega)$ of dimension N and M, respectively, such that

$$U^N = \text{span}\{\psi_j\}_{j\in[N]}, \quad V^M = \text{span}\{\varphi_r\}_{r\in[M]}.$$

Then, the resulting discrete PG formulation of (4.2) is

$$\text{find } (\widehat{\mathbf{u}}, \widehat{p}) \in [U^N]^2 \times U^N : \quad \begin{cases} a(\widehat{\mathbf{u}}, \mathbf{v}) + b(\mathbf{v}, \widehat{p}) = \int_\Omega \mathbf{f}\cdot\mathbf{v} & \forall \mathbf{v} \in [V^M]^2 \\ b(\widehat{\mathbf{u}}, q) = 0 & \forall q \in V^M. \end{cases}$$

Then, evaluating the momentum and the mass conservation equations of (4.2) on the trial and test functions, yields a linear system of dimension $3M \times 3N$. Indeed, expanding the velocity and the pressure with respect to the trial basis yields

$$\widehat{u_1} = \sum_{j_1\in[N]} \widehat{u}_{1,j_1}\psi_{j_1}, \quad \widehat{u_2} = \sum_{j_2\in[N]} \widehat{u}_{2,j_2}\psi_{j_2}, \quad \widehat{p} = \sum_{j_3\in[N]} \widehat{p}_{j_3}\psi_{j_3}.$$

Then, the first M linear equations are obtained considering the momentum equation and a test function of the form $\mathbf{v} = [\varphi_r, 0]^\intercal$, namely,

$$\sum_{j_1\in[N]} \widehat{u}_{1,j_1} \int_\Omega \nabla\psi_{j_1}\cdot\nabla\varphi_r - \sum_{j_3\in[N]} \widehat{p}_{j_3} \int_\Omega \psi_{j_3}\frac{\partial\varphi_r}{\partial x_1} = \int_\Omega f_1\varphi_r, \quad \forall r \in [M]. \tag{4.5}$$

Analogously, considering the momentum equation and a test function of the form $\mathbf{v} = [0, \varphi_r]^\intercal$ yields the second set of M linear equations

$$\sum_{j_2\in[N]} \widehat{u}_{2,j_2} \int_\Omega \nabla\psi_{j_2}\cdot\nabla\varphi_r - \sum_{j_3\in[N]} \widehat{p}_{j_3} \int_\Omega \psi_{j_3}\frac{\partial\varphi_r}{\partial x_2} = \int_\Omega f_2\varphi_r, \quad \forall r \in [M]. \tag{4.6}$$

The third set of M linear equations is obtained by considering the mass conservation and a test function of the form $q = \varphi_r$, namely,

$$-\sum_{j_1\in[N]} \widehat{u}_{1,j_1} \int_\Omega \frac{\partial\psi_{j_1}}{\partial x_1}\varphi_r - \sum_{j_2\in[N]} \widehat{u}_{2,j_2} \int_\Omega \frac{\partial\psi_{j_2}}{\partial x_2}\varphi_r = 0, \quad \forall r \in [M]. \tag{4.7}$$

Finally, we notice that, thanks to the regularity of the trial and test functions, and to the fact that they vanish on the boundary $\partial\Omega$, the following relations hold

$$\int_\Omega \psi_j \frac{\partial\varphi_r}{\partial x_k} = -\int_\Omega \frac{\partial\psi_j}{\partial x_k}\varphi_r, \quad k = 1, 2, \ \forall j \in [N], \ \forall r \in [M].$$

Collecting the equations (4.5), (4.6) and (4.7), yields the $3M \times 3N$ linear system

$$\begin{bmatrix} \mathbf{L} & \mathbf{0} & \mathbf{T}_1 \\ \mathbf{0} & \mathbf{L} & \mathbf{T}_2 \\ \mathbf{T}_1 & \mathbf{T}_2 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \mathbf{0} \end{bmatrix}, \tag{4.8}$$

with $\mathbf{u}_k = (u_{k,j})_{j\in[N]}$, for $k = 1, 2$, $\mathbf{p} = (p_j)_{j\in[N]}$ and where $\mathbf{L}, \mathbf{T}_1, \mathbf{T}_2 \in \mathbb{R}^{M\times N}$ are defined as

$$L_{rj} = \int_\Omega \nabla\psi_j \cdot \nabla\varphi_r, \quad (T_k)_{rj} = \int_\Omega \frac{\partial\psi_j}{\partial x_k}\varphi_r, \quad \forall j \in [N],\ \forall r \in [M],\ k = 1, 2.$$

In practice, when considering the full-PG $\mathcal{PS}$ approach, we choose $U^N = \text{span}(\mathcal{P}^L)$, with $N = (2^{L+1} - 1)^2$ and $V^M = \text{span}(\mathcal{S}^R)$, with $M = R^2$, where $\mathcal{P}^L$ and $\mathcal{S}^R$ are the bases defined in Section 2.4. In the full-PG $\mathcal{SP}$ case, the role of the trial and test functions is inverted. In particular, we provide an *equal-order* approximation for the velocity and the pressure.

Remark 4.1.1. We normalize the trial and test functions for the pressure with respect to the $L^2(\Omega)$-norm. In particular, explicit computations show that

$$\|\mathcal{P}_{\ell,\mathbf{k}}\|_{L^2(\Omega)} = 2^{-(\ell+\frac{5}{2})}, \quad \|\mathcal{S}_{\mathbf{r}}\|_{L^2(\Omega)} = \frac{1}{\pi|\mathbf{r}|}, \quad \forall \ell, \mathbf{k}, \mathbf{r},$$

with $\mathcal{P}_{\ell,\mathbf{k}}$ and $\mathcal{S}_{\mathbf{r}}$ defined as in (2.34) and (2.35), respectively. This is equivalent to pre- and post-multiplying the full-PG stiffness matrix in (4.8) by suitable diagonal matrices. Numerical evidence shows that this strategy is able to considerably reduce the condition number of the stiffness matrix, thus making the procedure more stable (see also Remark 4.1.2). □

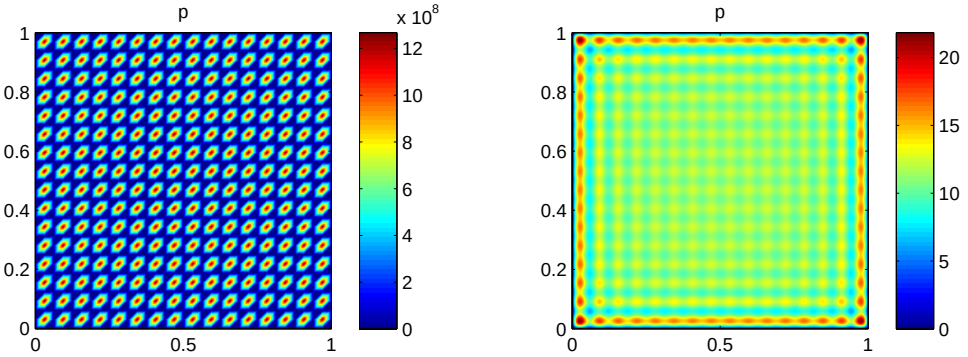
Finally, we notice that the discretization presented here can be generalized by considering three different trial spaces relative to u_1, u_2 and p, of different dimensions, instead of a unique trial space U^N. The same generalization holds for the test functions.

4.1.3 Numerical assessment of full-PG

We assess the performances of full-PG $\mathcal{PS}$ and of full-PG $\mathcal{SP}$ on problem (4.1) with exact solution defined as in (4.3)-(4.4) and with $M = N$. In particular, we set $L = 4$ and $R = 2^{L+1} - 1 = 31$, corresponding to $N = M = 961$. The $L^2(\Omega)$-norm of the relative error on the velocity and on the pressure is shown in Table 4.1, in the columns labeled as "$M = N$". The velocity is well captured, but, unfortunately, the computed pressure is totally unreliable, especially for the choice $\mathcal{PS}$. In Figure 4.2 we plot the pressure computed by the two approaches. In

	full-PG $\mathcal{PS}$		full-PG $\mathcal{SP}$	
	$M = N$	$M = 4N$	$M = N$	$M \approx 4.1N$
$\widehat{u_1}$	3.2e–03	4.0e–03	1.5e–03	9.0e–04
$\widehat{u_2}$	3.2e–03	4.0e–03	1.7e–03	9.0e–04
$\widehat{p}$	1.3e+09	2.0e–02	3.3e+01	5.8e–03

Table 4.1: Relative errors with respect to $L^2(\Omega)$-norm for full-PG $\mathcal{PS}$ and $\mathcal{SP}$ applied to the Stokes problem (4.1) with exact solution (4.3)-(4.4), $N = 961$ and different values of M.



(a) (b)

Figure 4.2: Pressure computed with the full-PG approach, with $M = N$: $\mathcal{PS}$ (left) and $\mathcal{SP}$ (right) approach.

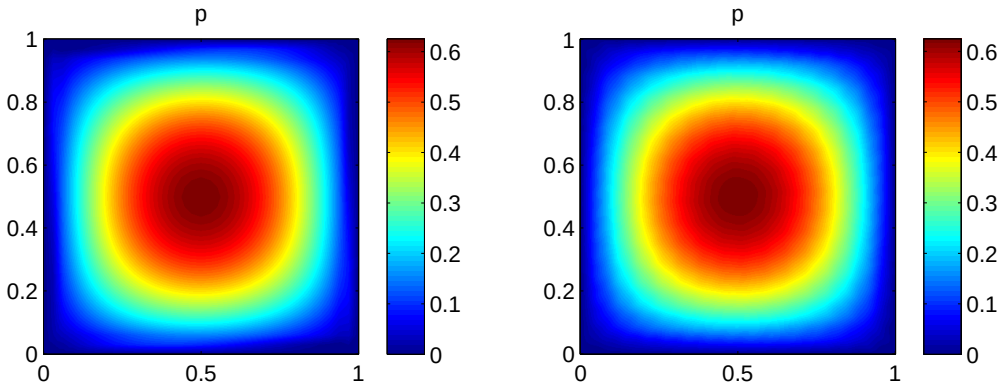
both cases, a strong instability occurs. This depends on the fact that the stiffness matrix is singular in both cases, having a non-trivial kernel of *spurious pressures*, namely

$$\exists p_0 \in U^N \setminus \{0\} : \quad b(p_0, \mathbf{v}) = 0, \quad \forall \mathbf{v} \in [V^M]^2. \tag{4.9}$$

In particular, it turns out that the subspace of the pressures satisfying (4.9) is one-dimensional in both the $\mathcal{SP}$ and $\mathcal{PS}$ cases (employing the command `null` of MATLAB®).

Due to the particular shape of the spurious pressures, this phenomenon can be classified as a *checkerboard instability*, already observed in the PG discretization of the Stokes problem (see, e.g., [EG13, Section 4.2.3]). Figure 4.2 clearly highlights this phenomenon.

Enriching the test space In order to overcome the checkerboard instability, we enrich the test space by choosing $M > N$. In Table 4.1 we show the results for both the full-PG $\mathcal{PS}$ and $\mathcal{SP}$ approaches, with $M = 4N$ and $M \approx 4.1N$, respectively. In particular, in the $\mathcal{PS}$ case, we set $L = 4$ and $R = 2(2^{L+1} - 1) = 62$, corresponding to $N = 961$ and $M = 3844$. In the $\mathcal{SP}$ case, we let $R = 31$ and $L = 5$,



(a) (b)

Figure 4.3: Pressure computed with the full-PG approach, with $M > N$: $\mathcal{PS}$ (left) and $\mathcal{SP}$ (right) approach.

corresponding to $N = 961$ and $M = 3969$. In both cases, the resulting full-PG solution is computed using the backslash \ command of MATLAB®, i.e., solving a least-squares problem. The checkerboard instability disappears in both cases, and the pressure is well-captured (see Figure 4.3).

Remark 4.1.2. With reference to Remark 4.1.1, we compare the condition number of the full-PG stiffness matrix in (4.8) before and after the $L^2(\Omega)$-normalization of the trial and test functions associated with the pressure, for $M > N$. In the $\mathcal{PS}$ case, after the $L^2(\Omega)$-normalization, the condition number decreases from $8.2 \cdot 10^2$ to 66.2, whereas in the $\mathcal{SP}$ case it decreases from $4.3 \cdot 10^4$ to 33.1. In both cases, the benefit due to the normalization is evident. This comparison is meaningless when $M = N$, since the stiffness matrix is singular. □

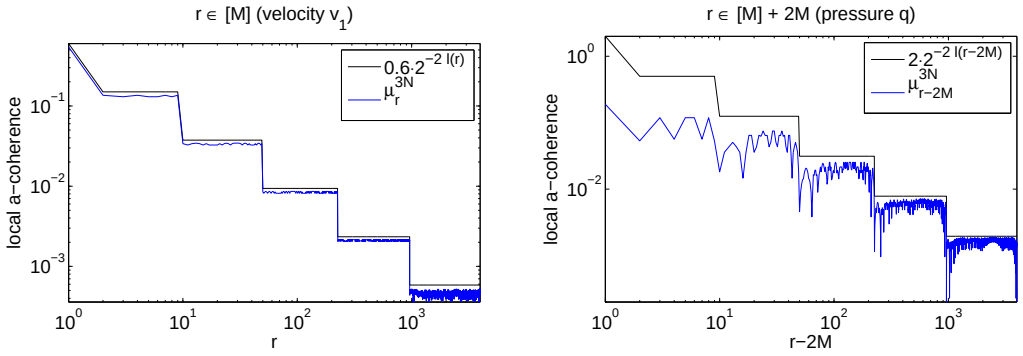
4.1.4 Numerical assessment of CORSING $\mathcal{SP}$

Finally, we carry out a numerical assessment of the CORSING $\mathcal{SP}$ strategy, choosing $R = 31$ and $L = 5$, corresponding to the full-PG approach with $N = 961$ and $M = 3969$. First, in Figure 4.4 we numerically show that the vector

$$\nu_r^{3N} := \begin{cases} 0.6 \cdot 2^{-2\ell(r)} & \forall r \in [M], \\ 0.6 \cdot 2^{-2\ell(r-M)} & \forall r \in [M] + M, \\ 2 \cdot 2^{-2\ell(r-2M)} & \forall r \in [M] + 2M, \end{cases}$$

is an upper bound to the local a-coherence, for $N = 961$ and $M = 3844$.

Notice that we need different upper bounds for the velocities and for the pressure test indices. Moreover, we plot only the tests relative to v_1 (the first



(a) (b)

Figure 4.4: Upper bound to the local a-coherence for CORSING $\mathcal{SP}$ applied to problem (4.1), for the velocity, (a) and the pressure, (b).

M rows of the full-PG stiffness matrix), since the results for the tests associated with v_2 (the second group of M rows of the full-PG stiffness matrix) are identical.

In Figure 4.5, we show the results obtained after 50 runs of CORSING $\mathcal{SP}$. We choose TS = 50%, 60%, 70%, corresponding to $m = 1441, 1153, 864$, respectively. The sparsity level is $s = 100$ in all three cases and the test selection is carried out avoiding repetitions. The ESP values are very good, and the velocities are also very well reconstructed. However, observing the pressure of the worst solution in the successful cluster (Figure 4.5, right) we notice significant oscillations for TS = 60% and, especially, for TS = 70%. A deeper understanding of this phenomenon is currently under investigation.

4.2 Multi-dimensional ADR problems

In this section, we deal with ADR problems of the form (2.2) in dimension $d > 2$. We generalize the CORSING $\mathcal{HS}$ approach presented in Chapter 2 to higher dimensions employing *tensorization* for both the trial and the test spaces. The resulting strategy is named CORSING $\mathcal{QS}$. We choose to adopt tensorization since, on the one hand, it allows for an immediate generalization of the local a-coherence estimates from the one-dimensional to the d-dimensional case; on the other hand, the assembly of the resulting stiffness matrix is easily implemented, thanks to the algebraic properties of the Kronecker product.

In Section 4.2.1 we deal with tensorization for general trial and test bases and we specialize this approach to the case of hierarchical hat functions and sine functions in Section 4.2.2. Afterwards, we present local a-coherence estimates for the d-dimensional case and introduce a tensorized strategy for the selection of the test functions in Section 4.2.3. In Section 4.2.4 we provide an

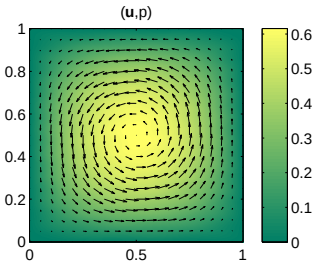
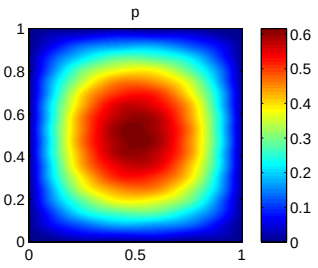
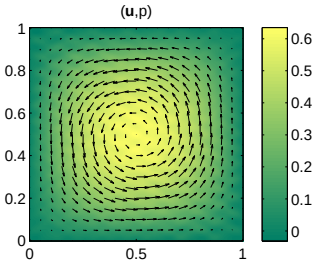
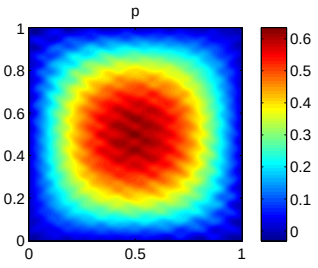
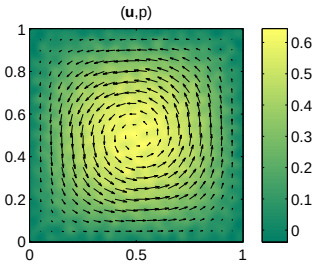
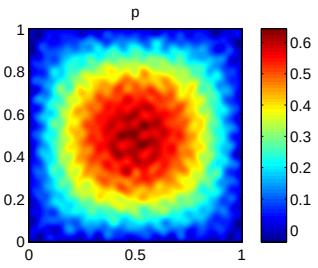
TS = 50% m = 1441 ESP = 1.00 L^2-rel. err. $\widehat{u_1}$: 1.0e–03 ± 5.7e–05 $\widehat{u_2}$: 9.7e–04 ± 5.7e–05 $\widehat{p}$: 1.9e–02 ± 2.6e–03		
TS = 60% m = 1153 ESP = 1.00 L^2-rel. err. $\widehat{u_1}$: 1.1e-03 ± 2.1e-04 $\widehat{u_2}$: 1.1e-03 ± 2.8e-04 $\widehat{p}$: 2.0e-02 ± 5.7e-03		
TS = 70% m = 864 ESP = 1.00 L^2-rel. err. $\widehat{u_1}$: 1.4e-03 ± 5.3e-04 $\widehat{u_2}$: 1.5e-03 ± 5.6e-04 $\widehat{p}$: 3.8e-02 ± 1.2e-02		

Figure 4.5: Assessment of CORSING $\mathcal{SP}$ on the Stokes problem (4.1), with exact solution (4.3)-(4.4): statistical analysis of the results (left), solution with the worst pressure obtained over 50 runs (right).

analytical proof of the classical inf-sup property for the full-PG $\mathcal{QS}$ discretization in the case of the two-dimensional Poisson equation. Finally, a numerical assessment of the CORSING $\mathcal{QS}$ approach is carried out in the two-dimensional (Section 4.2.5) and the three-dimensional (Section 4.2.6) case.

4.2.1 Tensorization

Let $N_1, M_1 \in \mathbb{N}$ and consider two bases $\{\psi_j\}_{j\in[N_1]}$ and $\{\varphi_q\}_{q\in[M_1]}$ defined over the one-dimensional domain $(0,1)$. Then, considering the d-dimensional domain $\Omega = (0,1)^d$, we define the tensorized trial basis as

$$\widehat{\psi_{\mathbf{j}}} := \bigotimes_{k=1}^{d} \psi_{j_k}, \quad \forall \mathbf{j} \in [N_1]^d,$$

and the tensorized test basis as

$$\widehat{\varphi_{\mathbf{q}}} := \bigotimes_{k=1}^{d} \varphi_{q_k}, \quad \forall \mathbf{q} \in [M_1]^d, \tag{4.10}$$

where, $\otimes$ denotes the standard tensor product of functions, defined as

$$\left(\bigotimes_{k=1}^{d} f_k\right)(\mathbf{x}) := \prod_{k=1}^{d} f_k(x_k), \quad \forall \mathbf{x} \in [0,1]^d.$$

Moreover, we denote $N := N_1^d$ and $M := M_1^d$.

Considering the generic ADR problem (2.2) with $\Omega = (0,1)^d$, we compute the explicit expression for the stiffness matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ associated with the full-PG discretization with respect to two tensorized families of trial and test functions.

First, we recall the definition of *Kronecker product* for matrices.

Definition 4.1. Given two matrices $\mathbf{X} \in \mathbb{R}^{k\times\ell}, \mathbf{Y} \in \mathbb{R}^{p\times q}$, their *Kronecker product* is defined as

$$\mathbf{X}\otimes\mathbf{Y} := \begin{bmatrix} X_{11}\mathbf{Y} & \cdots & X_{1\ell}\mathbf{Y} \\ \vdots & \ddots & \vdots \\ X_{k1}\mathbf{Y} & \cdots & X_{k\ell}\mathbf{Y} \end{bmatrix} \in \mathbb{R}^{kp\times\ell q}.$$

The next lemma shows how to rewrite the stiffness matrix associated with the d-dimensional problem (2.2), as a suitable combination of Kronecker products of three stiffness matrices $\mathbf{L}, \mathbf{T}, \mathbf{R}$ of dimension $M_1 \times N_1$, associated with the one-dimensional Laplace, the pure transport, and the pure reactive equation, respectively.

Lemma 4.2. *Consider the d-dimensional ADR problem (2.2) on $\Omega = (0,1)^d$, with constant data $\eta, \rho \in \mathbb{R}$ and $\mathbf{b} \in \mathbb{R}^d$. Then, the stiffness matrix associated with the full-PG discretization with respect to the bases $\{\widehat{\psi}_{\mathbf{j}}\}_{\mathbf{j}\in[N_1]^d}$ and $\{\widehat{\varphi}_{\mathbf{q}}\}_{\mathbf{q}\in[M_1]^d}$ is given by*

$$\mathbf{A} := \eta \sum_{k=1}^{d} \mathbf{R}^{(k-1)\otimes} \otimes \mathbf{L} \otimes \mathbf{R}^{(d-k)\otimes} + \sum_{k=1}^{d} b_k \mathbf{R}^{(k-1)\otimes} \otimes \mathbf{T} \otimes \mathbf{R}^{(d-k)\otimes} + \rho \mathbf{R}^{d\otimes},$$

where $\mathbf{L}, \mathbf{T}, \mathbf{R} \in \mathbb{R}^{M_1 \times N_1}$ are defined as

$$L_{qj} := (\psi_j', \varphi_q'), \quad T_{qj} := (\psi_j', \varphi_q), \quad R_{qj} := (\psi_j, \varphi_q), \quad \forall j \in [N_1], \forall q \in [M_1],$$

and

$$\mathbf{R}^{n\otimes} := \underbrace{\mathbf{R} \otimes \cdots \otimes \mathbf{R}}_{n \text{ times}}, \quad \forall n \in \mathbb{N}.$$

Proof. First, with straightforward calculations, it is possible to show that the following relations hold

$$\begin{aligned}
(\nabla\widehat{\psi}_{\mathbf{j}}, \nabla\widehat{\varphi}_{\mathbf{q}}) &= \sum_{k=1}^{d} (\psi_{j_k}', \varphi_{q_k}') \prod_{\ell\neq k} (\psi_{j_\ell}, \varphi_{q_\ell}), \\
(\mathbf{b}\cdot\nabla\widehat{\psi}_{\mathbf{j}}, \widehat{\varphi}_{\mathbf{q}}) &= \sum_{k=1}^{d} b_k (\psi_{j_k}', \varphi_{q_k}) \prod_{\ell\neq k} (\psi_{j_\ell}, \varphi_{q_\ell}), \quad \forall \mathbf{b} \in \mathbb{R}^d, \\
(\widehat{\psi}_{\mathbf{j}}, \widehat{\varphi}_{\mathbf{q}}) &= \prod_{k=1}^{d} (\psi_{j_k}, \varphi_{q_k}).
\end{aligned}$$

Then, recalling that the entries of $\mathbf{A}$ associated with problem (2.2) are defined by

$$a(\widehat{\psi}_{\mathbf{j}}, \widehat{\varphi}_{\mathbf{q}}) = \eta(\nabla\widehat{\psi}_{\mathbf{j}}, \nabla\widehat{\varphi}_{\mathbf{q}}) + (\mathbf{b}\cdot\nabla\widehat{\psi}_{\mathbf{j}}, \widehat{\varphi}_{\mathbf{q}}) + \rho(\widehat{\psi}_{\mathbf{j}}, \widehat{\varphi}_{\mathbf{q}}),$$

using the Kronecker product for matrices, and the lexicographic ordering for the d-dimensional trial and test bases, we obtain the thesis. □

4.2.2 The $\mathcal{QS}$ trial and test combination

We introduce a new combination of trial and test functions, denoted $\mathcal{QS}$. This is built up with the tensorization strategy presented above, with the initial choice corresponding to the $\mathcal{HS}$ approach, introduced in Chapter 2.

For every $(\mathbf{l}, \mathbf{k}) \in \mathbb{N}^d \times \mathbb{N}^d$, with $0 \le \mathbf{k} \le 2^{\mathbf{l}} - 1$ (the operations and the inequalities are understood component-wise), we have

$$\mathcal{Q}_{\mathbf{l},\mathbf{k}}(\mathbf{x}) := \prod_{i=1}^{d} \mathcal{H}_{\ell_i,k_i}(x_i), \quad \forall \mathbf{x} \in [0,1]^d.$$

In order to fulfil Hypothesis 1, introduced in Section 3.1.2, we normalize the tensorized test family. In fact, the family $\{\widehat{\varphi}_{\mathbf{q}}\}$, defined in (4.10) with $\varphi_q = \mathcal{S}_q$, is already orthogonal with respect to the $H^1(\Omega)$-inner product $(\nabla u, \nabla v)$, and the $H^1(\Omega)$-seminorm can be easily computed as

$$|\widehat{\varphi}_{\mathbf{q}}|^2_{H^1(\Omega)} = (\nabla\widehat{\varphi}_{\mathbf{q}}, \nabla\widehat{\varphi}_{\mathbf{q}}) = \sum_{k=1}^{d} (\mathcal{S}'_{q_k}, \mathcal{S}'_{q_k}) \prod_{\ell\neq k} (\mathcal{S}_{q_\ell}, \mathcal{S}_{q_\ell})$$
$$= \sum_{k=1}^{d} \prod_{\ell\neq k} \frac{1}{(\pi q_\ell)^2} = \frac{\|\mathbf{q}\|_2^2}{\pi^{2(d-1)} \prod_{k=1}^{d} q_k^2}.$$

Hence, we obtain the normalized sine functions

$$\mathcal{S}_{\mathbf{q}}(\mathbf{x}) = \frac{\pi^{d-1}}{\|\mathbf{q}\|_2} \prod_{k=1}^{d} q_k \mathcal{S}_{q_k}(x_k), \quad \forall \mathbf{x} \in [0,1]^d.$$

The tensorized family of normalized sine functions with $M_1 = R$ is denoted

$$\mathcal{S}^R := \{\mathcal{S}_{\mathbf{q}} : \|\mathbf{q}\|_\infty \leq R\},$$

according to the definition given in Chapter 2 for the two-dimensional case.

For a fixed $L \in \mathbb{N}$ and chosen $N_1 = 2^{L+1} - 1$, we denote the resulting trial family of tensorized hierarchical hat functions as

$$\mathcal{Q}^L := \{\mathcal{Q}_{\mathbf{l},\mathbf{k}} : 0 \leq \mathbf{k} \leq 2^{\mathbf{l}} - 1,\ \|\mathbf{l}\|_\infty \leq L\}.$$

The basis $\mathcal{Q}^1$ is plotted in Figure 4.6. The reason for the letter $\mathcal{Q}$ for the trial basis is due to the fact that $\mathcal{Q}^L$ spans the space of piecewise bilinear continuous function over a regular Cartesian grid of step $h = 2^{-L-1}$ on $[0,1]^d$ and, in the FE literature, the standard Lagrangian basis functions associated with this space is usually referred to as $\mathbb{Q}_1$ elements (see, e.g., [EG13]).

We also notice that the basis $\mathcal{Q}^L$ has already been employed for the numerical approximation of PDEs, e.g., in the context of *Sparse Grids* (see [BG04]).

Unfortunately, the basis $\mathcal{Q}^L$ is not orthogonal with respect to the inner product $(\nabla u, \nabla v)$ and, at the moment, estimating a good lower bound for the quantity κ_s defined in (3.23) is an open issue.

Remark 4.2.1. For practical purposes such as the post-processing of the solution, the hierarchical basis $\mathcal{Q}^L$ is very easy to handle. Indeed, we can compute the values of the functions at the Cartesian grid knots by applying a sparse matrix, having a Kronecker product structure. In particular, in the one-dimensional case, for every $\mathbf{w} \in \mathbb{R}^{N_1}$, with $N_1 = 2^{L+1} - 1$, we define

$$[\mathbf{Tw}]_{j+1} = (\Psi\mathbf{w})(jh), \quad \forall j = 0, \ldots, N_1 + 1, \tag{4.11}$$

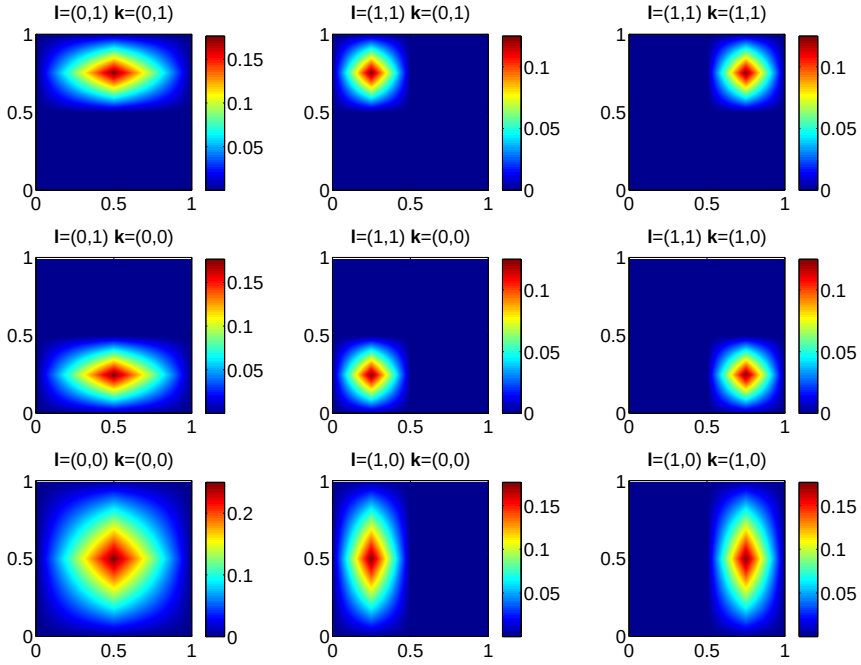


Figure 4.6: The basis $\mathcal{Q}^1$.

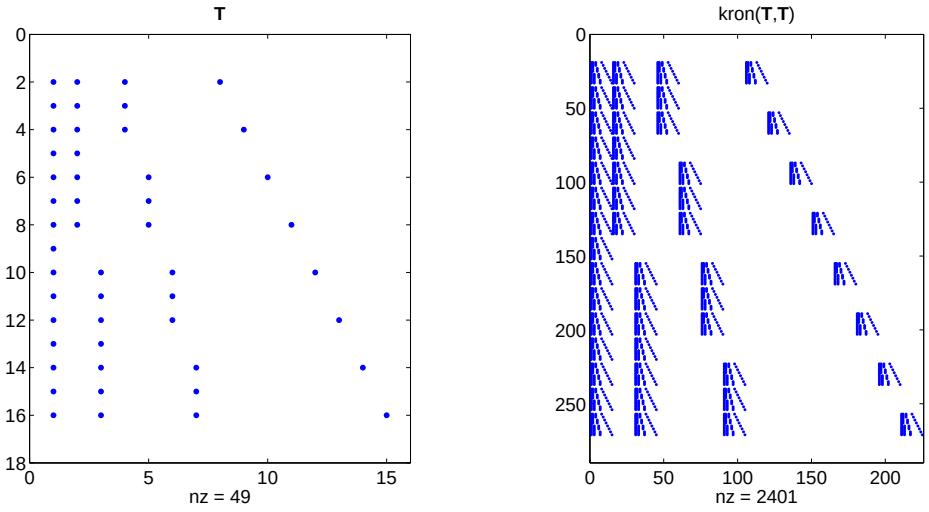


Figure 4.7: Sparsity pattern of the matrix $\mathbf{T}$ defined in (4.11) (left) and $\mathbf{T} \otimes \mathbf{T}$ (right).

where Ψ is the reconstruction operator (see Definition 3.2) and $h = 2^{-L-1}$. Then, in dimension d, the transformation matrix is simply obtained as $\mathbf{T}^{d\otimes}$. The coefficients in **Tw** are extremely simple to handle in MATLAB® using efficient built-in functions such as, e.g., `reshape`, `interp1`, `interp2`, `interp3` and `interpn`. In Figure 4.7 the sparsity pattern of $\mathbf{T}$ and of $\mathbf{T}\otimes\mathbf{T}$, for $L = 3$, are shown. □

4.2.3 Local a-coherence upper bound and tensorized randomization

In the $\mathcal{QS}$ case in dimension d, the local a-coherence can be estimated as follows. For the sake of simplicity, let us focus on the case of the d-dimensional Laplace equation (in the ADR case, analogous considerations can be made). Starting from the relation

$$(\nabla\mathcal{Q}_{\mathbf{l},\mathbf{k}}, \nabla\mathcal{S}_{\mathbf{q}}) = \frac{\pi^{d-1}}{\|\mathbf{q}\|_2}\prod_{i=1}^{d} q_i \cdot \sum_{i=1}^{d}(\mathcal{H}'_{\ell_i,k_i}, \mathcal{S}'_{q_i})\prod_{j\neq i}(\mathcal{H}_{\ell_j,k_j}, \mathcal{S}_{q_j}),$$

recalling the one-dimensional upper bounds

$$\begin{aligned}
&|(\mathcal{H}'_{\ell,k}, \mathcal{S}'_q)|^2 \lesssim \frac{1}{q},\\
&|(\mathcal{H}_{\ell,k}, \mathcal{S}_q)|^2 \sim \frac{1}{q^4}|(\mathcal{H}_{\ell,k}, \mathcal{S}''_q)|^2 = \frac{1}{q^4}|(\mathcal{H}'_{\ell,k}, \mathcal{S}'_q)|^2 \lesssim \frac{1}{q^5},
\end{aligned}$$

and using the inequality

$$\left(\sum_{i=1}^{d} x_i\right)^2 \leq d\sum_{i=1}^{d} x_i^2, \quad \forall \mathbf{x}\in\mathbb{R}^d,$$

we obtain the following d-dimensional estimate, independent of $(\mathbf{l},\mathbf{k})$,

$$\begin{aligned}
|(\nabla\mathcal{Q}_{\mathbf{l},\mathbf{k}}, \nabla\mathcal{S}_{\mathbf{q}})|^2 &\lesssim \frac{d\pi^{2(d-1)}}{\|\mathbf{q}\|_2^2}\prod_{i=1}^{d} q_i^2 \cdot \sum_{i=1}^{d}|(\mathcal{H}'_{\ell_i,k_i}, \mathcal{S}'_{q_i})|^2\prod_{j\neq i}|(\mathcal{H}_{\ell_j,k_j}, \mathcal{S}_{q_j})|^2\\
&\lesssim \frac{d\pi^{2(d-1)}}{\|\mathbf{q}\|_2^2}\prod_{i=1}^{d} q_i^2 \cdot \sum_{i=1}^{d}\frac{1}{q_i}\prod_{j\neq i}\frac{1}{q_j^5}\\
&= d\pi^{2(d-1)}\sum_{i=1}^{d}\frac{q_i}{\|\mathbf{q}\|_2^2}\prod_{j\neq i}\frac{1}{q_j^3}.
\end{aligned}$$

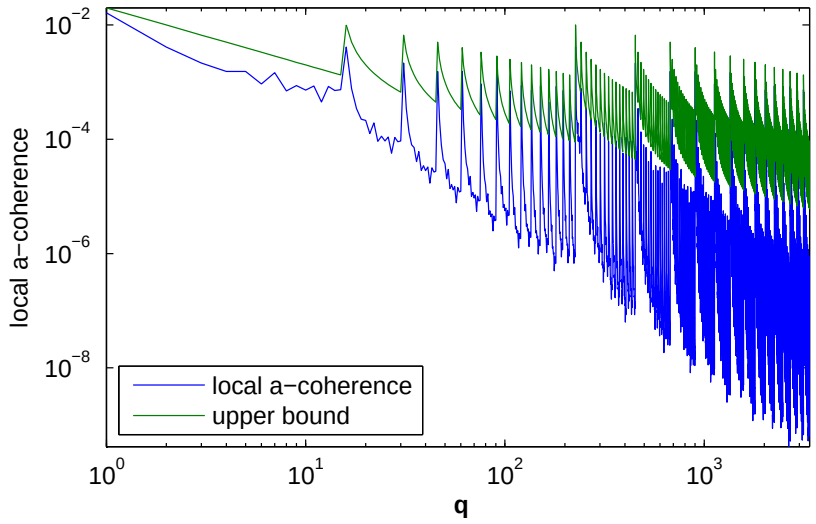


Figure 4.8: Local a-coherence upper bound (4.14) for the three-dimensional Poisson problem.

Now, exploiting the inequality $q_i/\|\mathbf{q}\|_2^2 \leq 1/q_i$, for every $i \in [d]$, we obtain

$$|(\nabla \mathcal{Q}_{\mathbf{l},\mathbf{k}}, \nabla \mathcal{S}_{\mathbf{q}})|^2 \lesssim d\pi^{2(d-1)} \sum_{i=1}^{d} \frac{1}{q_i} \prod_{j \neq i} \frac{1}{q_j^3}$$

$$= d\pi^{2(d-1)} \prod_{i=1}^{d} \frac{1}{q_i} \cdot \sum_{i=1}^{d} \prod_{j \neq i} \frac{1}{q_j^2} \lesssim d^2 \pi^{2(d-1)} \prod_{i=1}^{d} \frac{1}{q_i}. \tag{4.12}$$

This leads us to the choice

$$\nu_{\mathbf{q}}^{N} \sim \prod_{i=1}^{d} \frac{1}{q_i}, \tag{4.13}$$

that is far from being optimal (recall that the theory presented in Chapter 3 does not require $\boldsymbol{\nu}^{N,M}$ to be sharp), but it has a very interesting advantage: $\nu_{\mathbf{q}}^{N}$ has a *separable* form, i.e., it is a tensor product of one-dimensional upper bounds $\nu_q^{N_1} \sim 1/q$.

Remark 4.2.2. Observing estimate (4.12), we notice that the choice of the upper bound (4.13) hides a constant factor $d^2\pi^{2(d-1)}$, that depends exponentially on d. In other words, we face the *curse of dimensionality*. Nevertheless, this estimate seems to be very pessimistic for moderate values of d. In Figure 4.8 we compare upper bound (4.13) with the local a-coherence, showing in particular that, for $d = 3$ and $L = 3$, corresponding to $N_1 = 15$ and $N = 3375$, it holds

$$\mu_{\mathbf{q}}^{3375} \leq \frac{0.02}{q_1 q_2 q_3}, \quad \forall \mathbf{q} \in [15]^3, \tag{4.14}$$

in the case of the Poisson problem (the multi-indices $\mathbf{q}$ are ordered lexicographically). How to deal with the case $d \gg 1$, remains an open issue. □

Tensorized randomization Thanks to the separable form of the upper bound in (4.13), we apply the following heuristic strategy for the test selection procedure:

1. choose m of the form $m = m_1^d$, with $m_1 \in \mathbb{N}$;
2. for every $k = 1, \dots, d$, draw m_1 indices $\tau_1^k, \dots, \tau_{m_1}^k$ at random according to the probability density $p_q \sim 1/q$ and define $\mathcal{T}^k := \{\tau_1^k, \dots, \tau_{m_1}^k\}$;
3. define the set of m selected multi-indices as $\mathcal{T} := \mathcal{T}^1 \times \cdots \times \mathcal{T}^d$.

This *grid* structure of the random multi-indices in $\mathcal{T}$ allows for an efficient assembly of the stiffness matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ as a suitable combination of Kronecker product of $m_1 \times N_1$ submatrices of $\mathbf{L}, \mathbf{T}, \mathbf{R}$ defined in Lemma 4.2. For example, in the case of the three-dimensional Poisson equation, we have

$$\mathbf{A} = \mathbf{L}_{\mathcal{T}^1} \otimes \mathbf{R}_{\mathcal{T}^2} \otimes \mathbf{R}_{\mathcal{T}^3} + \mathbf{R}_{\mathcal{T}^1} \otimes \mathbf{L}_{\mathcal{T}^2} \otimes \mathbf{R}_{\mathcal{T}^3} + \mathbf{R}_{\mathcal{T}^1} \otimes \mathbf{R}_{\mathcal{T}^2} \otimes \mathbf{L}_{\mathcal{T}^3},$$

where $\mathbf{L}_{\mathcal{T}^k}$ is the submatrix of $\mathbf{L}$ identified by the rows in $\mathcal{T}^k$ and $\mathbf{R}_{\mathcal{T}^k}$ is defined analogously, for $k = 1, 2, 3$.

Due to this particular structure of $\mathbf{A}$, we can avoid its storage, implementing only the matrix-vector multiplication and making use of the algebraic property

$$(\mathbf{X} \otimes \mathbf{Y})\mathrm{vec}(\mathbf{U}) = \mathrm{vec}(\mathbf{Y}^\intercal \mathbf{U} \mathbf{X}),$$

where $\mathrm{vec}(\mathbf{U})$ is the *vectorization* of $\mathbf{U}$, i.e., a column vector made of the columns of $\mathbf{U}$ stacked atop one another from left to right (see, e.g., [Dem97]). In this way, only the storage of $\mathcal{O}(d)$ one-dimensional CORSING matrices (of dimension $m_1 \times N_1$) is needed.

4.2.4 Well posedness of full-PG $\mathcal{QS}$ for the 2D Poisson problem

It is possible to show that the full-PG $\mathcal{QS}$ formulation fulfills the classical inf-sup property in the case of the Poisson equation with homogeneous boundary conditions. In particular, we provide a generalization of Proposition 2.5, based on a similar linear algebra argument.

Lemma 4.3. *Let $\mathbf{A} \in \mathbb{R}^{M \times N}, \mathbf{K}_\psi \in \mathbb{R}^{N \times N}$ and $\mathbf{K}_\varphi \in \mathbb{R}^{M \times M}$. If there exist two invertible matrices $\mathbf{M}_\psi \in \mathbb{R}^{N \times N}, \mathbf{M}_\varphi \in \mathbb{R}^{M \times M}$ such that*

$$\mathbf{K}_\psi = \mathbf{M}_\psi^\intercal \mathbf{M}_\psi, \quad \mathbf{K}_\varphi = \mathbf{M}_\varphi^\intercal \mathbf{M}_\varphi$$

then, the following equivalence holds

$$\inf_{\mathbf{u}\in\mathbb{R}^N}\sup_{\mathbf{v}\in\mathbb{R}^M}\frac{\mathbf{v}^\intercal\mathbf{A}\mathbf{u}}{(\mathbf{v}^\intercal\mathbf{K}_\varphi\mathbf{v})^{\frac{1}{2}}(\mathbf{u}^\intercal\mathbf{K}_\psi\mathbf{u})^{\frac{1}{2}}}=\sigma_{\min}(\mathbf{M}_\varphi^{-\intercal}\mathbf{A}\mathbf{M}_\psi^{-1}).$$

Proof. Consider the substitution $\widetilde{\mathbf{v}}=\mathbf{M}_\varphi\mathbf{v}$ and $\widetilde{\mathbf{u}}=\mathbf{M}_\psi\mathbf{u}$. Then, plugging $\mathbf{v}=\mathbf{M}_\varphi^{-1}\widetilde{\mathbf{v}}$ and $\mathbf{u}=\mathbf{M}_\psi^{-1}\widetilde{\mathbf{u}}$ into the inf-sup expression, one obtains

$$\inf_{\mathbf{u}\in\mathbb{R}^N}\sup_{\mathbf{v}\in\mathbb{R}^M}\frac{\mathbf{v}^\intercal\mathbf{A}\mathbf{u}}{(\mathbf{v}^\intercal\mathbf{K}_\varphi\mathbf{v})^{\frac{1}{2}}(\mathbf{u}^\intercal\mathbf{K}_\psi\mathbf{u})^{\frac{1}{2}}}=\inf_{\widetilde{\mathbf{u}}\in\mathbb{R}^N}\sup_{\widetilde{\mathbf{v}}\in\mathbb{R}^M}\frac{\widetilde{\mathbf{v}}^\intercal\mathbf{M}_\varphi^{-\intercal}\mathbf{A}\mathbf{M}_\psi^{-1}\widetilde{\mathbf{u}}}{\|\widetilde{\mathbf{v}}\|_2\|\widetilde{\mathbf{u}}\|_2}=\sigma_{\min}(\mathbf{M}_\varphi^{-\intercal}\mathbf{A}\mathbf{M}_\psi^{-1}).$$

□

Theorem 4.4. *For every $L\in\mathbb{N}$, defined $N_1=2^{L+1}-1$, there exists a constant $\widetilde{\alpha}>0$ such that the following inf-sup condition holds*

$$\inf_{u\in U^N}\sup_{v\in V^N}\frac{(\nabla u,\nabla v)}{|u|_{H^1(\Omega)}|v|_{H^1(\Omega)}}\geq\widetilde{\alpha},\tag{4.15}$$

where $U^N=\operatorname{span}(\mathcal{Q}^L)$ and $V^N=\operatorname{span}(\mathcal{S}^{N_1})$. In particular, $\widetilde{\alpha}\geq 16/\pi^4$.

Proof. We consider the following one-dimensional trial and test functions, without any rescaling with respect to the $H^1(\Omega)$-seminorm and any hierarchical structure on the hat function basis

$$\psi_j(x):=\max(1-|x-x_j^h|/h,0),\quad \varphi_q(x):=\sin(q\pi x),\quad \forall x\in[0,1],\ \forall j,q\in[N_1],$$

with $h=1/(N_1+1)$ and $x_j^h:=jh$, for every $j\in[N_1]$. Moreover, we build the corresponding two-dimensional bases through tensorization, as described in Section 4.2.1, namely

$$\widehat{\psi_{\mathbf{j}}}:=\psi_{j_1}\otimes\psi_{j_2},\quad \widehat{\varphi_{\mathbf{q}}}:=\varphi_{q_1}\otimes\varphi_{q_2},\quad \forall\mathbf{q},\mathbf{j}\in[N_1]^2.$$

Now, we define three matrices playing a key role in the proof. The first one is the well known discrete sine transform matrix $\mathbf{S}\in\mathbb{R}^{N_1\times N_1}$, defined as

$$S_{ij}:=\sin(ij\pi h),\quad \forall i,j\in[N_1].$$

It is symmetric and fulfills the property $\mathbf{S}^2=\frac{1}{2h}\mathbf{I}$. Moreover, we define two tridiagonal matrices $\mathbf{T}_1,\mathbf{T}_2\in\mathbb{R}^{N_1\times N_1}$ as

$$(T_1)_{ij}:=\begin{cases}2/h & \text{if } i=j\\ -1/h & \text{if } |i-j|=1\\ 0 & \text{otherwise,}\end{cases}\quad \forall i,j\in[N_1],$$

and

$$(T_2)_{ij} := \begin{cases} 2h/3 & \text{if } i = j \\ h/6 & \text{if } |i-j| = 1 \\ 0 & \text{otherwise.} \end{cases} \quad \forall i, j \in [N_1].$$

It turns out that they are diagonalized by $\mathbf{S}$. In particular, we have

$$\mathbf{D}_1 = \mathbf{S}\mathbf{T}_1\mathbf{S}, \quad \mathbf{D}_2 = \mathbf{S}\mathbf{T}_2\mathbf{S}, \tag{4.16}$$

where $\mathbf{D}_1, \mathbf{D}_2 \in \mathbb{R}^{N_1 \times N_1}$ are diagonal matrices defined as

$$(D_1)_{ij} = \delta_{ij}\frac{1}{h^2}(1-\cos(\pi j h)), \quad (D_2)_{ij} = \delta_{ij}\frac{1}{6}(2+\cos(\pi j h)), \quad \forall i, j \in [N_1].$$

We also introduce the diagonal matrix $\mathbf{D}_3 \in \mathbb{R}^{N_1 \times N_1}$, defined as

$$(D_3)_{ij} = \delta_{ij}\frac{(\pi j)^2}{2}, \quad \forall i, j \in [N_1].$$

Then, we consider the one-dimensional Poisson equation and define the corresponding full-PG stiffness matrix $\mathbf{L} \in \mathbb{R}^{N_1 \times N_1}$ and the stiffness matrices associated with the one-dimensional trial and test bases $\mathbf{L}_\psi, \mathbf{L}_\varphi \in \mathbb{R}^{N_1 \times N_1}$, respectively, as

$$L_{ij} := (\psi_j', \varphi_i'), \quad (L_\psi)_{ij} := (\psi_j', \psi_i'), \quad (L_\varphi)_{ij} := (\varphi_j', \varphi_i').$$

Analogously, we define the reaction matrices $\mathbf{R}, \mathbf{R}_\psi, \mathbf{R}_\varphi \in \mathbb{R}^{N_1 \times N_1}$ as

$$R_{ij} := (\psi_j, \varphi_i), \quad (R_\psi)_{ij} := (\psi_j, \psi_i), \quad (R_\varphi)_{ij} := (\varphi_j, \varphi_i).$$

Straightforward computations show the following algebraic relations to hold:

$$\begin{aligned} \mathbf{L} &= \mathbf{S}\mathbf{T}_1, & \mathbf{R} &= \tfrac{1}{2}\mathbf{D}_3^{-1}\mathbf{S}\mathbf{T}_1, \\ \mathbf{L}_\psi &= \mathbf{T}_1, & \mathbf{R}_\psi &= \mathbf{T}_2, \\ \mathbf{L}_\varphi &= \mathbf{D}_3, & \mathbf{R}_\varphi &= \tfrac{1}{2}\mathbf{I}. \end{aligned} \tag{4.17}$$

In particular, we explicitly compute $\mathbf{R}$, exploiting that $\varphi_i'' = -(i\pi)^2\varphi_i$, as

$$R_{ij} = (\psi_j, \varphi_i) = -\tfrac{1}{(\pi i)^2}(\psi_j, \varphi_i'') = \tfrac{1}{(\pi i)^2}(\psi_j', \varphi_i') = \tfrac{1}{2}(D_3)_{ii}^{-1}L_{ij}.$$

Hence, $\mathbf{B} = \frac{1}{2}\mathbf{D}_3^{-1}\mathbf{L} = \frac{1}{2}\mathbf{D}_3^{-1}\mathbf{S}\mathbf{T}_1$.

Now, we define the corresponding matrices for the two-dimensional framework, $\widehat{\mathbf{A}}, \widehat{\mathbf{K}}_\psi, \widehat{\mathbf{K}}_\varphi \in \mathbb{R}^{N \times N}$, defined as

$$\widehat{A}_{\mathbf{qj}} := (\nabla\widehat{\psi}_{\mathbf{j}}, \nabla\widehat{\varphi}_{\mathbf{q}}), \quad (\widehat{K}_\psi)_{\mathbf{jk}} := (\nabla\widehat{\psi}_{\mathbf{j}}, \nabla\widehat{\psi}_{\mathbf{k}}), \quad (\widehat{K}_\varphi)_{\mathbf{ql}} := (\nabla\widehat{\varphi}_{\mathbf{q}}, \nabla\widehat{\varphi}_{\mathbf{l}}),$$

where the multi-indices $\mathbf{j}, \mathbf{k}, \mathbf{q}, \mathbf{l}$ are implicitly ordered lexicographically.

Using this notation, (4.15) is equivalent to its discrete counterpart

$$\inf_{\mathbf{u}\in\mathbb{R}^N} \sup_{\mathbf{v}\in\mathbb{R}^N} \frac{\mathbf{v}^\intercal \widehat{\mathbf{A}}\mathbf{u}}{(\mathbf{v}^\intercal \widehat{\mathbf{K}}_\varphi \mathbf{v})^{\frac{1}{2}} (\mathbf{u}^\intercal \widehat{\mathbf{K}}_\psi \mathbf{u})^{\frac{1}{2}}} \geq \widetilde{\alpha}. \tag{4.18}$$

Thanks to the tensorial nature of the two-dimensional bases, we have the following algebraic relations

$$\begin{aligned} \widehat{\mathbf{A}} &= \mathbf{L}\otimes\mathbf{R} + \mathbf{R}\otimes\mathbf{L}, \\ \widehat{\mathbf{K}}_* &= \mathbf{L}_*\otimes\mathbf{R}_* + \mathbf{R}_*\otimes\mathbf{L}_*, \quad * = \psi, \varphi. \end{aligned}$$

Hence, exploiting the algebraic relations (4.17) and the following property of the Kronecker product

$$(\mathbf{X}\otimes\mathbf{Y})(\mathbf{Z}\otimes\mathbf{W}) = \mathbf{XZ}\otimes\mathbf{YW}, \quad \forall \mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \in \mathbb{R}^{N_1\times N_1},$$

we obtain

$$\begin{aligned} \widehat{\mathbf{A}} &= \mathbf{L}\otimes\mathbf{R} + \mathbf{R}\otimes\mathbf{L} \\ &= \mathbf{S}\mathbf{T}_1 \otimes \tfrac{1}{2}\mathbf{D}_3^{-1}\mathbf{S}\mathbf{T}_1 + \tfrac{1}{2}\mathbf{D}_3^{-1}\mathbf{S}\mathbf{T}_1 \otimes \mathbf{S}\mathbf{T}_1 \\ &= \tfrac{1}{2}(\mathbf{I}\otimes\mathbf{D}_3^{-1} + \mathbf{D}_3^{-1}\otimes\mathbf{I})(\mathbf{S}\otimes\mathbf{S})(\mathbf{T}_1\otimes\mathbf{T}_1). \end{aligned}$$

Moreover, using (4.16), we have

$$\begin{aligned} \widehat{\mathbf{K}}_\psi &= \mathbf{L}_\psi\otimes\mathbf{R}_\psi + \mathbf{R}_\psi\otimes\mathbf{L}_\psi \\ &= \mathbf{T}_1\otimes\mathbf{T}_2 + \mathbf{T}_2\otimes\mathbf{T}_1 \\ &= \mathbf{S}^{-1}\mathbf{D}_1\mathbf{S}^{-1}\otimes\mathbf{S}^{-1}\mathbf{D}_2\mathbf{S}^{-1} + \mathbf{S}^{-1}\mathbf{D}_2\mathbf{S}^{-1}\otimes\mathbf{S}^{-1}\mathbf{D}_1\mathbf{S}^{-1} \\ &= 16h^4(\mathbf{S}\mathbf{D}_1\mathbf{S}\otimes\mathbf{S}\mathbf{D}_2\mathbf{S} + \mathbf{S}\mathbf{D}_2\mathbf{S}\otimes\mathbf{S}\mathbf{D}_1\mathbf{S}) \\ &= 16h^4(\mathbf{S}\otimes\mathbf{S})(\mathbf{D}_1\otimes\mathbf{D}_2 + \mathbf{D}_2\otimes\mathbf{D}_1)(\mathbf{S}\otimes\mathbf{S}), \end{aligned}$$

and, finally, with similar arguments, it holds that

$$\widehat{\mathbf{K}}_\varphi = \mathbf{L}_\varphi\otimes\mathbf{R}_\varphi + \mathbf{R}_\varphi\otimes\mathbf{L}_\varphi = \tfrac{1}{2}(\mathbf{D}_3\otimes\mathbf{I} + \mathbf{I}\otimes\mathbf{D}_3).$$

Now, defining the $N\times N$ matrices

$$\widehat{\mathbf{T}}_1 := \mathbf{T}_1\otimes\mathbf{T}_1, \quad \widehat{\mathbf{S}} := \mathbf{S}\otimes\mathbf{S}, \quad \widehat{\mathbf{D}}_1 := \mathbf{D}_1\otimes\mathbf{D}_1,$$

we obtain the relation

$$\widehat{\mathbf{S}}\widehat{\mathbf{T}}_1\widehat{\mathbf{S}} = (\mathbf{S}\otimes\mathbf{S})(\mathbf{T}_1\otimes\mathbf{T}_1)(\mathbf{S}\otimes\mathbf{S}) = \mathbf{S}\mathbf{T}_1\mathbf{S}\otimes\mathbf{S}\mathbf{T}_1\mathbf{S} = \widehat{\mathbf{D}}_1, \tag{4.19}$$

and defining the $N \times N$ diagonal matrices

$$\begin{aligned}\widehat{\mathbf{D}}_2 &:= \tfrac{1}{2}(\mathbf{I} \otimes \mathbf{D}_3^{-1} + \mathbf{D}_3^{-1} \otimes \mathbf{I}),\\ \widehat{\mathbf{D}}_\psi &:= 16h^4(\mathbf{D}_1 \otimes \mathbf{D}_2 + \mathbf{D}_2 \otimes \mathbf{D}_1),\\ \widehat{\mathbf{D}}_\varphi &:= \tfrac{1}{2}(\mathbf{D}_3 \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{D}_3),\end{aligned}$$

yields the following identities

$$\widehat{\mathbf{A}} = \widehat{\mathbf{D}}_2\widehat{\mathbf{S}}\,\widehat{\mathbf{T}}_1, \quad \widehat{\mathbf{K}}_\psi = \widehat{\mathbf{S}}\,\widehat{\mathbf{D}}_\psi\widehat{\mathbf{S}}, \quad \widehat{\mathbf{K}}_\varphi = \widehat{\mathbf{D}}_\varphi.$$

As a consequence, (4.18) becomes

$$\inf_{\mathbf{u}\in\mathbb{R}^N} \sup_{\mathbf{v}\in\mathbb{R}^N} \frac{\mathbf{v}^\intercal\widehat{\mathbf{D}}_2\widehat{\mathbf{S}}\widehat{\mathbf{T}}_1\mathbf{u}}{(\mathbf{v}^\intercal\widehat{\mathbf{D}}_\varphi\mathbf{v})^{\frac{1}{2}}(\mathbf{u}^\intercal\widehat{\mathbf{S}}\,\widehat{\mathbf{D}}_\psi\widehat{\mathbf{S}}\mathbf{u})^{\frac{1}{2}}} \geq \widetilde{\alpha}. \tag{4.20}$$

Applying Lemma 4.3 with the factorizations

$$\widehat{\mathbf{K}}_\psi = (\widehat{\mathbf{D}}_\psi^{\frac{1}{2}}\widehat{\mathbf{S}})^\intercal\widehat{\mathbf{D}}_\psi^{\frac{1}{2}}\widehat{\mathbf{S}}, \quad \widehat{\mathbf{K}}_\varphi = (\widehat{\mathbf{D}}_\varphi^{\frac{1}{2}})^\intercal\widehat{\mathbf{D}}_\varphi^{\frac{1}{2}},$$

recalling relation (4.19) and using equality $\widehat{\mathbf{S}}^{-1} = 4h^2\widehat{\mathbf{S}}$, an equivalent formulation of the inf-sup property (4.20) is

$$\sigma_{\min}(\widehat{\mathbf{G}}) \geq \widetilde{\alpha},$$

with $\widehat{\mathbf{G}} := 4h^2\widehat{\mathbf{D}}_\varphi^{-\frac{1}{2}}\widehat{\mathbf{D}}_2\widehat{\mathbf{D}}_1\widehat{\mathbf{D}}_\psi^{-\frac{1}{2}}$. Since $\widehat{\mathbf{G}}$ is a diagonal matrix, its minimum singular value coincides with its smallest absolute diagonal entry, that we explicitly estimate from below in the following.

Being $\widehat{\mathbf{G}}$ product of diagonal matrices, we have

$$\widehat{G}_{k,k} = 4h^2\frac{(\widehat{D}_2)_{k,k}(\widehat{D}_1)_{k,k}}{\sqrt{(\widehat{D}_\varphi)_{k,k}(\widehat{D}_\psi)_{k,k}}}, \quad \forall k \in [N].$$

Then, defining the functions

$$d_1(x) := 1 - \cos(\pi x), \quad d_2(x) := \frac{1}{6}(2 + \cos(\pi x)), \quad d_3(x) := \frac{\pi^2 x^2}{2}, \quad \forall x \in [0,1],$$

and

$$\mathcal{G}(x_1,x_2) := \frac{1}{\sqrt{2}}\frac{d_1(x_1)d_1(x_2)\sqrt{d_3(x_1)+d_3(x_2)}}{d_3(x_1)d_3(x_2)\sqrt{d_1(x_1)d_2(x_2)+d_2(x_1)d_1(x_2)}}, \tag{4.21}$$

and exploiting the relation

$$(\widetilde{D}\otimes\widetilde{E})_{N_1(i-1)+j,N_1(i-1)+j} = \widetilde{D}_{ii}\widetilde{E}_{jj},$$

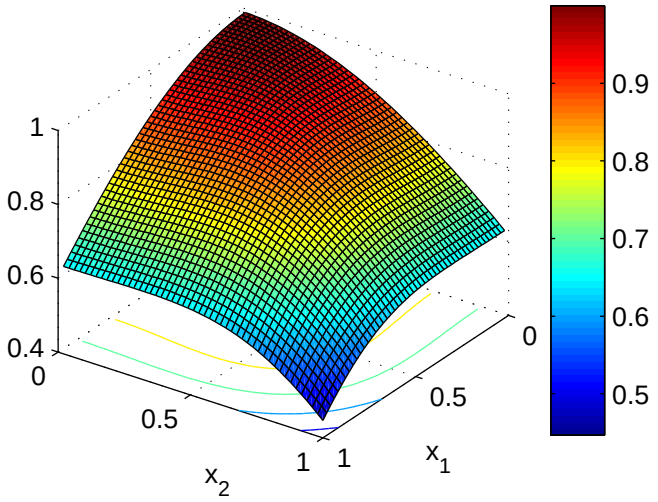


Figure 4.9: Surface plot of the function $\mathcal{G}$ defined in (4.21).

for all diagonal matrices $\widetilde{\mathbf{D}}, \widetilde{\mathbf{E}} \in \mathbb{R}^{N_1 \times N_1}$, it can be easily checked that

$$\widehat{G}_{N_1(i-1)+j, N_1(i-1)+j} = \mathcal{G}(x_i^h, x_j^h), \quad \forall i, j \in [N_1].$$

The function $\mathcal{G}(x_1, x_2)$ does not depend on h and its surface plot is reported in Figure 4.9. Hence,

$$\min_{k \in [N]} \widehat{G}_{k,k} \geq \min_{(x,y) \in [0,1]^2} \mathcal{G}(x_1, x_2).$$

The final step is to estimate the minimum of $\mathcal{G}$ over $[0,1]^2$ from below. By noticing that

$$\frac{1}{6} \leq d_2(x) \leq \frac{1}{2}, \quad \frac{4}{\pi^2} \leq \frac{d_1(x)}{d_3(x)} \leq 1, \quad \forall x \in [0,1],$$

where the second inequality holds by continuity at the singular point $x = 0$, we obtain

$$\mathcal{G}(x_1, x_2) \geq \frac{d_1(x_1)}{d_3(x_1)} \frac{d_1(x_2)}{d_3(x_2)} \sqrt{\frac{d_3(x_1) + d_3(x_2)}{d_1(x_1) + d_1(x_2)}} \geq \frac{4}{\pi^2} \cdot \frac{4}{\pi^2} \cdot 1 = \frac{16}{\pi^4},$$

which implies the thesis. □

Remark 4.2.3. The optimal value of the constant $\widetilde{\alpha}$ is actually $8\sqrt{3}/\pi^3 \approx 0.45$. Indeed, looking at the plot in Figure 4.9 of the function $\mathcal{G}$, we notice that its minimum is reached at $(1,1)$. Hence, $\widetilde{\alpha} \geq \mathcal{G}(1,1) = 8\sqrt{3}/\pi^3$. Even if it is graphically evident where this minimum is reached $\mathcal{G}(1,1)$, proving it in a rigorous

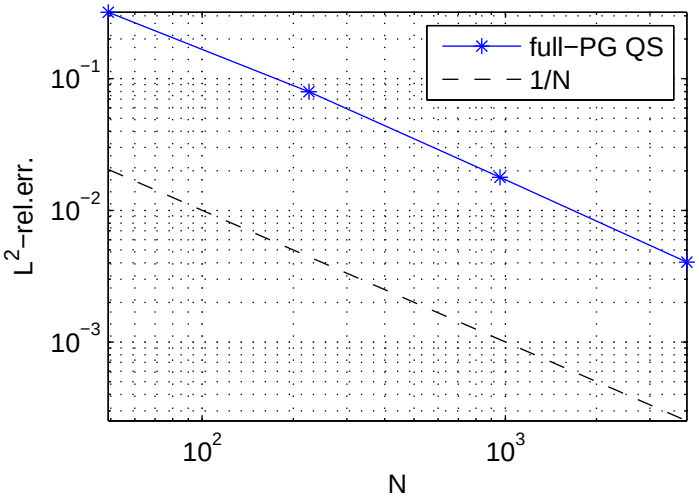


Figure 4.10: Convergence trend of the full-PG $\mathcal{QS}$ method in the two-dimensional case.

way is not so trivial (see the MathOverflow question `http://mathoverflow.net/questions/219575`). □

Finally, we observe that, using a similar argument, a statement analogous to Theorem 4.4 holds in the $\mathcal{SQ}$ case.

4.2.5 Numerical results for the 2D case

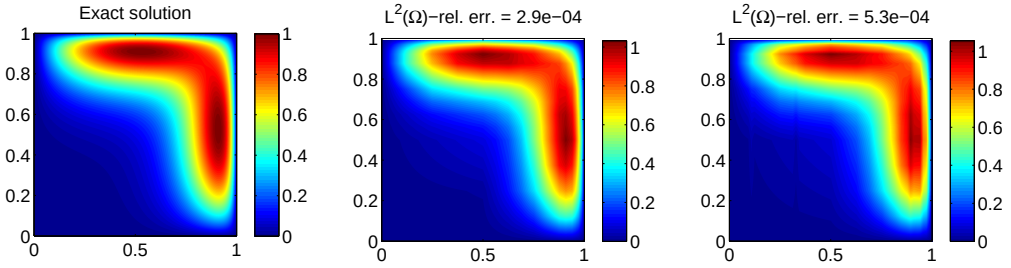
We assess the performances of the CORSING $\mathcal{QS}$ method with the tensorized test selection strategy described in Section 4.2.3, applied to the advection-dominated example (2.38), with $\eta = 0.1$, $\mathbf{b} = [1,1]^\intercal$ and with exact solution defined as in (2.39).

Convergence of full-PG $\mathcal{QS}$ First, we check that the full-PG $\mathcal{QS}$ method reaches the best approximation error with respect to the $L^2(\Omega)$-norm in Figure 4.10. The figure shows the $L^2(\Omega)$-norm of the relative error associated with the full-PG $\mathcal{QS}$ solution for $N_1 = M_1 = 2^{L+1} - 1$ and $L = 2,3,4,5$, corresponding to $N_1 = 7,15,31,63$ and trial space dimension $N = 49,225,961,3969$, respectively. The error decays according to the asymptotical trend $1/N$.

To show that this is indeed the desired trend, we recall a standard estimate for the interpolation error on a uniform tensor product grid of mesh-size h (see, e.g., [BS08, Theorem 4.6.11]), given by

$$\|w - \Pi_h^1 w\|_{L^2(\Omega)} \lesssim h^2 |w|_{H^2(\Omega)}, \quad \forall w \in H^2(\Omega), \tag{4.22}$$

where Π_h^1 is the Lagrangian interpolant of order 1. Thanks to the regularity of the exact solution, we conclude that full-PG reaches the interpolation error with



(a) (b) (c)

Figure 4.11: Assessment of CORSING $\mathcal{QS}$ on the 2D advection-dominated problem (2.38): exact solution, (a); best solution in the first cluster, (b); worst solution in the first cluster, (c).

respect to the $L^2(\Omega)$-norm. Indeed, the mesh-size is $h = 1/N_1 = 1/\sqrt{N}$ in the full-PG $\mathcal{QS}$ case.

Remark 4.2.4. We also notice that the error associated with the full-PG solution is optimal, thanks to Theorem 2.3 and Theorem 4.4. □

Assessment of CORSING $\mathcal{QS}$ We apply the CORSING $\mathcal{QS}$ method using the structured randomization described in Section 4.2.3, with $L = 5$, corresponding to $N_1 = 63$ and a trial space of dimension $N = 3969$. Then, we fix a sparsity level $s = 50$ and TS = 70%, corresponding to $m_1 = 34$ and $m = 1156$. The resulting statistical analysis over 50 runs yields an ESP = 0.96, and a mean $L^2(\Omega)$-norm of the relative error in the first cluster of $3.3 \cdot 10^{-4}$, with corresponding standard deviation $6.4 \cdot 10^{-5}$. In Figure 4.11, the exact solution (a) is compared with the best and the worst solutions in the first cluster ((b) and (c), respectively). The two boundary layers are well captured in both cases, but small artifacts appear in the worst case scenario (Figure 4.11, (c)). These are probably due to the tensorized test selection, that makes the random experiments slightly "biased".

4.2.6 Numerical results for the 3D case

In this final numerical assessment, we provide preliminary results for the application of CORSING $\mathcal{QS}$ in dimension $d = 3$. We consider the Poisson equation as a model problem, with exact solution

$$u(\mathbf{x}) := (x_1 - x_1^2)(x_2 - x_2^2)(x_3 - x_3^2), \quad \forall \mathbf{x} \in [0,1]^3. \tag{4.23}$$

Convergence of full-PG $\mathcal{QS}$ First, we check the convergence of the full-PG $\mathcal{QS}$ method showing that it shares the same trend as the interpolation error with respect to the $L^2(\Omega)$-norm (Figure 4.12). In particular, we compute the full-PG $\mathcal{QS}$

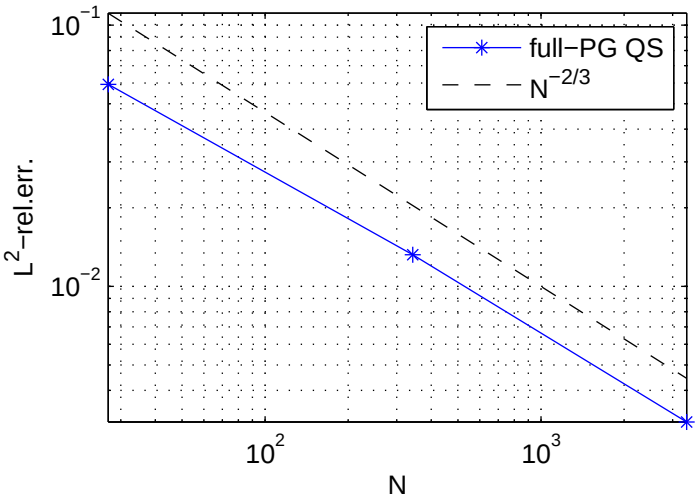


Figure 4.12: Convergence trend of the full-PG $\mathcal{QS}$ approach in the three dimensional case.

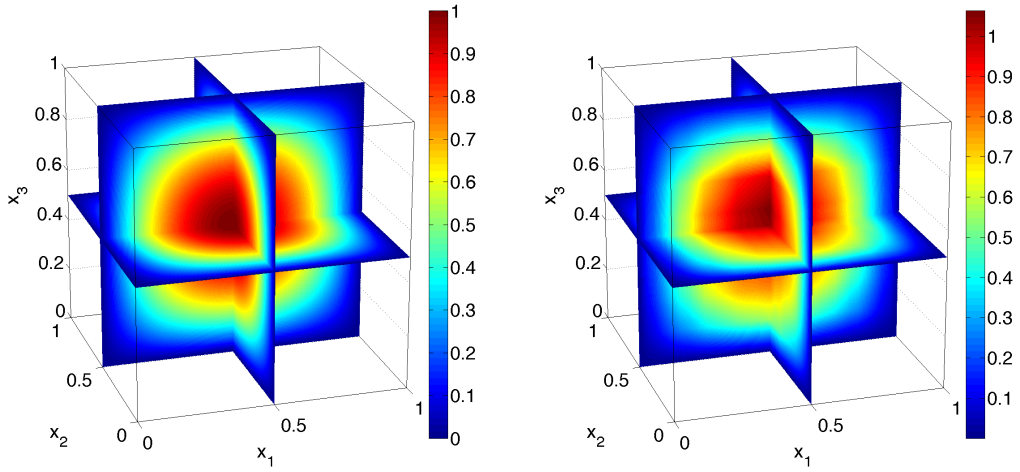


Figure 4.13: Comparison between the exact solution (4.23) of the three-dimensional Poisson problem, (a), and the worst solution in the cluster of the CORSING $\mathcal{QS}$ method, (b).

solution with $N_1 = M_1 = 2^{L+1} - 1$ and $L = 1, 2, 3$, corresponding to $N_1 = 3, 7, 15$ and $N = 27, 343, 3375$. The error asymptotically decays like $N^{-2/3}$. Hence, recalling (4.22) and observing that the mesh-size is $h = N_1^{-1} = N^{-1/3}$, we conclude that the full-PG method reaches the interpolation error in the $L^2(\Omega)$-norm

Assessment of CORSING $\mathcal{QS}$ We apply the CORSING $\mathcal{QS}$ method fixing the hierarchical level $L = 4$, corresponding to $N_1 = 31$ and a total trial space dimension $N = 29791$. Then we choose $m_1 = 16$, leading to a total number of test functions $m = 4096$ and TS = 86% and fix $s = 200$. We carry out 50 random experiments, obtaining ESP = 0.78. The resulting mean $L^2(\Omega)$-relative error is $3.9 \cdot 10^{-2}$, with a standard deviation $1.1 \cdot 10^{-2}$.

We show the worst solution belonging to the successful cluster in Figure 4.13. The main features of the solution are well captured, confirming the applicability of the CORSING strategy to the three-dimensional case.

Conclusions

We have shown how CS can be applied to reduce the computational cost associated with the PG discretization of a PDE.

In particular, we applied CORSING to the one-dimensional ADR problem in Chapter 2, with a suitable choice of the trial and test spaces leading to well posed full-PG formulations (Propositions 2.5 and 2.7). The extensive numerical analysis of the one-dimensional case carried out in Section 2.3 shows that the CORSING procedure is accurate and robust, both in terms of classical error measures, such as the $L^2(\Omega)$-norm and the $H^1(\Omega)$-seminorm, as well as with respect to statistical tools, such as the indices TS and ESP. Moreover, the results of the thorough numerical comparison against full-PG, FE, an SVD-based approach and the best s-term approximation error provide an additional confirmation of the reliability of CORSING.

A comparison between the deterministic (D-CORSING) and randomized (R-CORSING) variant of the method showed that the first one could be affected by instability, leading to aliasing phenomena (Figure 2.6), or modest ESP rates (Figure 2.11), whereas the second one showed to be more robust and reliable.

Afterwards, we generalized the CORSING methodology to the two-dimensional ADR problem (Section 2.4). In particular, in Section 2.4.4 showed the advantages of CORSING with respect to the full-PG approach, both in terms of memory and computational time, on an advection-dominated example.

In Chapter 3, we presented a rigorous formalization and provided a theoretical analysis of the R-CORSING method, based on the local a-coherence and the RISP. In particular, in Theorem 3.11 we showed that a sufficient condition for the RISP to hold with high probability in a given s-sparse set is that m and s be linear dependent, up to logarithmic factors. On the contrary, at the moment we have been only able to prove (Theorem 3.12) a *uniform* RISP (i.e., a RISP holding in *all* possible s-sparse sets) assuming a (suboptimal) quadratic dependence between m and s. Afterwards, an s-sparse RIP result in high probability is proved in Theorem 3.19, assuming an optimal linear dependence between m and s. Exploiting these theorems, we proved a recovery result in expectation (Theorem 3.16) and one in probability (Theorem 3.17).

The hypotheses of this general theory have been explicitly checked in the

case of CORSING $\mathcal{SH}$ and $\mathcal{HS}$ applied to one-dimensional ADR equations with constant coefficients and of CORSING $\mathcal{HS}$ applied to a one-dimensional diffusion equation, with nonconstant diffusivity. Finally, we numerically assessed the theoretical hypotheses for CORSING $\mathcal{PS}$ applied to the two-dimensional Poisson equation.

Considering more challenging benchmarks, we showed that CORSING can be successfully applied to the Stokes problem (Section 4.1) after a suitable stabilization strategy for the full-PG approach. Moreover, we generalized the methodology through tensorization to the multi-dimensional ADR problem in Section 4.2, introducing the CORSING $\mathcal{QS}$ strategy. First, the inf-sup property is proved for the full-PG $\mathcal{QS}$ approach applied to two-dimensional Poisson problem (Theorem 4.4). Then, after the introduction of a tensorized test selection strategy, that shows how to apply CORSING in dimension $d > 2$, a preliminary validation of CORSING $\mathcal{QS}$ on the three-dimensional Poisson equation is provided.

Future developments

We present a series of open challenges related to the extension and the improvement of the CORSING methodology.

Domains with complex geometries The extension to a domain $\Omega \subseteq \mathbb{R}^d$, with $d = 2, 3$, characterized by a complex geometry is a delicate issue. Mimicking the approach followed for $\Omega = (0,1)^d$, we can build, on the one hand, a hierarchical basis over a set of nested triangulations and, on the other hand, a family of global basis functions, playing the role of sine functions. This could be addressed following the approaches proposed in [Osw13] and [CHQZ07], respectively.

ADR with nonconstant coefficients The case of ADR with nonconstant coefficients needs to be analyzed in more detail. In particular, even though the theoretical results for the one-dimensional diffusion equation with nonconstant diffusive term seem promising (see Section 3.3.5), a numerical validation of CORSING is still lacking. Probably, numerical integration for the assembly of the stiffness matrix could not be avoided in this case.

Stabilization in the advection-dominated case In the advection-dominated case, the CORSING procedure could benefit of the use of stabilization techniques (see,e.g., [BDG06, CDW12] and the references therein).

Exploiting the hidden structure of A The stiffness matrix $\mathbf{A}$ associated with the CORSING discretization exhibits a remarkable structure in the $\mathcal{SH}$, $\mathcal{HS}$, $\mathcal{SQ}$ and $\mathcal{QS}$ cases (e.g., recall the decomposition $\mathbf{A} = \mathbf{TS}$ in (2.24) and Lemma 4.2). This structure could be exploited in order to optimize the linear algebra operations involved in the OMP algorithm, such as the matrix-vector multiplications associated with $\mathbf{A}$ and $\mathbf{A}^\intercal$. Moreover, an interesting related direction could be to apply efficient recovery algorithm such as that proposed in [GHI+13] in the case of Fourier measurements, that employs $\mathcal{O}(s \log N)$ samples and runs in $\mathcal{O}(s \log^2 N)$ time. A computational cost for the recovery phase that scales sublinearly in N

could make the CORSING method highly competitive in the context of numerical methods for PDEs.

Application of the theory in the multi-dimensional case The hypotheses of the theory presented in Section 3.2 have been numerically checked in the case of CORSING $\mathcal{PS}$ and $\mathcal{QS}$, but a rigorous theoretical verification is still open.

Nonorthonormal test functions Numerical evidence shows that CORSING can be performed also when the test functions are not orthonormal (consider, e.g., the CORSING $\mathcal{SP}$ approach). Though, at this stage, we cannot provide any theoretical justification for this.

Dictionaries *vs* Bases Another possible extension is the use of *dictionaries* instead of bases, i.e., families of possibly linearly dependent functions. This is particularly meaningful in the case of trial functions, since the redundancy of the dictionary generates very sparse representations of the solution. Moreover, this could improve the stability of the hierarchical multiscale decomposition (see [Osw13]).

More challenging benchmarks Finally, we would like to apply the CORSING strategy to more challenging settings, such as *nonlocal* problems or *boundary integral equations* [Bon99], whose discretization could give rise to full stiffness matrices, or nonlinear PDEs, such as the well known *Navier-Stokes* equations.

Acknowledgements

I deeply thank my advisors Simona Perotto and Stefano Micheletti for having introduced me to the world of research with incredible enthusiasm and passion. They trusted me, giving me freedom to follow my creativity and time to develop my ideas. At the same time, they guided me in crucial moments, when important decisions had to be made. They taught me how to look at old problems with new eyes - how to think "out of the box". I consider them not only talented researchers, but also special persons, capable of great humanity and humility.

I thank Fabio Nobile, for having hosted me at École polytechnique fédérale de Lausanne in fall 2014. I learnt from him that tough problems must be attacked without fear, always looking for insightful and simple explanations to complex phenomena. My passion for research increased exponentially during my stay in Lausanne.

I gratefully acknowledge Holger Rauhut, Wolfgang Dahmen and the members of their research groups for the great hospitality received in RWTH Aachen University during my short but very fruitful visit in September 2015.

Finally, I express my gratitude to the Reviewers for their constructive comments and suggestions.

List of acronyms

AFEM	Adaptive Finite Element Method
BOS	Bounded Orthonormal System
CORSING	COmpRessed SolvING
CS	Compressed Sensing
ESP	Empirical Success Probability
FE	Finite Elements
full-PG	Full Petrov-Galerkin
$\mathcal{HS}$	Hat functions *vs.* Sine functions
OMP	Orthogonal Matching Pursuit
PDE	Partial Differential Equation
PG	Petrov–Galerkin
$\mathcal{PS}$	Pyramids *vs.* Sine functions
$\mathcal{QS}$	$\mathbb{Q}_1$ polynomials *vs.* Sine functions
RIP	Restricted Isometry Property
RISP	Restricted Inf-Sup Property
$\mathcal{SH}$	Sine functions *vs.* Hat functions
$\mathcal{SP}$	Sine functions *vs.* Pyramids
$\mathcal{SQ}$	Sine functions *vs.* $\mathbb{Q}_1$ polynomials
TS	Test Savings

Bibliography

[AB72] A.K. Aziz and I. Babuška. *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, chapter 1. Academic Press, New York, 1972.

[AH15] B. Adcock and A.C. Hansen. Generalized Sampling and Infinite-Dimensional Compressed Sensing. *Found. Comput. Math.*, pages 1–61, 2015.

[AHPR13] B. Adcock, A.C. Hansen, C. Poon, and B. Roman. Breaking the coherence barrier: asymptotic incoherence and asymptotic sparsity in compressed sensing. *arXiv:1302.0561*, 2013.

[AW02] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569–579, 2002.

[BBF13] D. Boffi, F. Brezzi, and M. Fortin. *Mixed Finite Element Methods and Applications*, volume 44 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2013.

[BBRS15] J.-L. Bouchot, B. Bykowski, H. Rauhut, and C. Schwab. Compressed sensing petrov-galerkin approximations for parametric pdes. Technical Report 2015-09, Seminar for Applied Mathematics, ETH, Zürich, 2015.

[BDG06] P.B. Bochev, C.R. Dohrmann, and M.D. Gunzburger. Stabilization of low-order mixed finite elements for the Stokes equations. *SIAM J. Numer. Anal.*, 44(1):82–101, 2006.

[BF91] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.

[BG04] H.J. Bungartz and M. Griebel. Sparse grids. *Acta Numer.*, 13:147–269, 2004.

[BMP15] S. Brugiapaglia, S. Micheletti, and S. Perotto. Compressed solving: A numerical approximation technique for elliptic PDEs based on Compressed Sensing. *Comput. Math. Appl.*, 70(6):1306–1335, 2015.

[BNMP15] S. Brugiapaglia, F. Nobile, S. Micheletti, and S. Perotto. A theoretical study of COmpRessed SolvING for advection-diffusion-reaction problems. Technical Report MOX-Report No. 42/2015, Politecnico di Milano, Dip. di Matematica, 2015.

[Bon99] M. Bonnet. Boundary integral equation methods for solids and fluids. *Meccanica*, 34(4):301–302, 1999.

[BS08] S.C. Brenner and R. Scott. *The mathematical theory of finite element methods*, volume 15. Springer Science & Business Media, 2008.

[Bub13] I.G. Bubnov. Report on the works of professor Timoshenko which were awarded the Zhuranskyi Prize. In *Symposium of the institute of communication engineers*, volume 81, 1913.

[CBL89] S. Chen, S. A. Billings, and W. Luo. Orthogonal least squares methods and their application to non-linear system identification. *Internat. J. Control*, 50(5):1873–1896, 1989.

[CCM+15] A. Chkifa, A. Cohen, G. Migliorati, F. Nobile, and R. Tempone. Discrete least squares polynomial approximation with random evaluations - application to parametric and stochastic elliptic PDEs. *ESAIM: M2AN*, 49(3):815–837, 2015.

[CDD15] A. Cohen, W. Dahmen, and R. DeVore. Orthogonal matching pursuit under the restricted isometry property. *arXiv preprint arXiv:1506.04779*, 2015.

[CDL13] A. Cohen, M.A. Davenport, and D. Leviatan. On the stability and accuracy of least squares approximations. *Found. Comput. Math.*, 13(5):819–834, 2013.

[CDW12] A. Cohen, W. Dahmen, and G. Welper. Adaptivity and variational stabilization for convection-diffusion equations. *ESAIM Math. Model. Numer. Anal.*, 46(05):1247–1273, 2012.

[CGM01] R. Coifman, F. Geshwind, and Y. Meyer. Noiselets. *Appl. Comput. Harmon. Anal.*, 10(1):27–44, 2001.

[Che52] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sums of observations. *Ann. Math. Stat.*, 23:409–507, 1952.

[Chl87] E.F.F. Chladni. *Entdeckungen über die Theorie des Klanges*. Zentralantiquariat der DDR, 1787.

[CHQZ07] C.G. Canuto, M.Y. Hussaini, A.M. Quarteroni, and T.A. Zang. *Spectral Methods: Evolution to Complex Geometries and Applications to Fluid Dynamics (Scientific Computation)*. Springer-Verlag New York, Inc., 2007.

[Chr02] O. Christensen. *An Introduction to Frames and Riesz Bases*. Appl. Numer. Harmon. Anal. Birkhäuser Boston, 2002.

[CQ82] C. Canuto and A. Quarteroni. Approximation results for orthogonal polynomials in Sobolev spaces. *Math. Comp.*, 38(157):67–86, 1982.

[CR07] E.J. Candès and J.K. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23:969–985, 2007.

[CRT06] E.J. Candès, J.K. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[CT65] J.W. Cooley and J.W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19(90):297–301, 1965.

[CW08] E.J. Candès and M.B. Wakin. An introduction to compressive sampling. *IEEE Signal Process. Mag.*, 25(2):21–30, 2008.

[Dah97] W. Dahmen. Wavelet and multiscale methods for operator equations. *Acta Numer.*, 6:55–228, 1997.

[Dau92] I. Daubechies. *Ten Lectures on Wavelets*, volume 61. SIAM, Philadelphia, 1992.

[DDT+08] M.F. Duarte, M.A. Davenport, D. Takhar, J.N. Laska, T. Sun, K. E Kelly, and R.G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Process. Magazine*, 25(2):83, 2008.

[Dem97] J.W. Demmel. *Applied numerical linear algebra*. Siam, 1997.

[DeV98] R.A. DeVore. Nonlinear approximation. *Acta Numer.*, 7:51–150, 1998.

[DHSW12] W. Dahmen, C. Huang, C. Schwab, and G. Welper. Adaptive Petrov–Galerkin methods for first order transport equations. *SIAM J. Num. Anal.*, 50(5):2420–2445, 2012.

[DL92] D.L. Donoho and B.F. Logan. Signal recovery and the large sieve. *SIAM J. Appl. Math.*, 52(2):577–591, 1992.

[DO11] A. Doostan and H. Owhadi. A non-adapted sparse approximation of PDEs with stochastic inputs. *J. Comput. Phys.*, 230(8):3015–3034, 2011.

[Don06] D.L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52:1289–1306, 2006.

[DS89] D.L. Donoho and P.B. Stark. Uncertainty principles and signal recovery. *SIAM J. Appl. Math.*, 49(3):906–931, 1989.

[EB02] M. Elad and A.M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. Inform. Theory*, 48(9):2558–2567, 2002.

[EG13] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159. Springer Science+Business Media, 2013.

[Ela10] M. Elad. *Sparse and Redundant Representations: from Theory to Applications in Signal and Image Processing*. Springer Science+Business Media, New York, 1st edition, 2010.

[Fou22] J. Fourier. *Theorie analytique de la chaleur, par M. Fourier*. Chez Firmin Didot, père et fils, 1822.

[FR11] M. Fornasier and H. Rauhut. Compressive Sensing. In O. Scherzer, editor, *Handbook of Mathematical Methods in Imaging*, pages 187–228. Springer Science+Business Media, New York, 2011.

[FR13] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Appl. Numer. Harmon. Anal. Springer Science+Business Media, New York, 2013.

[FS81] J.H. Friedman and W. Stuetzle. Projection pursuit regression. *J. Amer. Statist. Assoc.*, 76(376):817–823, 1981.

[Gal15] B.G. Galerkin. Series solution of some problems of elastic equilibrium of rods and plates. *Vestnik inzhenerov i tekhnikov*, 19(7):897–908, 1915.

[GHI+13] B. Ghazi, H. Hassanieh, P. Indyk, D. Katabi, E. Price, and L. Shi. Sample-optimal average-case Sparse Fourier Transform in two dimensions. In *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, pages 1258–1265. IEEE, 2013.

[GKL06] K. Guo, G. Kutyniok, and D. Labate. Sparse multidimensional representations using anisotropic dilation and shear operators. *Wavelets und Splines (Athens, GA, 2005), G. Chen und MJ Lai, eds., Nashboro Press, Nashville, TN*, pages 189–201, 2006.

[GL13] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, 4th edition, 2013.

[GMS03] A.C. Gilbert, S. Muthukrishnan, and M.J. Strauss. Approximation of functions over redundant dictionaries using coherence. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 243–252. Society for Industrial and Applied Mathematics, 2003.

[GP09] J.-L. Guermond and B. Popov. An optimal L^1-minimization algorithm for stationary Hamilton-Jacobi equations. *Commun. Math. Sci.*, 7(1):211–238, 2009.

[Gra10] L. Grasedyck. Hierarchical singular value decomposition of tensors. *SIAM J. Matrix Anal. Appl.*, 31(4):2029–2054, 2010.

[Gue04] J.-L. Guermond. A finite element technique for solving first-order PDEs in L^P. *SIAM J. Numer. Anal.*, 42(2):714–737 (electronic), 2004.

[GW12] M.J. Gander and G. Wanner. From Euler, Ritz, and Galerkin to Modern Computing. *SIAM Rev.*, 54(4):627–666, 2012.

[Haa10] A. Haar. Zur theorie der orthogonalen funktionensysteme. *Math. Ann.*, 69(3):331–371, 1910.

[HL67] R.R. Hocking and R.N. Leslie. Selection of the best subset in regression analysis. *Technometrics*, 9(4):531–540, 1967.

[HS09] M. Herman and T. Strohmer. High-resolution radar via compressed sensing. *IEEE Trans. Signal Process.*, 57(6):2275–2284, 2009.

[JMPY10] S. Jokar, V. Mehrmann, M.E. Pfetsch, and H. Yserentant. Sparse approximate solution of partial differential equations. *Appl. Numer. Math.*, 60:452–472, 2010.

[JV11] L. Jacques and P. Vandergheynst. Compressed Sensing: "When Sparsity Meets Sampling". In G. Cristobal, P. Schelkens, and H. Thienpont, editors, *Optical and Digital Image Processing - Fundamentals and Applications*. Wiley-VCH, Weinheim, Germany, 2011.

[Kut12] G. Kutyniok. Compressed sensing: Theory and applications. *arXiv:1203.3815*, 2012.

[KV02] K. Kunisch and S. Volkwein. Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. *SIAM J. Numer. Anal.*, 40(2):492–515, 2002.

[KW14] F. Krahmer and R. Ward. Stable and robust sampling strategies for compressive imaging. *IEEE Trans. Image Process.*, 23(2):612–622, 2014.

[Lav88] J.E. Lavery. Nonoscillatory solution of the steady-state inviscid Burgers' equation by mathematical programming. *J. Comput. Phys.*, 79(2):436–448, 1988.

[Lav89] J.E. Lavery. Solution of steady-state one-dimensional conservation laws by mathematical programming. *SIAM J. Numer. Anal.*, 26(5):1081–1089, 1989.

[LDSP08] M. Lustig, D.L. Donoho, J.M. Santos, and J.M. Pauly. Compressed sensing mri. *IEEE Signal Process. Magazine*, 25(2):72–82, 2008.

[LH70] L.R. LaMotte and R.R. Hocking. Computational efficiency in the selection of regression variables. *Technometrics*, 12(1):83–93, 1970.

[Li09] J. Li. Penalty finite element approximations for the Stokes equations by l2 projection. *Math. Methods Appl. Sci.*, 32(4):470–479, 2009.

[Liv12] E.D. Livshitz. On the optimality of the orthogonal greedy algorithm for μ-coherent dictionaries. *J. Approx. Theory*, 164(5):668–681, 2012.

[LM72] J.L. Lions and E. Magenes. *Non-Homogeneous Boundary Value Problems and Applications*, volume I. Springer-Verlag, Berlin, 1972.

[Log65] B.F. Logan. *Properties of high-pass signals*. PhD thesis, Columbia university., 1965.

[Mal99] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic press, 1999.

[MZ93] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12):3397–3415, 1993.

[Nat95] B.K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.

[Neč62] J. Nečas. Sur une méthode pour résoudre les équations aux dérivées partielles du type elliptique, voisine de la variationnelle. *Ann. Sc. Norm. Super. Pisa Cl. Sci.*, 16(4):305–326, 1962.

[NSV09] R.H. Nochetto, K.G. Siebert, and A. Veeser. Theory of adaptive finite element methods: an introduction. In *Multiscale, nonlinear and adaptive approximation*, pages 409–542. Springer, 2009.

[Nyq28] H. Nyquist. Certain topics in telegraph transmission theory. *Trans. Amer. Inst. Electr. Eng.*, 47(2):617–644, 1928.

[Ose10] I.V. Oseledets. Approximation of $2^d \times 2^d$ matrices using tensor decomposition. *SIAM J. Matrix Anal. Appl.*, 31(4):2130–2145, 2010.

[Osw13] P. Oswald. *Multilevel finite element approximation: theory and applications*. Springer-Verlag, 2013.

[Pet40] G.I. Petrov. Application of Galerkin's method to a problem of the stability of the flow of a viscous fluid. *Priklad. Matem. Mekh.*, 4:3–12, 1940.

[PHD14] J. Peng, J. Hampton, and A. Doostan. A weighted ℓ_1-minimization approach for sparse polynomial chaos expansions. *J. Comput. Phys.*, 267:92–111, 2014.

[PM93] W.B. Pennebaker and J.L. Mitchell. *JPEG: Still image data compression standard*. Springer Science & Business Media, 1993.

[PRK93] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the 27th Annual Asilomar Conference on Signals, Systems, and Computers*, pages 40–44, 1993.

[Qua14] A. Quarteroni. *Numerical Models for Differential Problems*, volume 8 of *MS&A*. Springer-Verlag Italia, Milan, 2nd edition, 2014.

[QV08] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*, volume 23 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2008.

[Rau10] H. Rauhut. Compressive sensing and structured random matrices. In M. Fornasier, editor, *Theoretical Foundations and Numerical Methods for Sparse Recovery*, volume 9 of *Radon Series on Computational and Applied Mathematics*, pages 1–92. deGruyter, 2010.

[Ric] Rice University. Compressive Sensing Resources. `http://dsp.rice.edu/cs`.

[Rit08] W. Ritz. Über eine neue methode zur lösung gewisser variationsprobleme der mathematischen physik. *J. Reine Angew. Math.*, 135:1–61, 1908.

[Rit09] Walter Ritz. Theorie der transversalschwingungen einer quadratischen platte mit freien rändern. *Ann. Phys.*, 333(4):737–786, 1909.

[RS14] H. Rauhut and C. Schwab. Compressive Sensing Petrov-Galerkin approximation of high-dimensional parametric operator equations. *arXiv preprint arXiv:1410.4929*, 2014.

[Rub09] R. Rubinstein. Omp-Box v10. `http://www.cs.technion.ac.il/~ronrubin/software.html`, 2009.

[RZE08] R. Rubinstein, M. Zibulevsky, and M. Elad. Efficient implementation of the k-SVD algorithm using batch orthogonal matching pursuit. Technical Report CS-2008-08, Technion, Computer Science Department, 2008.

[SCD02] J.-L. Starck, E.J. Candès, and D.L. Donoho. The curvelet transform for image denoising. *IEEE Trans. Image Process.*, 11(6):670–684, 2002.

[Sha49] C.E. Shannon. Communication in the presence of noise. *Proc. IRE*, 37(1):10–21, 1949.

[SSN+14] K. Sargsyan, C. Safta, H.N. Najm, B.J. Debusschere, D. Ricciuto, and P. Thornton. Dimensionality reduction for complex models via Bayesian Compressive Sensing. *Int. J. Uncertain. Quantif.*, 4(1), 2014.

[Tem01] R. Temam. *Navier-Stokes equations: theory and numerical analysis*, volume 343. AMS, 2001.

[Tem03] V.N. Temlyakov. Nonlinear methods of approximation. *Found. Comput. Math.*, 3(1):33–107, 2003.

[Tim13] S. Timoshenko. Sur la stabilité des systèmes élastiques. *Annales des Ponts et Chaussées*, 9:496–566, 1913.

[Tol12] G.P. Tolstov. *Fourier Series*. Dover Books on Mathematics. Dover Publications Inc., Mineola, N.Y., 2012.

[Tro04] J.A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.

[Tro11] J.A. Tropp. Improved analysis of the subsampled randomized hadamard transform. *Adv. Adapt. Data Anal.*, 3:115–126, 2011.

[Tro12] J.A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[vdBF07] E. van den Berg and M.P. Friedlander. SPGL1: A solver for large-scale sparse reconstruction, 2007. `http://www.cs.ubc.ca/labs/scl/spgl1`.

[vdBF08] E. van den Berg and M.P. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comput.*, 31(2):890–912, 2008.

[Wel74] L. Welch. Lower bounds on the maximum cross correlation of signals (corresp.). *IEEE Trans. Inform. Theory*, pages 397–399, 1974.

[YK13] X. Yang and G.E. Karniadakis. Reweighted ℓ_1-minimization method for stochastic elliptic differential equations. *J. Comput. Phys.*, 248:87–108, 2013.

[Yse86] H. Yserentant. On the multi-level splitting of finite element spaces. *Numer. Math.*, 49(4):379–412, 1986.

[Zha11] T. Zhang. Sparse recovery with orthogonal matching pursuit under rip. *IEEE Trans. Inform. Theory*, 57(9):6215–6221, 2011.