

On oracle-type local recovery guarantees in compressed sensing

Ben Adcock, Claire Boyer, Simone Brugiapaglia

▶ **To cite this version:**

Ben Adcock, Claire Boyer, Simone Brugiapaglia. On oracle-type local recovery guarantees in compressed sensing. Information and Inference, 2020. hal-01811947v2

HAL Id: hal-01811947
https://hal.science/hal-01811947v2

Submitted on 14 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Information and Inference: A Journal of the IMA (2020) Page 1 of 47
doi:10.1093/imaiai/drn000

On oracle-type local recovery guarantees in compressed sensing

BEN ADCOCK,
Simon Fraser University, Burnaby, BC, Canada
ben_adcock@sfu.ca

CLAIRE BOYER
Ecole normale supérieure, Paris, PSL University, France
claire.boyer@sorbonne-universite.fr

AND

SIMONE BRUGIAPAGLIA*
Concordia University, Montreal, QC, Canada
*Corresponding author: simone.brugiapaglia@concordia.ca

[Received on 23 March 2020]

We present improved sampling complexity bounds for stable and robust sparse recovery in compressed sensing. Our unified analysis based on ℓ^1 minimization encompasses the case where (i) the measurements are block-structured samples in order to reflect the structured acquisition that is often encountered in applications; (ii) the signal has an arbitrary structured sparsity, by results depending on its support S. Within this framework and under a random sign assumption, the number of measurements needed by ℓ^1 minimization can be shown to be of the same order than the one required by an oracle least-squares estimator. Moreover, these bounds can be minimized by adapting the variable density sampling to a given prior on the signal support and to the coherence of the measurements. We illustrate both numerically and analytically that our results can be successfully applied to recover Haar wavelet coefficients that are sparse in levels from random Fourier measurements in dimension one and two, which can be of particular interest in imaging problems. Finally, a preliminary numerical investigation shows the potential of this theory for devising adaptive sampling strategies in sparse polynomial approximation.

Keywords: compressed sensing, structured sparsity, oracle-type inequality, variable density sampling, local recovery guarantee.

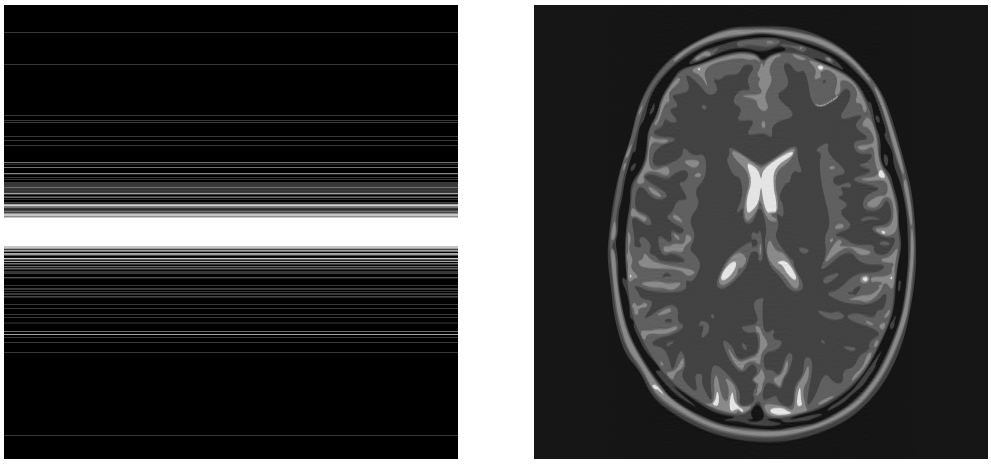
1. Introduction

1.1 *Motivations*

Standard Compressed Sensing (CS) concerns the recovery of a sparse vector x from linear measurements. The theory of CS is well-established, with one of its signature results being the existence of suitable decoders (for instance, based on convex optimization) that achieve recovery from near-optimal numbers of suitably-chosen measurements (e.g. Gaussian random measurements) scaling linearly with the sparsity and logarithmically with the ambient dimension.

However, many applications of CS exhibit more structure than sparsity alone. Hence there is a need to understand its performance for more structured signal models. With this in mind, the purpose of this paper is to provide sufficient conditions on the number of measurements required in order to ensure recovery of a structured sparse signal $x \in \mathbb{C}^n$ with random signs, via quadratically-constrained Basis

© The author 2020. Published by Oxford University Press on behalf of the Institute of Mathematics and its Applications. All rights reserved.



(a) (b)

Figure 1. Example of structured sampling in 2D MRI. In (a), samples are taken along horizontal lines in the acquisition domain. In (b), the reconstruction of a synthetic image has been generated using (qBP) and samples located as in (a).

Pursuit:

$$\min_{z\in\mathbb{C}^n} \|z\|_1 \qquad \text{such that} \qquad \|y-Az\|_2 \leqslant \eta.$$

Our analysis has two key features. First, we consider a general type of measurement matrix A, based on a block structure. Although the sampling is random, the block structure is a way to get a theoretical setting closer to realistic sampling than standard CS measurement matrix constructions: often in applications, only certain sampling patterns are allowed and those "joint" measurements are modeled here by the block structure of A. For instance, in Magnetic Resonance Imaging (MRI), a common practice is to acquire samples along radial lines or straight lines, as illustrated in Figure 1. Moreover, in many other applications, such as ultrasound imaging or interferometry, the acquisition is often constrained to specific sampling patterns [17, 19]. In addition, the block structure of A considered in this paper can also model the case of multiple-sensor data acquisition, considered in applications such as parallel MRI [24]. It should be noted that the block-structured acquisition encompasses standard CS strategies in which isolated measurements are sampled from a given isometry (see for instance [14, Chapter 12]).

Second, our analysis provides recovery guarantees that are local with respect to the support of the sparse vector x; in particular, they impose no signal model (e.g. the sparse model). As a result, they allow one to consider certain structured sparsity models for x. In recent years, structured sparsity has been proved to be a more appropriate prior than standard sparsity when dealing with real-world problems, such as imaging [2]. Structured sparsity is a key feature to consider in order to devise optimal sampling strategies for image reconstruction [3]. Moreover, it is able to leverage and theoretically legitimize block-structured sensing [7, 13].

A number of recent works have considered local recovery guarantees in CS [7, 13]. However, unlike in the classical setting, it was unknown whether the corresponding measurement conditions were optimal. Note that in the case of Gaussian measurements, in [5], a theoretical phase transition has been identified around the statistical dimension, denoted here by δ, of the descent cone associated to the ℓ^1 norm: if the number of measurements m is such that $m \geqslant \delta + \sqrt{n}$ then basis pursuit succeeds in

recovering an s-sparse vector x, if $m \leqslant \delta - \sqrt{n}$ then basis pursuit fails to recover x, both cases with high probability. However, as soon as the sampling is more structured, meaning that the sampling is based either on isolated measurements from a structured isometry (e.g. such as the Fourier transform), or on blocks of structured measurements, there is no optimality guarantee on the required number of measurements to ensure recovery. To address this, in this paper our measurement conditions for the (quadratically-constrained) Basis Pursuit decoder are compared with those of the oracle least-squares estimator. The latter relies on *a priori* knowledge of the support of x (hence the term 'oracle'), something which is of course not available to the former.

1.2 *Comparison with existing results*

In the seminal paper [9], under a random sign assumption on the signal to reconstruct, the authors proposed to draw uniformly at random rows from an isometry $A_0 = (d_k^*)_{1\leqslant k\leqslant n}$, leading to stable reconstruction with probability at least $1-\varepsilon$ with the following required number of measurements:

$$m \gtrsim s \cdot n \max_k \|d_k\|_\infty^2 \cdot \ln(n/\varepsilon)\,. \tag{1.1}$$

This result can be of interest when considering totally incoherent transforms such as the Fourier matrix for which $n \max_k \|d_k\|_\infty^2 = O(1)$. However, this is not relevant anymore in the case of coherent transforms, such as the Fourier-Haar transform used to model MRI acquisition, where $n \max_k \|d_k\|_\infty^2 = O(n)$.

In this paper, our results include the previous ones, but are also extended to: (i) the case of variable density sampling; (ii) stability robustness results when measurements are corrupted with bounded noise; (iii) structured measurements using blocks of measurements; (iv) optimization of the sampling density with respect to prior information on the signal support, such as structured sparsity.

In [7, 13], acquisition of i.i.d. blocks of measurements was introduced to model structured acquisition closer to applications constraints. In this paper, we allow non-identically distributed blocks of measurements, requiring only independence between the blocks. Note that in [13], this extended setting was also considered to handle parallel acquisition. However, the main results in [7, 13] were involving the maximum between two quantities in the required number of measurements. The latter prevents theoretically and numerically from any minimization of the obtained bound with respect to the way of drawing measurements. In this paper, with an additional assumption on the signal sign randomness, we derive a bound on the number of measurements depending only on one quantity making its minimization easier and analytically explicit. New results on optimal sampling strategies are presented showing that they should not only depend on the coherence of the sensing matrix (as in standard CS, see for instance [11] and [14, Chapter 12]) but also on the prior on the signal structure.

1.3 *Contributions*

In this paper, we extend the setting of block-structured sensing introduced in [6] and also considered in [7, 13] to the case where the blocks are not identically distributed and in which some of them can be deterministically chosen. In this setting, we derive stable and robust recovery guarantees for CS, while considering a random sign assumption on the signal of interest. The recovery guarantees are *local* (or equivalently, *nonuniform* or *signal-based*), in the sense that they ensure the recovery with high probability for a fixed signal.

Let S be the set of s largest absolute entries of a given signal and let F be the probability model used to draw random (possibly block-structured) measurements from a finite-dimensional isometry A_0. Our recovery guarantees are based on two notions of local coherence, denoted as $\Lambda(S,F)$ and $\Theta(S,F)$ such

that $\Lambda(S,F) \leqslant \Theta(S,F)$ and on a global coherence measure $\Gamma(F)$ (these three quantities are formally introduced in Definition 2.3). Using this notation, an *oracle-type inequality* is a condition on the number of measurements of the form

$$m \gtrsim \Lambda(S,F) \cdot \ln(|S|/\varepsilon), \tag{1.2}$$

which guarantees robust recovery from noisy measurements via the oracle least-squares estimator. First, we prove stable and robust recovery for CS with probability at least $1-\varepsilon$ under the condition

$$m \gtrsim \Theta(S,F) \cdot \ln^2(n/\varepsilon). \tag{1.3}$$

Moreover, we show that the oracle-type inequality

$$m \gtrsim \Lambda(S,F) \cdot \ln(n/\varepsilon), \tag{1.4}$$

is sufficient to guarantee stable and robust recovery under the extra assumption $\Lambda(S,F) \gtrsim \Gamma(F) \cdot \ln(n/\varepsilon)$, which is verified in cases of practical interest. In the case of standard sparsity and uniform random sampling as in [9], conditions (1.3) and (1.4) are implied by (1.1) and, more in general, they refine the measurement conditions given in [10, 15, 18] for the case of variable density sampling and standard sparsity.

A main consequence of Conditions (1.3) and (1.4) is to give an optimal sampling strategy in order to minimize the required number of measurements while taking into account prior information on the support S, such as structured sparsity. We derive a closed form expression for the drawing probability, i.e. how to choose the measurements, in Section 4. This is a substantial contribution since in previous CS approaches, only variable density sampling based on the sensing transform coherence was performed. Here, the optimal strategy is shown to be not only dependent on the sampling coherence but also on the signal structure.

For illustrative purposes, let us briefly describe the implications of our contribution to the case of random isolated Fourier measurements for the recovery a one-dimensional signal that is assumed to be sparse in levels with respect to the Haar transform (this case study is discussed in detail in Section 5.1). Let us assume the signal to have sparsities in levels $(s_j)_{0\leqslant j\leqslant J}$, i.e., $s_j = |\Omega_j \cap S|$, where $(\Omega_j)_{0\leqslant j\leqslant J}$ are Haar wavelet subbands. Moreover, let us divide the space of Fourier frequencies $\{-n/2+1,\dots,n/2\}$ into subbands $(W_j)_{0\leqslant j\leqslant J}$ and denote as $j(k)$ the frequency band associated with the k-th frequency. Then, minimizing the quantity $\Lambda(S,F)$ in (1.4) leads to drawing the k-th frequency with probability

$$\pi_k = \frac{2^{-j(k)}\sum_{j=1}^{J} 2^{-|j-j(k)|}s_j}{\sum_{\ell=1}^{n} 2^{-j(\ell)}\sum_{j'=1}^{J} 2^{-|j(\ell)-j'|}s_{j'}}. \tag{1.5}$$

The resulting sufficient condition on the number of measurements is

$$m \gtrsim \left(\sum_{j=1}^{J} s_j + \sum_{\substack{j'=1 \\ j'\neq j}}^{J} 2^{-|j-j'|}s_{j'} \right) \cdot \ln(s)\ln(n/\varepsilon). \tag{1.6}$$

This improves the previous conditions from [2, 7] by decreasing the interference between the sparsities in levels by a square-root factor. Namely, we have $2^{-|j-j'|}$ instead of $2^{-|j-j'|/2}$ in (1.6). Analogous considerations hold for the case of the two-dimensional Fourier-Haar transform with Fourier measurements structured along vertical or horizontal lines, discussed in Section 5.2.

Finally, the explicit dependence of the optimal sampling measure on the signal support allows for adaptive sampling strategies, which can be particularly relevant when the signal structure or the sampling coherence are not known *a priori*. Preliminary numerical experiments in Section 5.3 illustrate how our analysis can be applied to derive adaptive sampling strategies for sparse polynomial approximation.

1.4 *Organization of the paper*

In Section 2, we introduce the setting, the considered recovery algorithm, and the sampling strategy adopted to be compatible with physically-constrained acquisition. The definitions of crucial quantities involved in the theoretical analysis are also given therein. In Section 3, the main results are presented and compared with the oracle case. A main consequence of this work is discussed in Section 4, in which an optimal sampling strategy is proposed. Section 5 gathers illustrations of the results, in particular in the case of Fourier-wavelets transforms encountered in MRI applications and in function interpolation, where an adaptive sampling strategy is proposed. The proofs of the main results are organized in Appendices A-F.

2. Setting

In this section, we describe the formal setting of the paper. After introducing some standard notation in Section 2.1, we discuss the recovery strategies considered in the case of noiseless and noisy measurements in Section 2.2. In Section 2.3, we describe the sampling strategies analyzed in this paper; in particular, one can consider block-structured sampling (Section 2.3 (i)) and isolated measurements (Section 2.3 (ii)) from a finite-dimensional isometry. Finally, in Section 2.4 the main technical ingredients of the proposed theoretical analysis are introduced. We also discuss the random sign assumption on the signal to recover and define three key quantities (denoted by Θ, Λ, and Γ) that will play a major role in the analysis carried out in Section 3.

2.1 *Notation*

In this paper, n denotes the dimension of the signal to reconstruct. The notation $S \subseteq \{1,\dots,n\}$ refers to the support of the signal to reconstruct and define $S^c := \{1,\dots,n\} \setminus S$. The vectors $(e_i)_{1\leqslant i\leqslant d}$ denote the vectors of the canonical basis of $\mathbb{R}^d$, where d will be equal to n or $\sqrt{n}$, depending on the context. For every $x \in \mathbb{C}^d$, we define x_S to be the restriction of x to the components in S. Notice that x_S may be a $|S|$-dimensional or a d-dimensional vector, depending on the context; in the second case, the entries of x_S in S^c are set to be zero. Moreover, we set P_S to be the matrix defined by the linear projection $P_S x = x_S$ for every $x \in \mathbb{C}^d$. Again, P_S can be a $d \times d$ or a $|S| \times d$ matrix, depending on whether x_S is considered as a $|S|$-dimensional or as a d-dimensional vector. Observe that when x is supported on S, then also $x = P_S^* x_S$ holds. We will use the shorthand notation M_S to denote the matrix MP_S^*. Similarly, if M_k denotes a matrix indexed by k, then $M_{k,S} = M_k P_S^*$. For any matrix M, for any $1 \leqslant p,q \leqslant \infty$, the operator norm $\|M\|_{p\to q}$ is defined as

$$\|M\|_{p\to q} = \sup_{\|v\|_p \leqslant 1} \|Mv\|_q,$$

with $\|\cdot\|_p$ and $\|\cdot\|_q$ denoting the standard ℓ_p and ℓ_q norms. Note that for a matrix $M \in \mathbb{R}^{n\times n}$,

$$\|M\|_{\infty\to\infty} = \max_{1\leqslant i\leqslant n} \|e_i^* M\|_1.$$

The function sign : $\mathbb{R}^n \to \mathbb{R}^n$ is defined by

$$(\mathrm{sign}(x))_i = \begin{cases} 1 & \text{if} \quad x_i > 0 \\ -1 & \text{if} \quad x_i < 0 \\ 0 & \text{if} \quad x_i = 0, \end{cases}$$

and Id (or Id_n) will denote the (n-dimensional) identity matrix. We denote by $\mathscr{R}(M)$ the range of the matrix M and by $M^\dagger$ the left pseudo-inverse of M, meaning that if M has full column rank, $M^\dagger = (M^*M)^{-1}M^*$.

2.2 *Recovery techniques*

Let $x \in \mathbb{C}^n$ be supported on $S \subset \{1,\dots,n\}$. In the case of noiseless measurements, the collected data y can be written as follows

$$y = Ax, \tag{2.1}$$

where A is the sampling matrix. In order to recover x, we consider ℓ^1-minimization with equality constraint, also known as the Basis Pursuit (BP) optimization program:

$$\min_{z\in\mathbb{C}^n} \|z\|_1 \quad \text{such that} \quad y = Az. \tag{BP}$$

In the case where observations are corrupted with noise, we will assume the noise to be bounded. In particular, we will assume that there exists $\eta > 0$, supposed to be known, such that

$$y = Ax + \varepsilon, \qquad \|\varepsilon\|_2 \leqslant \eta. \tag{2.2}$$

In order to estimate x, we then consider the ℓ^1-minimization problem with inequality constraint, also called quadratically-constrained Basis Pursuit (qBP):

$$\min_{z\in\mathbb{C}^n} \|z\|_1 \quad \text{such that} \quad \|y - Az\|_2 \leqslant \eta. \tag{qBP}$$

2.3 *Sampling strategy*

GENERAL SETTING Given some distributions $(F_\ell)_{1\leqslant\ell\leqslant m}$ respectively on sets of $p_\ell \times n$ matrices, with $p_\ell \geqslant 1$ for $\ell = 1,\dots,m$, the sampling strategy consists in drawing m independent matrices $B_1,\dots,B_m$ where $B_\ell \sim F_\ell$ for $\ell = 1,\dots,m$ and forming the sensing matrix as follows:

$$A = \frac{1}{\sqrt{m}} \begin{pmatrix} B_1 \\ \vdots \\ B_m \end{pmatrix}, \qquad \text{with} \qquad B_\ell \sim F_\ell, \quad \text{for}\, \ell = 1,\dots,m. \tag{2.3}$$

We assume the sampling to be isotropic, in the sense that

$$\mathbb{E}(A^*A) = \mathbb{E}\left(\frac{1}{m}\sum_{\ell=1}^{m} B_\ell^* B_\ell\right) = \mathrm{Id}.$$

This abstract setting can be specialized to the case where we are given an orthogonal matrix $A_0 \in \mathbb{C}^{n\times n}$ with rows $(d_i^*)_{1\leqslant i\leqslant n}$ representing the set of possible linear measurements imposed by a specific sensor device. In particular, this framework encompasses the two following cases.

(i) **Block-structured sampling from a finite-dimensional isometry.**

Let $(\mathscr{I}_k)_{1\leqslant k\leqslant M}$ denote a partition of the set $\{1,\dots,n\}$, i.e. a family of disjoint subsets

$$\mathscr{I}_k \subset \{1,\dots,n\} \quad \text{s.t.} \quad \bigsqcup_{k=1}^{M} \mathscr{I}_k = \{1,\dots,n\}.$$

The rows $(d_i^*)_{1\leqslant i\leqslant n} \in \mathbb{C}^n$ of A_0 are then partitioned accordingly into a block dictionary $(D_k)_{1\leqslant k\leqslant M}$, such that

$$D_k = (d_i^*)_{i\in\mathscr{I}_k} \in \mathbb{C}^{|\mathscr{I}_k|\times n}.$$

Define the random blocks $B_1,\dots,B_m$ to be i.i.d. copies of a random block B such that

$$\mathbb{P}(B = D_k/\sqrt{\pi_k}) = \pi_k, \qquad \text{for} \quad k = 1,\dots,M,$$

where $(\pi_k)_{1\leqslant k\leqslant M}$ is a discrete probability distribution on $\{1,\dots,M\}$. Note that in this case, all the distributions (F_ℓ)'s are the same one, characterizing the law of the random block B described right above. The sensing matrix A is then constructed by randomly drawing blocks as follows:

$$A = \frac{1}{\sqrt{m}}\,(B_\ell)_{1\leqslant \ell\leqslant m}\,. \tag{2.4}$$

Moreover, thanks to the renormalization, the random sensing matrix A satisfies

$$\mathbb{E}(A^*A) = \frac{1}{m}\sum_{\ell=1}^{m}\mathbb{E}(B_\ell^*B_\ell) = \mathbb{E}(B^*B) = \sum_{k=1}^{M} D_k^*D_k = A_0^*A_0 = \mathrm{Id}, \tag{2.5}$$

since A_0 is orthogonal and $(D_k)_{1\leqslant k\leqslant M}$ is a partition of the rows of A_0.

(ii) **Isolated measurements from a finite-dimensional isometry** (standard CS).

This is a particular case of the setting described in (i), which is standard in CS: each block corresponds to a row of the matrix $A_0 = (d_1|d_2|\dots|d_n)^*$. Therefore, the sensing matrix is constructed by stacking random vectors drawn from the set of row vectors $\{d_1^*,\dots d_n^*\}$ and can be written as follows:

$$A = \frac{1}{\sqrt{m}}\,(a_\ell^*)_{1\leqslant \ell\leqslant m}\,, \tag{2.6}$$

where the random vectors $(a_\ell)_{1\leqslant \ell\leqslant m}$ are i.i.d. copies of a random vector a such that

$$\mathbb{P}(a = d_j/\sqrt{\pi_j}) = \pi_j,$$

for all $1 \leqslant j \leqslant n$. Here again all the (F_ℓ)'s consists in the same distribution, designating the law of the random vector a. The isotropy condition, i.e. $\mathbb{E}(A^*A) = \mathbb{E}\left(\frac{a_\ell a_\ell^*}{\pi_\ell}\right) = \mathrm{Id}$, is also satisfied.

REMARK 2.1 The setting can be modified in order to encompass partial deterministic sampling. Consider some distributions $(F_\ell)_{m_0+1\leqslant \ell\leqslant m}$ respectively on sets of $p_\ell \times n$ matrices, with $p_\ell \geqslant 1$ for $\ell = m_0+1,\dots,m$, the sampling strategy consists in drawing $m-m_0$ independent matrices $B_{m_0+1},\dots,B_m$ where $B_\ell \sim F_\ell$ for $\ell = m_0+1,\dots,m$ and forming the sensing matrix as follows:

$$A = \frac{1}{\sqrt{m-m_0}}\begin{pmatrix} B_1 \\ \vdots \\ B_m \end{pmatrix}, \qquad \text{with} \qquad B_\ell \sim F_\ell, \quad \text{for}\, \ell = m_0+1,\dots,m, \tag{2.7}$$

while $B_1,\dots,B_{m_0}$ can be *deterministically* chosen. We still need the sampling to be isotropic, in the sense that

$$\mathbb{E}(A^*A) = \sum_{\ell=1}^{m_0} B_\ell^* B_\ell + \mathbb{E}\left(\frac{1}{m-m_0}\sum_{\ell=m_0+1}^{m} B_\ell^* B_\ell\right) = \mathrm{Id}. \tag{2.8}$$

This possible extension is motivated as follows. In applications where the sensing matrix is randomly extracted from a Fourier/wavelets transform, multi-level sampling strategies have been proved to be highly effective (see [2]). In particular, one may want to partition the Fourier space into levels and then *saturate* (i.e., fully sample) some of them [16]. Usually, the saturated levels are those corresponding to the lowest frequencies. However, saturating some levels using a fully random procedure as in (i) leads to a suboptimal sampling rate, due to the coupon collector effect. Allowing partial deterministic sampling of m_0 blocks is a simple way to circumvent this problem. To avoid heavy notation, we will state the main results and the proofs with $m_0 = 0$.

2.4 *Assumptions*

We assume that the signal we aim at reconstructing satisfies a random sign property, defined as follows.

ASSUMPTION 2.2 For any vector $x \in \mathbb{R}^n$ or $x \in \mathbb{C}^n$ supported on S, we will say that x satisfies the random sign assumption if $\mathrm{sign}(x_S)$ is respectively a Rademacher or Steinhaus sequence.

The following quantities are crucial in the recovery guarantees.

DEFINITION 2.3 Consider a block sampling strategy as previously described in (2.3) where $(B_k)_{1\leqslant k\leqslant m}$ are random blocks such that $B_k \sim F_k$. We denote the collection of probability distributions by $F = (F_k)_{1\leqslant k\leqslant m}$. Let $S \subset \{1,\dots,n\}$. Define the quantities $\Theta(S,F)$, $\Lambda(S,F)$, and $\Gamma(F)$ to be positive real numbers such that

$$\Theta(S,F) \geqslant \|B_\ell^* B_{\ell,S}\|_{\infty\to\infty} \qquad \text{a.s.} \quad B_\ell \sim F_\ell, \quad \ell = (m_0+)1,\dots,m, \tag{2.9}$$

$$\Lambda(S,F) \geqslant \left\|B_{\ell,S}^* B_{\ell,S}\right\|_{2\to 2} \qquad \text{a.s.} \quad B_\ell \sim F_\ell, \quad \ell = (m_0+)1,\dots,m, \tag{2.10}$$

$$\Gamma(F) \geqslant \|B_\ell\|_{1\to 2}^2 = \max_{1\leqslant i\leqslant n} \|B_\ell e_i\|_2^2 \qquad \text{a.s.} \quad B_\ell \sim F_\ell, \quad \ell = (m_0+)1,\dots,m. \tag{2.11}$$

Typically, but not always, $\Theta(S,F)$, $\Lambda(S,F)$ and $\Gamma(F)$ will be taken as the least-upper bounds. Note that since $\left\|B_S^* B_S\right\|_{2\to 2} \leqslant \left\|B_S^* B_S\right\|_{\infty\to\infty} \leqslant \|B^* B_S\|_{\infty\to\infty}$ (due to, e.g. [14, Lemma A.8 and Remark A.10]), if $\Theta(S,F)$ and $\Lambda(S,F)$ are taken as least-upper bounds, then

$$\Lambda(S,F) \leqslant \Theta(S,F). \tag{2.12}$$

For the sake of readability, sometimes we will simply use Θ, Λ, and Γ to refer to $\Theta(S,F)$, $\Lambda(S,F)$, and $\Gamma(S,F)$, respectively.

In the case of the block-structured finite setting, one considers a block dictionary $(D_k)_{1\leqslant k\leqslant M}$ as in Section 2.3(i). Given the quantities in Definition 2.3, one can derive the following upper bounds: let

$S \subset \{1,\dots,n\}$ and π be a probability distribution on $\{1,\dots,M\}$,

$$\Theta(S,\pi) \geqslant \max_{1\leqslant k\leqslant M} \frac{1}{\pi_k} \|D_k^* D_{k,S}\|_{\infty\to\infty} = \max_{1\leqslant k\leqslant M} \max_{1\leqslant i\leqslant n} \frac{\|e_i^* D_k^* D_{k,S}\|_1}{\pi_k}, \tag{2.13}$$

$$\Lambda(S,\pi) \geqslant \max_{1\leqslant k\leqslant M} \frac{1}{\pi_k} \left\|D_{k,S}^* D_{k,S}\right\|_{2\to 2}, \tag{2.14}$$

$$\Gamma(\pi) \geqslant \max_{1\leqslant k\leqslant M} \frac{1}{\pi_k} \|D_k\|_{1\to 2}^2 = \max_{1\leqslant k\leqslant M} \max_{1\leqslant i\leqslant n} \frac{1}{\pi_k} \|D_k e_i\|_2^2. \tag{2.15}$$

In the case of isolated measurements drawn from an isometry, one considers the rows $\left(d_k^*\right)_{1\leqslant k\leqslant n}$ of an orthogonal matrix A_0 as in Section 2.3(ii). Given the quantities in Definition 2.3, one can derive the following upper bounds: let $S \subset \{1,\dots,n\}$ and π be a probability distribution on $\{1,\dots,n\}$,

$$\Theta(S,\pi) \geqslant \max_{1\leqslant k\leqslant n} \frac{1}{\pi_k} \|d_k\|_\infty \|d_{k,S}\|_1, \tag{2.16}$$

$$\Lambda(S,\pi) \geqslant \max_{1\leqslant k\leqslant n} \frac{1}{\pi_k} \left\|d_{k,S}\right\|_2^2, \tag{2.17}$$

$$\Gamma(\pi) \geqslant \max_{1\leqslant k\leqslant n} \frac{1}{\pi_k} \|d_k\|_\infty^2. \tag{2.18}$$

In (2.18), one may recognize the standard definition of the global coherence in CS, see for instance [8, 14].

3. Main results

In this section, we derive recovery guarantees for (BP) and (qBP) under a random sign assumption on the signal . They reveal sufficient conditions on the required number of measurements, provided that one can evaluate Θ, Λ, and Γ defined in Section 2.4.

OVERVIEW OF THE MAIN RESULTS Throughout the section, our benchmark will be an oracle-type inequality discussed in Section 3.1, namely

$$m \gtrsim \Lambda(S,F) \cdot \ln(|S|/\varepsilon), \tag{3.1}$$

which is proved to be sufficient for the robust recovery of an S-sparse via oracle-least squares with probability at least $1-\varepsilon$ in Proposition 3.1. We will refer to (3.1) as an *oracle-type* inequality. Notice that the oracle-least squares estimator requires an *a priori* knowledge of S to recover the signal, whereas the (BP) and (qBP) programs do not.

In Sections 3.2 and 3.3 we make a first step towards oracle-type inequalities for CS. In particular, we show that

$$m \gtrsim \Theta(S,F) \cdot \ln^2(n/\varepsilon), \tag{3.2}$$

is sufficient for the exact recovery from noiseless measurements (Theorem 3.3) or robust recovery from noisy measurements (Theorem 3.6) of a signal supported on S with probability at least $1-\varepsilon$. Note that, besides the additional logarithmic factor, (3.2) is not necessarily an oracle-type inequality, in view of (2.12).

We progressively improve bounds on m for CS recovery via (BP) and (qBP) in Section 3.4, presenting three results towards oracle-type inequalities. Theorem 3.8 requires a bound of the form

$$m \gtrsim \max_{j \in S^c} \Lambda(S \cup \{j\}, F) \cdot \ln^2(n/\varepsilon),$$

which is of oracle type up to enlarging the support by one element. Theorems 3.9 and 3.10 achieve oracle-type requirements on m at the price of an extra assumption involving $\Lambda(S,F)$ and $\Gamma(F)$. They only differ by a logarithmic factor. In particular, in Theorem 3.9 robust recovery from noisy measurements (or exact recovery from noiseless measurement) is guaranteed with probability $1-\varepsilon$ if

$$m \gtrsim \Lambda(S,F) \cdot \ln^2(n/\varepsilon),$$

and provided that $\Lambda(S,F) \gtrsim \Gamma(F)$. Theorem 3.10 achieves the same recovery guarantees if

$$m \gtrsim \Lambda(S,F) \cdot \ln(n/\varepsilon),$$

and under the extra assumption $\Lambda(S,F) \gtrsim \Gamma(F)\ln(n/\varepsilon)$. These extra assumptions do not turn out to be restrictive in practice (see Section 5). Finally, we note that the assumptions of Theorem 3.8, 3.9, and 3.10 are sufficient to guarantee stable and robust recovery from noisy measurements for both (qBP) and the oracle least-squares estimator (see Remark 3.11).

3.1 *Preliminary: an oracle inequality*

In this section, we derive a lower bound on the number of measurements sufficient to obtain robust recovery using an oracle least-squares estimator.

PROPOSITION 3.1 Let $x \in \mathbb{R}^n$ or $\mathbb{C}^n$ be a vector supported on a set S of size s and suppose we are given noisy measurements $y = Ax + \varepsilon$, with $\|\varepsilon\|_2 \leqslant \eta$. Then, there exist universal constants $c_0, C_0 > 0$ such that, for every $0 < \varepsilon < 1$ and provided

$$m \geqslant c_0 \cdot \Lambda(S,F) \cdot \ln\left(\frac{2s}{\varepsilon}\right), \tag{3.3}$$

the following holds with probability at least $1-\varepsilon$: the matrix A_S has full column rank and the oracle-least squares estimator $x^\star \in \mathbb{C}^n$ of the system $y = Az$, defined by

$$x^\star_S = (A_S)^\dagger y, \quad x^\star_{S^c} = 0. \tag{3.4}$$

satisfies the error estimate

$$\|x - x^\star\|_2 \leqslant C_0 \eta. \tag{3.5}$$

Possible values for the constants are $C_0 = \sqrt{2}$ and $c_0 = 32/3$.

Proof. Recalling the definition (3.4) of $x^\star$ and the fact that both x and $x^\star$ are supported on S, one has

$$\begin{aligned}
\|x^\star - x\|_2 &= \|(A_S)^\dagger y - x_S\|_2 = \|(A_S)^\dagger (Ax+\varepsilon) - x_S\|_2 = \|(A_S)^\dagger \varepsilon + (A_S)^\dagger A_{S^c} x_{S^c}\|_2 \\
&\leqslant \|A_S^\dagger\|_{2\to 2}\|\varepsilon\|_2 + \|(A_S^* A_S)^{-1}\|_{2\to 2}\|A_S^* A_{S^c}\|_{1\to 2}\|x_{S^c}\|_1 \\
&\leqslant \frac{1}{\sigma_{\min}(A_S)}\|\varepsilon\|_2 + \|(A_S^* A_S)^{-1}\|_{2\to 2}\|A_S^* A_{S^c}\|_{1\to 2}\|x_{S^c}\|_1.
\end{aligned}$$

If $\|A_S^* A_S - P_S\|_{2\to 2} \leqslant \delta$, then $\sigma_{\min}(A_S) \geqslant \sqrt{1-\delta}$ and $\|(A_S^* A_S)^{-1}\|_{2\to 2} \leqslant \frac{1}{1-\delta}$. Using Lemma D.1, if

$$m \geqslant \frac{1+2\delta/3}{\delta^2/2} \cdot \Lambda(S,F) \cdot \ln\left(\frac{2s}{\varepsilon}\right),$$

then $\|A_S^* A_S - P_S\|_{2\to 2} \leqslant \delta$. Considering that $\|x_{S^c}\|_1 = 0$ and fixing $\delta = 1/2$ leads to the desired result with the specified constants.

□

REMARK 3.2 In the following, we are going to derive "oracle-type" estimate for signal recovery via (BP) and (qBP). In this paper, "oracle-type" estimates will refer to the bound on the number of measurements, i.e. bounds of the form (3.3). They will not concern the robustness bound obtained in (3.5), which we will discuss later.

Proposition 3.1 implies robust recovery of sparse vectors when measurements are corrupted with bounded noise. In fact, it is also possible to prove the stability of the oracle least-squares estimator with respect to the standard sparsity model by considering a condition on m slightly stronger than (3.3). For the sake of readability, in the following results we will bypass this additional technical difficulty by focusing only robust sparse recovery. For a more extended discussion on stability, we refer to Remark 3.11.

3.2 *Noiseless recovery*

Our main result for the success of (BP) with an abstract block-structured framework presented in Section 2.3 is the following.

THEOREM 3.3 Let $x \in \mathbb{R}^n$ or $\mathbb{C}^n$ be a vector supported on S, such that $\operatorname{sign}(x_S)$ forms a Rademacher or Steinhaus sequence. Let A be the random sensing matrix defined in (2.3) associated with parameter $\Theta(S,F)$. Suppose we are given the data $y = Ax$. Then, given $0 < \varepsilon < 1$ and provided

$$m \geqslant c_1 \cdot \Theta(S,F) \cdot \ln^2\left(\frac{6n}{\varepsilon}\right),$$

for c_1 a numerical constant (for instance $c_1 = 82$), the vector x is the unique minimizer of the basis pursuit program (BP) with probability at least $1-\varepsilon$.

The proof is given in Appendix A.1.

REMARK 3.4 In standard CS, the number of measurements usually depends on the degree of sparsity s (such that $|S| \leqslant s$) on the one hand and on the sampling coherence on the other hand. Here, the quantity $\Theta(S,F)$ is encapsulating both information. This way, the use of coherent transforms is not prohibited anymore as soon as the support structure is adapted to it. Note that in the case of isolated measurements described in Section 2.3-(ii), the quantity $\Theta(S,F)$ can be bounded from above as follows, leading to the standard CS-type estimate:

$$\Theta(S,F) \leqslant s \cdot \sup_{a\sim F} \|a\|_\infty^2.$$

However, we observe that this upper bound is too crude in general, except under the following assumptions:

- there is no structure in the signal sparsity, meaning that the only prior on S is that $|S| = s$;

- the sensing matrix A is totally incoherent, meaning that all the entries of A have the same magnitude, $\forall(i,j), |A_{ij}| = \|A\|_{1\to\infty} = 1$ (this for instance the case of subsampled Fourier matrix).

REMARK 3.5 The bound on the number of measurements in Theorem 3.3 depends only on one quantity, namely $\Theta(S,F)$, which, in turn, depends on the signal support and on the way of drawing blocks. On the contrary, in [7, 13], the authors derived similar results with a bound on m depending on the maximum between two quantities (one of them corresponding to $\Theta(S,F)$ in this paper). However, the authors [7, 13] have not considered the random sign assumption. The latter was useful to obtain a closed-form expression for the "coherence" $\Theta(S,F)$. We will see that this is key to design optimal drawing strategies in Section 4.

3.3 *Robustness*

Theoretical guarantees can also be obtained when the observation vector y is corrupted by noise.

THEOREM 3.6 Let $x \in \mathbb{R}^n$ or $\mathbb{C}^n$ be a vector supported on S, such that $\mathrm{sign}(x_S)$ forms a Rademacher or Steinhaus sequence and $|S| = s$. Let A be the random sensing matrix defined in (2.3) with parameter $\Theta(S,F)$. Suppose that the data y is given such that $y = Ax + \varepsilon$ with $\|\varepsilon\|_2 \leqslant \eta$ for some given $\eta > 0$. Then, given $0 < \varepsilon < 1$ and provided

$$m \geqslant c_1 \cdot \Theta(S,F) \cdot \ln^2\left(\frac{6n}{\varepsilon}\right),$$

for c_1 a numerical constant (for instance $c_1 = 889$), a minimizer $x^\sharp$ of (qBP) satisfies

$$\|x - x^\sharp\|_2 \leqslant (C_1 + C_2\sqrt{s})\eta, \tag{3.6}$$

with probability at least $1-\varepsilon$, where C_1, C_2 are numerical constants.

The proof of Theorem 3.6 is given in Appendix A.2.

REMARK 3.7 The proof of Theorem 3.6 reveals a more precise recovery guarantee, which accounts for the stability with respect to the standard sparsity model. Namely, it is possible to generalize the theorem to the case $x \in \mathbb{R}^n$ or $\mathbb{C}^n$ where S is a set of indices corresponding to the s largest absolute entries of x (or, more in general, any subset of s indices). Under the same conditions as in Theorem 3.6, then, with high probability, one has

$$\|x - x^\sharp\|_2 \leqslant (C_1 + C_2\sqrt{s})\eta + C_3\|x - x_S\|_1. \tag{3.7}$$

This matches the bound derived in the standard setting of CS, see for instance [14, Theorem 12.22], which may seem a bit disappointing. However, to our knowledge, this is the first result of stability and robustness derived in the case of the random sign assumption.

HOW MUCH ARE WE OFF THE ORACLE ESTIMATE OF PROPOSITION 3.1? Regarding the number of measurements m, the conditions required by Theorems 3.3 and 3.6 depend on $\Theta(S,F)$, instead of $\Lambda(S,F)$ as in Proposition 3.1. Therefore, by (2.12), we only obtained an upper bound to the oracle estimate.

Regarding the error bound (3.6), one can notice that the noise level η is amplified by an additional factor $\sqrt{s}$, when compared to (3.5). Nonuniform approaches are known to suffer from this extra $\sqrt{s}$-factor in the robustness bound, see for instance [14, Theorem 12.22], although this may be an artifact

of the proof strategies employed so far. To the best our knowledge, the only better bound of the form $c \cdot \eta$ for c some constant has been obtained in [21, Proposition 2.6]. However, such a bound can be obtained so far only for Gaussian measurements since they are based on the control of the minimum conic singular value of the sensing matrix. Controlling the statistical dimension of the descent cone for structured measurements still remains an open question. That is why in the sequel we will leave aside the question of improving the robustness bound of a $\sqrt{s}$-factor and we will focus on deriving oracle-type bounds in terms of number of measurements.

3.4 *Getting oracle-type bounds for the required number of measurements*

We start by presenting a first oracle-type estimate where the quantity $\Theta(S,F)$ is replaced with $\Lambda(S\cup\{j\},F)$, for some $j \in S^c$. As discussed in Proposition 3.1, if the support S is known, the least amount of measurements needed for robust recovery should be of the order of $\Lambda(S,F)$. Indeed, $\Lambda(S,F)$ controls the condition number of the sensing matrix A restricted to the space of vectors supported on S. Theorem 3.8 almost invoke this condition number by slightly enlarging S by one off-support component. The proof of this result is given in Appendix A.3.

THEOREM 3.8 Let $x \in \mathbb{R}^n$ or $\mathbb{C}^n$ be a vector supported on S, such that $\operatorname{sign}(x_S)$ forms a Rademacher or Steinhaus sequence. Let A be the random sensing matrix defined in (2.3) with parameter $\Lambda(S,F)$ and let $y = Ax + \varepsilon$, with $\|\varepsilon\|_2 \leqslant \eta$. Then, there exist constants $c_1, C_1, C_2 > 0$ such that the following holds. For every $0 < \varepsilon < 1$, if

$$m \geqslant c_1 \max_{j\in S^c} \Lambda(S\cup\{j\},F) \ln\left(\frac{6(s+1)(n-s)}{\varepsilon}\right) \ln\left(\frac{6n}{\varepsilon}\right),$$

then, with probability at least $1-\varepsilon$, a minimizer $x^\sharp$ of (qBP) satisfies

$$\|x - x^\sharp\|_2 \leqslant (C_1 + C_2\sqrt{s})\eta.$$

In particular, in the noiseless case (i.e., $\eta = 0$), x is exactly recovered via (BP) with probability at least $1-\varepsilon$ and with constant $c_1 = 19$.

In the following, we propose two oracle-type inequalities that improve Theorems 3.3 and 3.6, in which the bound of the required number of measurements actually depends on $\Lambda(S,F)$ instead of $\Theta(S,F)$, at the price of an extra assumption involving $\Lambda(S,F)$ and $\Gamma(F)$.

THEOREM 3.9 Let $x \in \mathbb{R}^n$ or $\mathbb{C}^n$ be a vector supported on S with $|S| = s \leqslant n/2$, such that $\operatorname{sign}(x_S)$ forms a Rademacher or Steinhaus sequence. Let A be the random sensing matrix defined in (2.3) associated with parameters $\Lambda(S,F)$ and $\Gamma(S,F)$. Suppose we are given the data $y = Ax + \varepsilon$ such that $\|\varepsilon\|_2 \leqslant \eta$. Then, there exist constants $c_1, c_2, C_1, C_2 > 0$ such that the following holds. For every $0 < \varepsilon < 1$, if

$$\Lambda(S,F) \geqslant c_1 \cdot \Gamma(F), \tag{3.8}$$

and if

$$m \geqslant c_2 \cdot \Lambda(S,F) \cdot \ln^2\left(\frac{3n}{\varepsilon}\right),$$

then, with probability at least $1-\varepsilon$, a minimizer $x^\sharp$ of (qBP) satisfies

$$\|x - x^\sharp\|_2 \leqslant (C_1 + C_2\sqrt{s})\eta.$$

Making an assumption on $\Lambda(S,F)$ stronger than (3.8) allows to "kill" an extra log factor in the required number of measurements. This is the purpose of the following theorem.

THEOREM 3.10 Let $x \in \mathbb{R}^n$ or $\mathbb{C}^n$ be a vector supported on S with $|S| = s \leqslant n/2$, such that $\operatorname{sign}(x_S)$ forms a Rademacher or Steinhaus sequence. Let A be the random sensing matrix defined in (2.3) associated with parameters $\Lambda(S,F)$ and $\Theta(S,F)$. Suppose we are given the data $y = Ax + \varepsilon$, with $\|\varepsilon\|_2 \leqslant \eta$. Then, there exist constants $c_1, c_2, C_1, C_2 > 0$ such that the following holds. For every $0 < \varepsilon < 1$, if

$$\Lambda(S,F) \geqslant c_1 \cdot \Gamma(F) \cdot \ln(3n/\varepsilon), \tag{3.9}$$

and if

$$m \geqslant c_2 \cdot \Lambda(S,F) \cdot \ln\left(\frac{3n}{\varepsilon}\right),$$

then, with probability at least $1-\varepsilon$, a minimizer $x^\sharp$ of (qBP) satisfies

$$\|x - x^\sharp\|_2 \leqslant (C_1 + C_2\sqrt{s})\eta.$$

In particular, in the noiseless case (i.e., $\eta = 0$) the signal x is exactly recovered via (BP) with probability at least $1-\varepsilon$ and with constants $c_1 = 50$ and $c_2 = 100$.

The proof of Theorem 3.10 is given in Appendix A.4. Moreover, the proof of Theorem 3.9 easily follows from that of Theorem 3.10. Although the bound on the number of measurements in Theorem 3.10 only depends on $\Lambda(S,F)$ that encapsulates support information and sampling coherence related to the support S, the global coherence is somewhat restricted by the extra condition (3.9). Finally, we note in passing that the numerical values proposed for the constants c_1 and c_2 in the statement of Theorem 3.10 could be further optimized.

We will see that Assumptions (3.8) or (3.9) can be satisfied in practice in some applications in Section 5. However, in the case of isolated measurements, by recalling (2.17) and (2.18) condition (3.9) can be rewritten as follows:

$$\sup_{a \sim F} \|a_S\|_2^2 \geqslant c_1 \cdot \sup_{a \sim F} \|a\|_\infty^2 \ln(3n/\varepsilon),$$

in which $\sup_{a\sim F} \|a\|_\infty^2$ is the global sampling coherence [8]. By subsampling an isometry as in the setting (ii) of Section 2.3, and considering totally incoherent sampling, as for instance sampling Fourier frequencies uniformly at random, the previous condition becomes

$$s \geqslant c_1 \cdot \ln(3n/\varepsilon),$$

which is not too restrictive in practice.

REMARK 3.11 The conditions involving m, Γ and Λ in Theorems 3.8, 3.9, or 3.10 are sufficient to guarantee stable and robust recovery estimates of the form (3.7). Moreover, it is possible to show that the same assumptions are also sufficient to guarantee stable and robust recovery for the oracle least-squares estimator. For further details, we refer the reader to the discussion in Appendix B.

4. Deriving an optimal sampling strategy

The purpose of this section is to employ the results in Section 3 to derive optimal sampling strategies within the framework of block-structured sampling and isolated measurements from a finite-dimensional

isometry illustrated in Section 2.3. Indeed, the bounds on m just presented in Section 3 have the remarkable feature to be the first ones in CS with generalized structured sampling, depending only on one quantity, namely either Θ or Λ. This is the key to derive theoretical optimal sampling strategies: minimizing such bounds on m with respect to the sampling distribution, leads to the first theoretical closed form for the sampling probabilities when using structured blocks of measurements. All the proofs are presented in Appendix E.

The case of block-structured sampling strategy was introduced in [7] and adapted to parallel acquisition in [13], where variable density sampling with structured acquisition was considered. However, based on the analysis proposed in these works, minimizing the bound on the number of measurements with respect to π is not trivial and no closed form for π could be analytically given in general. In the following, Propositions 4.1 and 4.2 are the first theoretical results in CS on an optimal way to perform structured acquisition while taking into account partially coherent transforms and structured sparsity. We start by deriving optimal sampling strategies from Theorems 3.3 and 3.6 in the finite setting described in Section 2.3 (i).

PROPOSITION 4.1 In a finite setting where we sample blocks in a partitioned isometry as in Section 2.3 (i), Equation (2.4), the drawing probability minimizing $\Theta(S,\pi)$ is the following:

$$\forall k \in \{1,\dots,M\}, \quad \pi_k = \pi_k^{\Theta} = \frac{\|D_k^* D_{k,S}\|_{\infty\to\infty}}{\sum_{\ell=1}^{M} \|D_\ell^* D_{\ell,S}\|_{\infty\to\infty}}. \tag{4.1}$$

With such a choice for π, the required number m of blocks of measurements in Theorems 3.3 and 3.6 can be then rewritten as follows:

$$m \geqslant c \cdot \sum_{\ell=1}^{M} \|D_\ell^* D_{\ell,S}\|_{\infty\to\infty} \cdot \ln^2(6n/\varepsilon), \tag{4.2}$$

for a suitable universal constant $c > 0$. These conditions ensure noiseless recovery via (BP), or stable and robust recovery via (qBP), with probability at least $1-\varepsilon$.

At the price of the extra assumption, as (3.8) or (3.9) on $\Lambda(S,\pi)$ and $\Gamma(\pi)$ in Theorems 3.9 or 3.10, one could choose a drawing probability minimizing $\Lambda(S,\pi)$ instead of $\Theta(S,\pi)$. This is the purpose of the following results, corresponding to the case of block-structured and isolated measurements, respectively. For the sake of simplicity, we just illustrate the derivation from Theorem 3.10.

PROPOSITION 4.2 In a finite setting where we sample blocks in a partitioned isometry as in Section 2.3 (i), Equation (2.4), assume that $|S| \leqslant n/2$ and that

$$1 \geqslant c_1 \cdot \max_{1\leqslant k\leqslant M} \frac{\|D_k\|_{1\to 2}^2}{\|D_{k,S}^* D_{k,S}\|_{2\to 2}} \ln(3n/\varepsilon), \tag{4.3}$$

for c_1 a universal constant. Then the optimal drawing probability minimizing $\Lambda(S,\pi)$ is

$$\forall k \in \{1,\dots,M\}, \quad \pi_k = \pi_k^{\Lambda} = \frac{\|D_{k,S}^* D_{k,S}\|_{2\to 2}}{\sum_{\ell=1}^{M} \|D_{\ell,S}^* D_{\ell,S}\|_{2\to 2}}. \tag{4.4}$$

With such a choice for π, the required number m of blocks of measurements in Theorem 3.10 can be then rewritten as follows:

$$m \geqslant c_2 \cdot \sum_{\ell=1}^{M} \|D_{\ell,S}^* D_{\ell,S}\|_{2\to 2} \cdot \ln(6n/\varepsilon), \tag{4.5}$$

for a suitable universal constant $c_2 > 0$. These conditions ensure noiseless recovery via (BP), or stable and robust recovery via (qBP), with probability at least $1 - \varepsilon$.

In the following, we specialize the previous result to the case of isolated measurements subsampled from an isometry, described in Section 2.3 (ii).

PROPOSITION 4.3 In a finite setting where we sample isolated measurements as in Section 2.3 (ii), Equation (2.6), the drawing probability minimizing $\Theta(S,\pi)$ is the following:

$$\forall k \in \{1,\dots,n\}, \quad \pi_k = \pi_k^{\Theta} = \frac{\|d_k\|_\infty \|d_{k,S}\|_1}{\sum_{\ell=1}^n \|d_\ell\|_\infty \|d_{\ell,S}\|_1}. \tag{4.6}$$

With such a choice for π, the required number m of measurements in Theorems 3.3 and 3.6 can be then rewritten as follows:

$$m \geqslant c \cdot \sum_{\ell=1}^n \|d_\ell\|_\infty \|d_{\ell,S}\|_1 \cdot \ln^2(6n/\varepsilon), \tag{4.7}$$

for a suitable universal constant $c > 0$. These conditions ensure noiseless recovery via (BP), or stable and robust recovery via (qBP), with probability at least $1 - \varepsilon$.

Let us put the previous result into context. Variable density strategies were introduced to deal with partially coherent transforms (such as in the case of MRI) [10, 15, 18]. The idea is to sample the more coherent atoms $(d_k)_k$ with higher probability by setting for all $k \in \{1,\dots,n\}$

$$\pi_k = \frac{\|d_k\|_\infty^2}{\sum_{\ell=1}^n \|d_\ell\|_\infty^2}, \tag{4.8}$$

meaning that we tend to sample more where the sampling transform is coherent. With such a choice, one can ensure robust recovery with probability $1 - \varepsilon$ with a number of measurements of the order

$$m \gtrsim s \cdot \sum_{\ell=1}^n \|d_\ell\|_\infty^2 \cdot \ln(n/\varepsilon).$$

For instance in the case of MRI, $\sum_{\ell=1}^n \|d_\ell\|_\infty^2 = O(\ln(n))$, making bounds on m more realistic than those based on the global coherence $n \max_k \|d_k\|_\infty^2 = O(n\ln(n))$. The sampling strategy defined by (4.6) refines (4.8) since it takes into account a more accurate notion of local coherence, involving both the coherence of the individual atoms $(d_k)_k$ and its interaction with prior information about the support of the signal. In this way, the sampling strategy can be optimized to recover signals with structured sparsity, as discussed in Sections 5.1 and 5.2 for the case of sparsity in levels. Moreover, the explicit dependence of the sampling probability on the support in (4.6) enables us to devise iterative adaptive sampling strategies (see Section 5.3).

Still in the case of isolated measurements, one can improve the previous proposition by deriving sampling strategies derived from oracle-type bounds on the number of measurements. The following result applies Theorem 3.10 to the setting illustrated in Section 2.3 (ii).

PROPOSITION 4.4 In a finite setting where we sample isolated measurements as in Section 2.3 (ii), Equation (2.6), assume that $|S| \leqslant n/2$ and that

$$1 \geqslant c_1 \cdot \max_{1 \leqslant k \leqslant n} \frac{\|d_k\|_\infty^2}{\|d_{k,S}\|_2^2} \ln(3n/\varepsilon), \tag{4.9}$$

for c_1 a universal constant. Then the optimal drawing probability minimizing $\Lambda(S,\pi)$ is

$$\forall k \in \{1,\dots,n\}, \quad \pi_k = \pi_k^{\Lambda} = \frac{\|d_{k,S}\|_2^2}{\sum_{\ell=1}^n \|d_{\ell,S}\|_2^2}. \tag{4.10}$$

The required number m of measurements in Theorem 3.10 can be then rewritten as follows:

$$m \geqslant c_2 \cdot \sum_{\ell=1}^n \|d_{\ell,S}\|_2^2 \cdot \ln(6n/\varepsilon), \tag{4.11}$$

for a suitable universal constant $c_2 > 0$. These conditions ensure noiseless recovery via (BP), or stable and robust recovery via (qBP), with probability at least $1-\varepsilon$.

Finally, one could resort to Theorem 3.8 in order to remove condition (4.9), at the price of overestimating the support size by one element and of an extra log factor.

PROPOSITION 4.5 Let $x \in \mathbb{C}^n$ be a vector supported on S, such that $|S| = s$ and $\operatorname{sign}(x_S)$ forms a Rademacher or Steinhaus sequence. Let A be the random sensing matrix defined in (2.3) with parameter $\Lambda(S,F)$. If $0 < \varepsilon < 1$ and

$$m \geqslant c_1 \cdot \max_{j \in S^c} \Lambda(S \cup \{j\}, F) \cdot \ln^2\left(\frac{6n}{\varepsilon}\right),$$

with $c_1 > 0$ a numerical constant, then, with probability at least $1-\varepsilon$, noiseless recovery via (BP) or stable and robust recovery via (qBP) is ensured. The probability π minimizing the previous bound on m reads:

$$\forall k \in \{1,\dots,n\}, \quad \pi_k = \pi_k^{\tilde{\Lambda}} = \frac{\max\limits_{j \in S^c} \|d_{k,S\cup\{j\}}\|_2^2}{\sum\limits_{\ell=1}^n \max\limits_{j \in S^c} \|d_{\ell,S\cup\{j\}}\|_2^2}. \tag{4.12}$$

With such a choice, the required number m of measurements can be rewritten as follows:

$$m \geqslant c_1 \cdot \sum_{\ell=1}^n \max_{j \in S^c} \|d_{\ell,S\cup\{j\}}\|_2^2 \cdot \ln^2\left(\frac{6n}{\varepsilon}\right).$$

5. Applications and numerical experiments

In this section, we discuss the application of the optimal sampling strategies derived in Section 4 to three concrete case studies. Firstly, we consider the setting of isolated measurements from the one-dimensional Fourier-Haar transform and derive an explicit sampling probability that promotes the recovery of signals that are sparse in levels (Section 5.1). Secondly, the case of two-dimensional Fourier-Haar transform with block-structured measurements is discussed in Section 5.2, of considerable interest in MRI. In particular, we analyze the setting where the Fourier space is sampled using random vertical (or horizontal) lines and where the signal exhibits a particular type of sparsity in levels (see Definition 5.2). Thirdly, we illustrate how the proposed derivation of optimal sampling measures can be applied to devise adaptive sampling strategies for one-dimensional function approximation from pointwise data (Section 5.3).

5.1 *Performing isolated measurements with Fourier-Haar transform*

Let us consider the finite-dimensional setting (ii) described in Section 2.3 and let $n = 2^{J+1}$ for some $J \in \mathbb{N}$. Set A_0 to be the Fourier-Haar transform, i.e. $A_0 = \mathscr{F}H^* = (d_1|d_2|\dots|d_n)^*$ where $\mathscr{F} \in \mathbb{C}^{n\times n}$ is the Fourier matrix relative to frequencies $\{-n/2+1,\dots,n/2\}$ and $H^* \in \mathbb{R}^{n\times n}$ is the inverse Haar transform. The matrix A_0 could be naturally chosen to model the acquisition in MRI, where the measurements are performed in the Fourier domain, and where MR images are considered sparse in the wavelet domain. In this framework, we aim at reconstructing the wavelet coefficients x of the MR image.

First, we introduce the notion of sparsity in levels for the coefficient vector x. We define the levels $(\Omega_j)_{0\leqslant j\leqslant J}$ as the partition of $\{1,\dots,n\}$ corresponding to the subbands of the Haar transform, defined as $\Omega_0 := \{1,2\}$ and, for $j = 1,\dots,J$, as

$$\Omega_j := \{2^j+1,\dots,2^{j+1}\}.$$

The subband Ω_0 corresponds to the scaling function and to the Haar mother wavelet. The subsequent subbands Ω_j with $j \geqslant 1$ correspond to wavelet functions dyadically refined at level j (see [4] for further details). The vector x is assumed to be sparse in levels, i.e. denoting by S the support of x, we suppose

$$|S\cap\Omega_j| = s_j,$$

meaning that restricted to the j-th level Ω_j, x is s_j-sparse.

In order to devise effective sampling strategies, we partition the set of Fourier frequencies into frequency subbands $(W_j)_{0\leqslant j\leqslant J}$, defined as $W_0 := \{0,1\}$ and, for $j = 1,\dots,J$, as

$$W_j := \{-2^j+1,\dots,-2^{j-1}\}\cup\{2^{j-1}+1\dots,2^j\}.$$

Notice that for every $j = 0,\dots,J$ we have

$$|\Omega_j| = |W_j| = 2^{\max(1,j)}.$$

For the sake of simplicity, we will relabel the Fourier frequencies $\{-n/2+1,\dots,n/2\}$ as $\{1,\dots,n\}$. With this convention, we denote as $j : \{1,\dots,n\} \to \{0,\dots,J\}$ the function that maps a Fourier frequency to the corresponding frequency band, i.e. $j(k) = \ell$ if $k \in W_\ell$.

In this setting, local coherence upper bounds are explicitly computable. We recall the following estimate, corresponding to [4, Lemma 1]:

$$\mu_{j,\ell} := \max_{k\in W_j}\max_{i\in\Omega_\ell}|(\mathscr{F}H^*)_{ki}|^2 \leqslant C\cdot 2^{-j}2^{-|j-\ell|}, \tag{5.1}$$

where $C > 0$ is a universal constant.

APPLICATION OF THE THEORY Suppose that random isolated measurements are performed in the Fourier domain according to some probability distribution π (we note in passing that, although useful for illustrative purposes, sampling single points in the Fourier domain is not feasible in practice). Then, thanks to the local coherence upper bound (5.1), one can estimate the quantity $\Theta(S,\pi)$ as (see [7, Corollary 4.4])

$$\Theta(S,\pi) = C\max_{1\leqslant k\leqslant n}\frac{1}{\pi_k}2^{-j(k)}\sum_{p=0}^{J}2^{-|j(k)-p|/2}s_p. \tag{5.2}$$

As a consequence, the probability distribution minimizing $\Theta(S,\pi)$ in this setting can be written as follows (see Lemma E.1):

$$\forall k \in \{1,\dots,n\}, \quad \pi_k^{\Theta} = \frac{2^{-j(k)}\sum_{p=0}^{J} 2^{-|j(k)-p|/2} s_p}{\sum_{\ell=1}^{n} 2^{-j(\ell)}\sum_{p=0}^{J} 2^{-|j(\ell)-p|/2} s_p}. \tag{5.3}$$

Note that the drawing probability distribution is a function of the frequency level. Assuming that $\mathrm{sign}(x_S)$ is a Steinhaus or Rademacher sequence and employing Theorems 3.3 and 3.6, the required number of measurements needed to ensure exact or stable and robust recovery with probability $1-\varepsilon$ is

$$m \gtrsim \sum_{j=0}^{J} \left(s_j + \sum_{\substack{p=0 \\ p\neq j}}^{J} 2^{-|j-p|/2} s_p \right) \cdot \ln^2(6n/\varepsilon), \tag{5.4}$$

which matches results in [7, Corollary 4.4] and [2].

Replacing $\Theta(\pi,S)$ with $\Lambda(\pi,S)$ leads to an improvement of condition (5.4), at the price of an extra assumption on the sparsity in levels of the signal.

COROLLARY 5.1 Let $A_0 = FH^*$ be the one-dimensional Fourier-Haar transform and consider the splitting of the Haar and of the Fourier spaces into subbands $(\Omega_j)_{0\leqslant j\leqslant J}$ and $(W_j)_{0\leqslant j\leqslant J}$ as described above. Suppose that $x \in \mathbb{C}^n$ is an S-sparse vector with random signs and sparsities in levels $(s_j)_{0\leqslant j\leqslant J}$. Fix $\varepsilon \in (0,1)$. Then, condition (3.9) holds if

$$\min_{0\leqslant j\leqslant J} s_j + \sum_{j'\neq j} s_{j'} 2^{-|j-j'|} \gtrsim \ln(3n/\varepsilon). \tag{5.5}$$

In this case, $\Lambda(S,\pi)$ can be chosen as follows

$$\Lambda(S,\pi) = \max_{1\leqslant k\leqslant n} \frac{1}{\pi_k} 2^{-j(k)} \sum_{p=0}^{J} 2^{-|j(k)-p|} s_p. \tag{5.6}$$

Then, the probability minimizing $\Lambda(S,\pi)$ is

$$\forall k \in \{1,\dots,n\}, \quad \pi_k^{\Lambda} = \frac{2^{-j(k)}\sum_{p=0}^{J} 2^{-|j(k)-p|} s_p}{\sum_{\ell=1}^{n} 2^{-j(\ell)}\sum_{p=0}^{J} 2^{-|j(\ell)-p|} s_p}, \tag{5.7}$$

and the recovery guarantees of Theorem 3.10 hold with probability at least $1-\varepsilon$ if

$$m \gtrsim \sum_{j=0}^{J} \left(s_j + \sum_{\substack{p=0 \\ p\neq j}}^{J} 2^{-|j-p|} s_p \right) \cdot \ln(3n/\varepsilon). \tag{5.8}$$

The proof of this result is given in Appendix F.1. Condition (5.5) can be easily verified in practice. Note that by (5.8), we improved (5.4) (and, consequently, the state-of-the-art results in [7] and [2] when the signal is sparse with random signs) by (i) decreasing the interferences between different levels and (ii) removing a log factor.

NUMERICAL EXPERIMENTS In this part, we illustrate the improvement in recovery when the sampling strategy takes into account local coherences of the measurement vectors *and* the structured sparsity of the signal to reconstruct. With this aim, we generate 100 random signals with wavelet coefficients in dimension $n = 2048$ with $s/n \approx 6\%$ and having a sparsity in levels structure as in Figure 3(a) with 4 subbands in total. We compare three different sampling schemes where isolated samples are randomly drawn according to:

- the probability distribution minimizing the global coherence of each measurement vector: $\forall k \in \{1,\dots,n\}$,
$$\pi_k^{\infty} = \frac{\|d_k\|_\infty^2}{\sum_{p=1}^n \|d_p\|_\infty^2};$$
- the probability distribution minimizing $\Theta(S,\pi)$: $\forall k \in \{1,\dots,n\}$,
$$\pi_k^{\Theta} = \frac{\|d_k\|_\infty \sum_{j=0}^J s_j \|d_{k,\Omega_j}\|_\infty}{\sum_{p=1}^n \|d_p\|_\infty \sum_{j=0}^J s_j \|d_{p,\Omega_j}\|_\infty};$$
- the probability distribution minimizing $\Lambda(S,\pi)$: $\forall k \in \{1,\dots,n\}$,
$$\pi_k^{\Lambda} = \frac{\sum_{j=0}^J s_j \|d_{k,\Omega_j}\|_\infty^2}{\sum_{p=1}^n \sum_{j=0}^J s_j \|d_{p,\Omega_j}\|_\infty^2}.$$

These three sampling distributions can be computed in the case of the Fourier-Haar transform. This is illustrated in Figure 2 for the 1D structured-sparsity pattern described in Figure 3(a). One can clearly see in Figure 2 that a certain prior on the sparsity structure in this case inclines the sensing in the low frequency domain. Reconstruction results are shown in Figure 3(b). For the same amount of measurements, the sampling strategy minimizing the bound on $\Theta(S,\pi)$ proposed in this paper is outperforming the standard CS strategy taking only into account the global coherence. In this setting, sampling according to $\Lambda(S,\pi)$ improves the reconstruction quality compared to the $\Theta(S,\pi)$-based strategy.

5.2 *Sensing vertical (or horizontal) lines in MRI*

We extend the analysis of Section 5 from the one-dimensional to the two-dimensional framework, leading a more realistic model for the MRI problem.

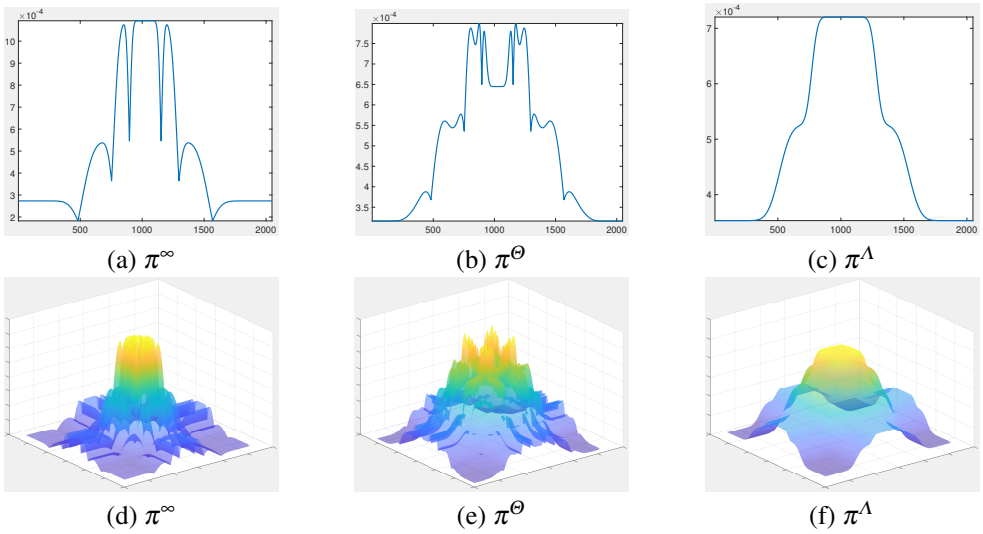
Let $n = 2^{2(J+1)}$ for some $J \in \mathbb{N}$ and denote the one-dimensional Fourier-Haar transform introduced in Section 5.1 as $\phi = \mathscr{F}H^* \in \mathbb{C}^{\sqrt{n}\times\sqrt{n}}$. Then, we define the two-dimensional Fourier-Haar transform as

$$A_0 = \phi \otimes \phi \in \mathbb{C}^{n\times n},$$

where $\otimes$ denotes the Kronecker product. The splitting of the wavelet multi-index space $\{1,\dots,\sqrt{n}\}^2$ and of the Fourier frequency space $\{-\sqrt{n}/2+1,\dots,\sqrt{n}/2\}^2$ into subbands is carried out by tensorizing the one-dimensional splitting presented in Section 5.1. Namely, we consider subbands defined, for $j,j' = 0,\dots,J$, as

$$\Omega_{j,j'} := \Omega_j \times \Omega_{j'} \quad \text{and} \quad W_{j,j'} := W_j \times W_{j'}.$$

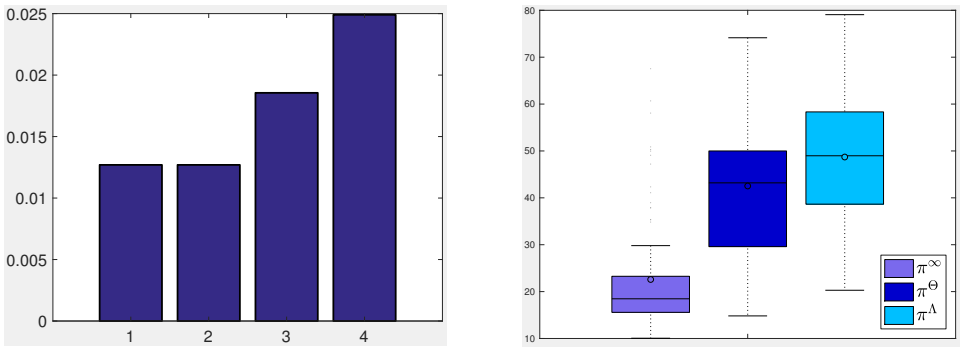
Moreover, we will represent the image wavelet coefficients as a vector $x \in \mathbb{C}^n$ or, equivalently, as a matrix $X = \mathbb{C}^{\sqrt{n}\times\sqrt{n}}$, such that $x = \mathrm{Vect}(X)$, where $\mathrm{Vect} : \mathbb{C}^{\sqrt{n}\times\sqrt{n}} \to \mathbb{C}^n$ is the vectorization operator



(a) π^{∞} (b) π^{Θ} (c) π^{Λ}

(d) π^{∞} (e) π^{Θ} (f) π^{Λ}

Figure 2. Comparison between sampling probability distributions chosen according to different strategies: in (a,b,c) for 1D signals with structured sparsity described in Figure 3(a), in (d,e,f) for 2D signals with the corresponding tensorized structured sparsity. In (a) and (d), the sampling probability distribution π^{∞} is optimized to minimize the global coherence, i.e. $\pi_k \propto \|a_k\|_{\infty}^2$; in (b) and (e), the sampling probability distribution π^{Θ} is optimized to minimize an upper bound to $\Theta(S,\pi)$, i.e. $\pi_k \propto \|a_k\|_{\infty} \sum_{j=0}^{J} s_j \|a_{k,\Omega_j}\|_{\infty}$; in (c) and (f), the sampling probability distribution π^{Λ} is optimized to minimize an upper bound to $\Lambda(S,\pi)$, i.e. $\pi_k \propto \sum_{j=0}^{J} \|a_k\|_{\infty} \|a_{k,\Omega_j}\|_{\infty}^2$; the probability distributions in (d), (e), and (f) are constructed accordingly in the 2D case.



(a) (b)

Figure 3. Boxplot of reconstruction PSNR in (b) of 100 random signals of length $n = 2048$ having a structured sparsity in the wavelet decomposition as in (a) ($s/n = 6\%$) from 25% measurements drawn according to π^{∞}, π^{Θ} and π^{Λ} as in Figure 2 (a,b,c). In (a), we represent the sparsity in levels structure of the randomly generated signals. The plot represents the percentage of nonzero coefficients for each subband.

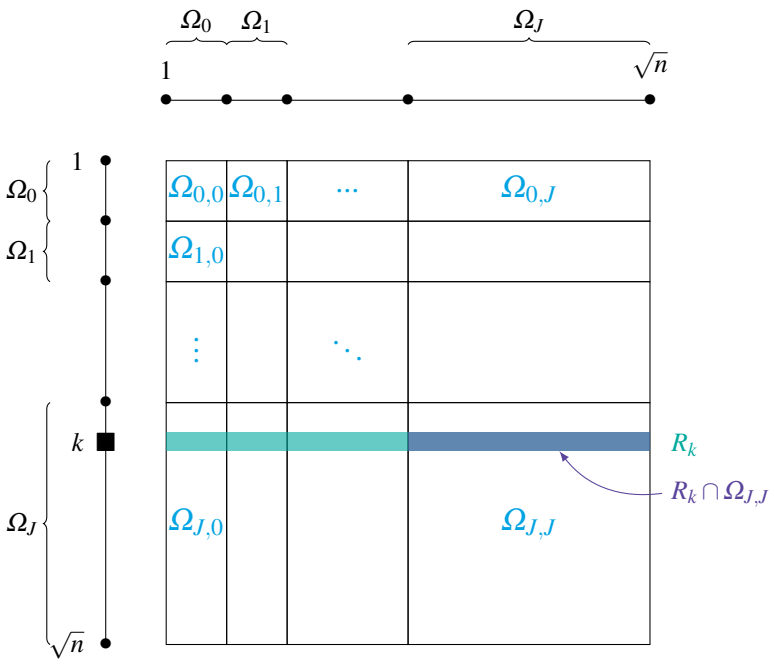


Figure 4. 2D view of the signal $x \in \mathbb{C}^n$ to reconstruct. The vector x can be reshaped into a $\sqrt{n} \times \sqrt{n}$ matrix. The indices can be partitioned according to the sets $\Omega_{j,j'} = \Omega_j \times \Omega_{j'}$ for $0 \leqslant j, j' \leqslant J$. R_k represents the coefficient indices corresponding to the k-th horizontal row.

that stacks the columns of a matrix atop, i.e., $\text{Vect}\,(x_1|\cdots|x_{\sqrt{n}}) = (x_1^*,\ldots,x_{\sqrt{n}}^*)^*$. Within this multi-level framework, we consider the following generalization of sparsity in levels.

DEFINITION 5.2 Let $x \in \mathbb{C}^n$ and $X \in \mathbb{C}^{\sqrt{n}\times\sqrt{n}}$ be such that $x = \text{Vect}\,(X)$. Then, given $S = \text{supp}(X)$, we define the following sparsities in levels for the vector x:

$$s_j^r := \max_{0\leqslant\ell\leqslant J} \max_{k\in\Omega_\ell} \left|S \cap \Omega_{\ell,j} \cap R_k\right|, \tag{5.9}$$

where $R_k := \{k\} \times \{1,\ldots,\sqrt{n}\}$ represents the set corresponding to the k-th horizontal line or row of the wavelet multi-index space $\{1,\ldots,\sqrt{n}\}^2$ (see Figure 4).

The sparsities in levels $(s_j^r)_{0\leqslant j\leqslant J}$ are intrinsically anisotropic. According to this definition, the wavelet coefficients restricted to the vertical subband $\{1,\ldots,\sqrt{n}\} \times \Omega_j$ and to a generic row can be at most s_j^r-sparse.

Equipped with this notion of structured sparsity, we consider measurements corresponding to vertical lines in the frequency space. This sensing strategy models MRI acquisition in a more realistic way than isolated Fourier measurements and corresponds to the setting (i) in Section 2.3. Specifically, we

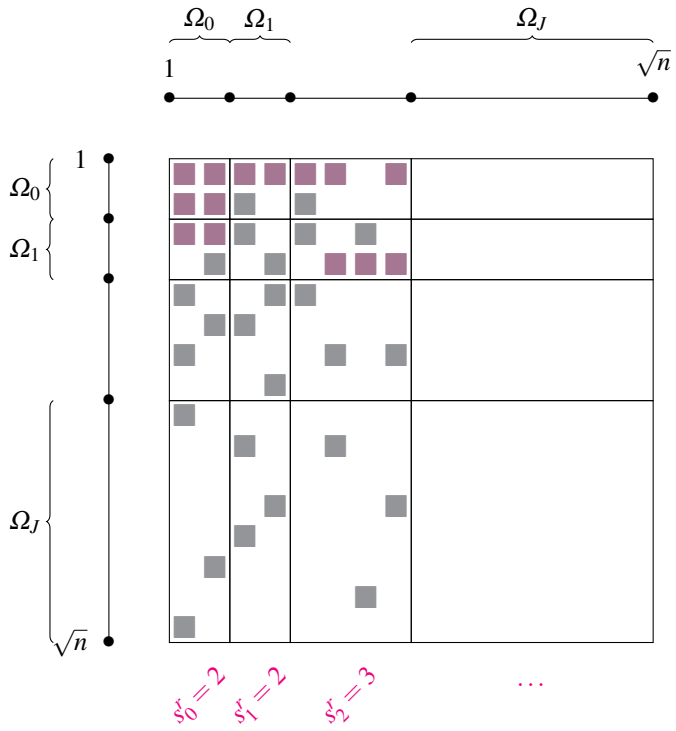


Figure 5. Example of a signal $x \in \mathbb{C}^n$ with support denoted by squares in grey for the first vertical subbands. The quantities $(s_j^r)_j$ are the maximal degrees of sparsity restricted to rows and respectively to vertical subbands $\cup_{j'} \Omega_{j',j}$. In the Figure, they are reached for support locations highlighted in red.

partition the set of n rows of the matrix A_0 into $\sqrt{n}$ blocks of size $\sqrt{n} \times n$ defined for $k = 1, \dots, \sqrt{n}$ as

$$D_k := \phi_{k,:} \otimes \phi = \left(\phi_{k,1}\phi | \cdots | \phi_{k,\sqrt{n}}\phi\right) \in \mathbb{C}^{\sqrt{n} \times n},$$

where $\phi_{k,:}$ denotes the k-th row of ϕ. Drawing the block D_k corresponds to sampling along the vertical line $\{1, \dots, \sqrt{n}\} \times \{k\}$ in the frequency space.

In this setting, the analysis based on the quantity $\Theta(S, \pi)$ leads to the following result.

COROLLARY 5.3 Let $A_0 \in \mathbb{C}^{n \times n}$ be the two-dimensional Fourier-Haar transform and consider the splitting of the Haar and of the Fourier spaces into subbands as defined above. Fix $\varepsilon \in (0,1)$. Suppose that $x \in \mathbb{C}^n$ is an S-sparse vector with random signs and associated structured sparsities in levels $(s_j^r)_{0 \leqslant j \leqslant J}$. Then, the same recovery guarantees of Theorem 3.6 hold with probability at least $1 - \varepsilon$ by drawing vertical lines according to the probability distribution $(\pi_k)_{1 \leqslant k \leqslant \sqrt{n}}$ defined as

$$\pi_k = \frac{2^{-j(k)} \sum_{j=0}^{J} 2^{-|j(k)-j|/2} s_j^r}{\sum_{\ell=1}^{\sqrt{n}} 2^{-j(\ell)} \sum_{j=0}^{J} 2^{-|j(\ell)-j|/2} s_j^r},$$

with a number m of drawn vertical lines of the order

$$m \gtrsim \sum_{j=0}^{J} \left(s_j^r + \sum_{\substack{j'=0 \\ j' \neq j}}^{J} 2^{-|j-j'|/2} s_{j'}^r \right) \cdot \ln^2\left(\frac{6n}{\varepsilon}\right). \tag{5.10}$$

In particular, the probability distribution $(\pi_k)_{1 \leqslant k \leqslant \sqrt{n}}$ is constant on each frequency subband W_j.

The proof of this results is given in Appendix F.3. Corollary 5.3 meets the results of [7, Corollary 4.10]: the bounds on the number of blocks of measurements are of the same order. Nevertheless, note that Corollary 5.3 requires an additional assumption on the sign randomness. Moreover, recall that Theorem 3.6 also ensures stable and robust recovery.

We can improve this result by analyzing the quantity $\Lambda(S,\pi)$.

COROLLARY 5.4 Let $A_0 \in \mathbb{C}^{n \times n}$ be the two-dimensional Fourier-Haar transform and consider the splitting of the Haar and of the Fourier spaces into subbands as defined above. Fix $\varepsilon \in (0,1)$. Suppose that $x \in \mathbb{C}^n$ is an S-sparse vector with random signs and associated structured sparsities $(s_j^r)_{0 \leqslant j \leqslant J}$. If

$$\min_{0 \leqslant j \leqslant J} \left(s_j^r + \sum_{\substack{j'=0 \\ j' \neq j}}^{J} 2^{-|j-j'|} s_{j'}^r \right) \cdot \log_2(\sqrt{n}) \gtrsim \ln(3n/\varepsilon),$$

then, the same recovery guarantees of Theorem 3.10 hold with probability at least $1-\varepsilon$ by drawing vertical lines according to the probability distribution $(\pi_k)_{1 \leqslant k \leqslant \sqrt{n}}$ defined as

$$\pi_k = \frac{2^{-j(k)} \sum_{j=0}^{J} 2^{-|j-j(k)|} s_j^r}{\sum_{\ell=1}^{\sqrt{n}} 2^{-j(\ell)} \sum_{j=0}^{J} 2^{-|j(\ell)-j|} s_j^r},$$

with a number m of drawn vertical lines of the order

$$m \gtrsim \left(\sum_{j=0}^{J} s_j^r + \sum_{\substack{j'=0 \\ j' \neq j}}^{J} 2^{-|j-j'|} s_{j'}^r \right) \cdot \log_2(\sqrt{n}) \cdot \ln\left(\frac{6n}{\varepsilon}\right). \tag{5.11}$$

In particular, the probability distribution $(\pi_k)_{1 \leqslant k \leqslant \sqrt{n}}$ is constant on each frequency subband W_j.

The proof of Corollary 5.4 is given in Appendix F.3.

Note that the bound (5.11) improves Corollary 5.3 by attenuating the interference between different subbands sparsities.

REMARK 5.5 Results analogous to those presented in this Section hold when sampling horizontal lines in the Fourier space, up to replacing the structured sparsities s_j^r in Definition 5.2 with

$$s_j^c := \max_{0 \leqslant \ell \leqslant J} \max_{k \in \Omega_\ell} |S \cap \Omega_{j,\ell} \cap C_k|,$$

where $C_k = \{1,\dots,\sqrt{n}\} \times \{k\}$ is the k-th vertical line of the wavelet multi-index space. In this case, sampling along the k-th horizontal line corresponds to draw a block $D_k = \phi \otimes \phi_{k,:}$. The proofs in Appendix F.3 can be easily adapted to this setting by interchanging the role of rows and columns in the Haar and in the Fourier multi-index spaces, respectively (or, equivalently, by switching the order of tensorization).

REMARK 5.6 There is a mismatch between the notation adopted here and in [7], where the role of rows and columns in the Haar and Fourier spaces is switched. As a result, horizontal Fourier sampling and structured sparsities s_j^c in [7] correspond to vertical Fourier sampling and structured sparsities s_j^r, respectively, in our framework. Although choosing one of these two conventions might be considered just a matter of taste, we have changed notation in order to adhere to the way the two-dimensional Fourier transform is performed by the MATLAB® command `fft2`. In our setting, fully-sampled measurements $y = A_0 x = (\mathscr{F}H^* \otimes \mathscr{F}H^*)x$ correspond to $Y = \mathscr{F}H^* X (\mathscr{F}H^*)^T$, where $x = \mathrm{Vect}(X)$ and $y = \mathrm{Vect}(Y)$. In particular, the block measurement $D_k x = ((\mathscr{F}H^*)_{k,:} \otimes \mathscr{F}H^*)x$ corresponds to the k-th column (or vertical line) of the matrix Y.

5.3 *Adaptive sampling for function approximation*

In this section, we examine the problem of approximating a function from random pointwise samples using standard sparsity with respect to Legendre polynomials [1, 12, 20]. In particular, we will see how the theoretical results shown in this paper can be employed to construct effective adaptive sampling strategies. For the sake of simplicity, we will focus on the one-dimensional case, although the strategies presented here can be generalized to multiple dimensions.

ADAPTIVE SAMPLING STRATEGIES Let $f : [-1,1] \to \mathbb{C}$ and let $\{L_j\}_{j=1}^n$ be the family of Legendre orthogonal polynomials normalized such that $\int_{-1}^1 L_j(x)L_k(x)dx = 2\delta_{jk}$. We aim at approximating f as a sparse expansion of Legendre polynomials from a fixed budget of m adaptively chosen pointwise samples.

In order to put ourselves in the framework of subsampled isometries, we consider the family of Gauss-Legendre quadrature points $\{g_j\}_{j=1}^n$ on $[-1,1]$ and their respective quadrature weights $\{w_j\}_{j=1}^n$ (we recall that $\{g_j\}_{j=1}^n$ are the roots of the polynomial L_{n+1} and the weights satisfy $\sum_{j=1}^n w_j = 2$). The resulting quadrature formula is exact on polynomials of degree less than or equal to $2n-1$. In particular, the matrix $A_0 \in \mathbb{R}^{n\times n}$, defined as

$$(A_0)_{ij} = (d_1|\cdots|d_n)^*, \quad d_i = \sqrt{\frac{w_i}{2}}(L_j(g_i))_{j=1}^n, \tag{5.12}$$

is orthogonal. Based on the theory presented in this paper, we consider two adaptive sampling strategies based on successive approximations of the support via ℓ^1 minimization. We refer to these strategies as (Adapt I) and (Adapt II). They are outlined in Algorithm 5.1 and described below.

Let us fix a target sparsity level $s \leqslant n$ and two numbers $K, m_1 \in \mathbb{N}$ such that $s \leqslant m_1 \leqslant N/K$. Both procedures draw m_1 samples K times, resulting in a total of $m = Km_1$ samples. At each iteration, we compute an approximation to the function f based on partial support information and we update the sampling measure accordingly. In particular, in the k-th iteration, (Adapt I) updates the sampling measure based on the support corresponding to the s entries of the $(k-1)$-th approximation with largest magnitude. On the other hand, (Adapt II) updates the measure by taking advantage of the $(k-1)s$ entries of the $(k-1)$-th approximation having largest magnitude. This difference corresponds to line 5 of Algorithm 5.1.

ADAPTIVE *vs.* NONADAPTIVE SAMPLING In order to evaluate the performance of adaptive sampling, we compare the following five sampling strategies:

(Adapt I) Adaptive sampling with fixed support size;

Algorithm 5.1 Adaptive sampling strategies (Adapt I) and (Adapt II)

Inputs: $s, n, K, m_1 \in \mathbb{N}$, $\eta > 0$

Output: $\hat{x} \in \mathbb{R}^n$

1: Draw m_1 samples $\{J_\ell^{(1)}\}_{\ell=1}^{m_1}$ i.i.d. uniformly from $\{1,\dots,N\}$ (i.e., $\pi_j^{(1)} = 1/N$)

2: Define $A^{(1)} = \sqrt{\frac{n}{m_1}}(d_{J_\ell^{(1)}})_{\ell=1}^{m_1} \in \mathbb{R}^{m_1 \times n}$ and $y^{(1)} = \sqrt{\frac{n}{m_1}}(\sqrt{\frac{w_\ell}{2}} f(g_{J_\ell^{(1)}}))_{\ell=1}^{m_1} \in \mathbb{R}^{m_1}$

3: Find $\hat{x}^{(1)} := \arg\min_{z \in \mathbb{R}^n} \|z\|_1$ s.t. $\|A^{(1)} z - y^{(1)}\|_2 \leqslant \eta$

4: **for** $k = 2, \dots, K$ **do**

5: Let $S^{(k-1)} = \begin{cases} \arg\min_{S:|S|=s} \|\hat{x}_S^{(k-1)} - \hat{x}^{(k-1)}\|_2, & \text{in the case of (Adapt I)} \\ \arg\min_{S:|S|=(k-1)s} \|\hat{x}_S^{(k-1)} - \hat{x}^{(k-1)}\|_2, & \text{in the case of (Adapt II)} \end{cases}$

6: Define the measure $\pi_j^{(k)} \propto \|d_{j,S^{(k-1)}}\|_2^2$ for every $j = 1,\dots,N$

7: Draw m_k indices $\{J_\ell^{(k)}\}_{\ell=1}^{m_1}$ i.i.d. according to $\pi^{(k)}$ from $\{1,\dots,N\}$

8: Define $A^{(k)} = (d_{J_\ell^{(k)}} c_\ell^{(k)})_{\ell=1}^{m_k}$ and $y^{(k)} = (\sqrt{\frac{w_\ell}{2}} f(g_{J_\ell^{(k)}}) c_\ell^{(k)})_{\ell=1}^{m_k}$, with $c_\ell^{(k)} = 1/\sqrt{m_1 \pi_{J_\ell^{(k)}}^{(k)}}$

9: Define $\tilde{A}^{(k)} = \frac{1}{\sqrt{k}} \begin{bmatrix} A^{(1)} \\ \vdots \\ A^{(k)} \end{bmatrix}$ and $\tilde{y}^{(k)} = \frac{1}{\sqrt{k}} \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(k)} \end{bmatrix}$

10: Find $\hat{x}^{(k)} := \arg\min_{z \in \mathbb{R}^n} \|z\|_1$ s.t. $\|\tilde{A}^{(k)} z - \tilde{y}^{(k)}\|_2 \leqslant \eta$

11: **end for**

12: Let $\hat{x} = \hat{x}^{(K)}$

(Adapt II) Adaptive sampling with increasing support size;

(Unif I) Random sampling from the continuous uniform measure over $(-1,1)$;

(Unif II) Uniform random sampling from A_0;

(Cheby) Random sampling from the continuous Chebyshev measure $\frac{1}{\pi\sqrt{1-t^2}}$ over $(-1,1)$;

We fix $n = 150$ and $s = 5$. For the strategies (Adapt I) and (Adapt II), we chose $m_1 = 15$, $K = 5$, leading to a total number of $m = Km_1 = 75$ adaptive measurements. We also fix the number of measurements as $m = 75$ for the nonadaptive strategies (Unif I), (Unif II), and (Cheby). Moreover, we fix $\eta = 0$ (using the `spg_bp` command of the SPGL1 MATLAB® package to solve basis pursuit [22, 23]). We run the following random experiment 100 times. For each run, we randomly generate a 5-sparse Legendre polynomial by selecting 5 indices uniformly at random form $\{1,\dots,n\}$ and by generating the respective coefficients independently at random according to the distribution $\mathcal{N}(0,1)$. Then, we apply the five strategies to recover the resulting function from pointwise samples.

The results of this comparison are shown in Figure 6, where we plot the ℓ^2 error of the approximate solution. Adaptive sampling slightly outperforms nonadaptive sampling, but the gain is not substantial.

We also visualize the evolution of the sampling measure $\pi^{(k)}$ as a function of the iteration k in Figure 7 for (Adapt I) and (Adapt II). We notice that the strategy (Adapt I) does not refine the measure from the third iteration. On the contrary, the measure is updated at each iteration in the case of (Adapt II), which takes advantage of more support information.

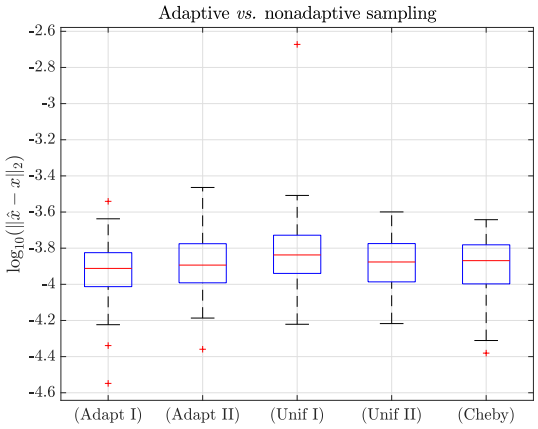


Figure 6. Approximation of f using adaptive strategies (Adapt I), (Adapt II) and nonadaptive strategies (Unif I), (Unif II), and (Cheby).

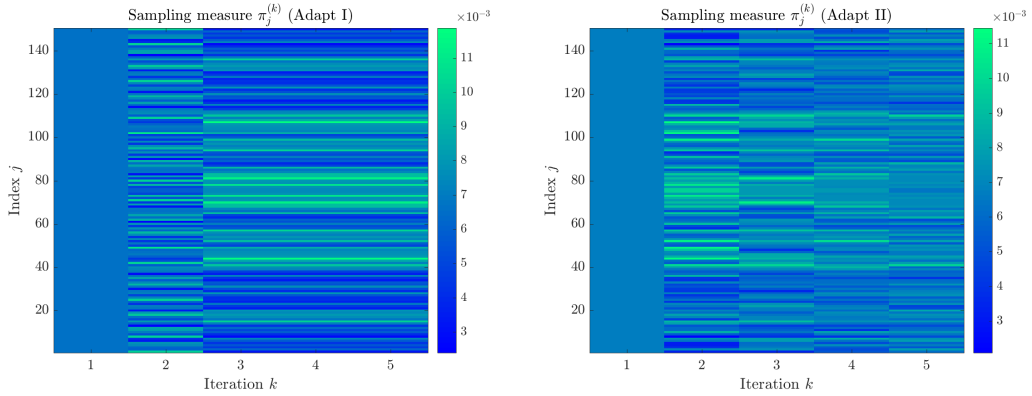


Figure 7. Evolution of the sampling measure $\pi^{(k)}$ as a function of the iteration k for adaptive strategies (Adapt I) and (Adapt II).

CONVERGENCE OF THE ADAPTED MEASURE An interesting feature of the proposed adaptive sampling schemes is its capacity to converge to the optimal sampling measure. In the next experiment, we consider a random 5-sparse combination of the first $n = 100$ Legendre polynomials. We compare the strategies (Adapt I) and (Adapt II) with $m_1 = 10$, $K = 5$, and $s = 5$. We start sampling from a uniform grid on $(-1,1)$ of $n^2 = 10000$ points (-1 and 1 are excluded). Moreover, we fix $\eta = 0$.

In Figure 8 we compare the measure $\pi^{(k)}$ for $k = 1,\dots,K$ with the Chebyshev measure π_{Cheb} on the uniform grid (i.e., such that $\pi_{\text{Cheb}}(t_j) \propto 1/\sqrt{1-t_j^2}$ for every point t_j of the uniform grid). We can see that for both (Adapt I) and (Adapt II) the adapted measure tends to approximate the Chebyshev measure. The convergence is more evident for the approach (Adapt II), where the approximate support gets larger and larger at each iteration.

A more quantitative convergence analysis is provided in Table 1, where we show the absolute error $\|\pi^{(k)} - \pi_{\text{Cheb}}\|_2$ as a function of the iteration $k = 1,\dots,K$ for both approaches. For (Adapt I), the error essentially stabilizes from the fourth iteration, whereas in the case of (Adapt II) it is monotonically

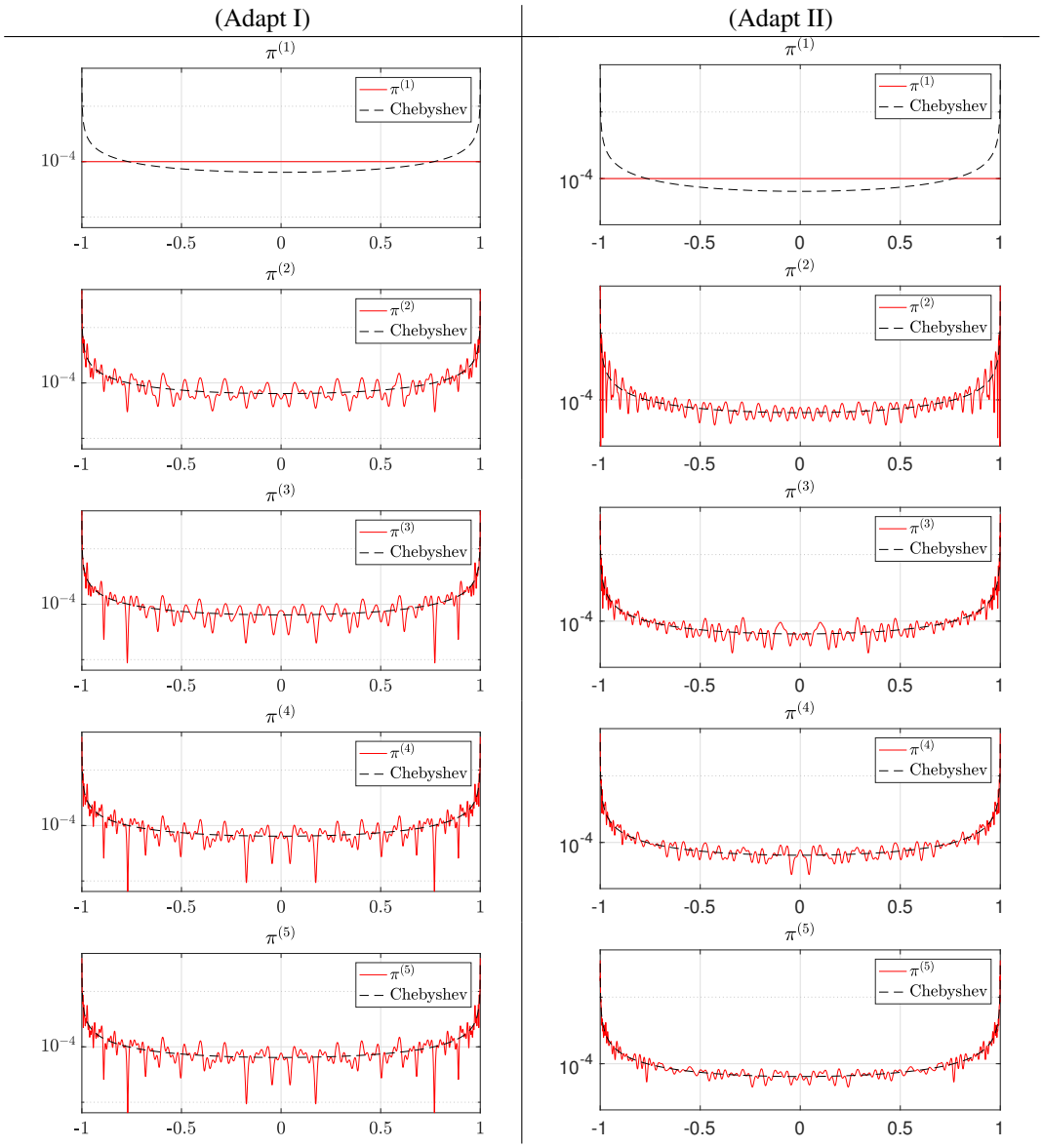


Figure 8. Convergence of the adapted measure $\pi^{(k)}$ to the Chebyshev measure for (Adapt I), on the left column, and for (Adapt II), on the right column.

decreasing. Comparing these data with Figure 8, we can see how (Adapt II) is able to correct the behavior of $\pi^{(k)}$ near the extrema ± 1 to a substantial extent at each iteration. On the contrary, in the case of (Adapt I) the measure $\pi^{(k)}$ exhibits severe oscillations near ± 1.

The convergence property of the proposed adaptive sampling strategies is particularly promising for high-dimensional approximation and, specifically, for the case of nontensorial domains, where the

	(Adapt I)	(Adapt II)
$\|\pi^{(1)} - \pi_{\text{Cheb}}\|_2$	1.0102e-02	1.0102e-02
$\|\pi^{(2)} - \pi_{\text{Cheb}}\|_2$	4.4442e-03	6.3474e-03
$\|\pi^{(3)} - \pi_{\text{Cheb}}\|_2$	4.2014e-03	3.3659e-03
$\|\pi^{(4)} - \pi_{\text{Cheb}}\|_2$	3.8730e-03	2.7679e-03
$\|\pi^{(5)} - \pi_{\text{Cheb}}\|_2$	3.8730e-03	2.3916e-03

Table 1. Distance between the adapted measure $\pi^{(k)}$ and the Chebyshev measure π_{Cheb} with respect to the ℓ^2 norm for (Adapt I) and (Adapt II).

optimal sampling measure is not known *a priori*. However, this is beyond the scope of this paper and is left for future investigation.

6. Conclusions

We have derived novel oracle-type inequalities for the number of measurements needed to guarantee stable and robust recovery for compressed sensing in the noisy setting. Our analysis relies on a random sign assumption for the signal to be recovered and encompasses the frameworks of block-structured and isolated random measurements, in particular subsampled from a finite-dimensional isometry.

The proposed analysis reveals a direct link between the number of measurements and the support of the signal to be recovered. This allows one to derive optimal sampling strategies in order to minimize the number of measurements and that are tailored to particular sparsity structure in the signal support.

We have derived optimal sampling strategies in the case of (i) subsampling one-dimensional Fourier-Haar transform via isolated measurements combined with the sparsity in levels structure and (ii) subsampling the two-dimensional Fourier-Haar transform via block-structured measurements combined with anisotropic (horizontal or vertical) sparsity in levels. Finally, we have shown how to perform adaptive sampling for one-dimensional function approximation from pointwise data.

All these results are based on a random sign assumption; for future work, it would be worth devising alternative proof techniques in order to remove it. The analysis of the extra assumptions (3.8) and (3.9) involving Λ and Γ could be also refined: are they necessary conditions to obtain oracle-type results? As for applications, the polynomial approximation example reveals an adaptive sampling strategy that is worth deepening and extending to the multivariate case as well. Besides, one could adapt the oracle-type study to the case of weighted ℓ^1 minimization which is of particular interest for polynomial approximation.

Acknowledgement

This work was supported by the Natural Sciences and Engineering Research Council of Canada [grant number 611675 to B.A. and S.B.]; and the Pacific Institute for the Mathematical Sciences (PIMS) ["PIMS Distinguished Visitor" program to C.B. and "PIMS Postdoctoral Training Centre in Stochastics" program to S.B].

The authors would like to thank Pierre Weiss for raising the optimal sampling concern and the anonymous referees for their insightful and constructive comments.

A. Proof of the main results

A.1 *Proof of Theorem 3.3*

In order to prove that x is the unique solution of (BP), one can use Fermat's rule for (BP). Under injectivity of A_S, this corresponds to the quest of a vector $v \in \mathscr{R}(A^*)$, called dual certificate, such that

$$\begin{cases} v_S & = \operatorname{sign}(x_S) \\ \|v_{S^c}\|_\infty & < 1. \end{cases}$$

A natural candidate for such a vector v is the dual certificate with minimal ℓ^2-norm, i.e.

$$v = A^*(A_S^\dagger)^* \operatorname{sign}(x_S) = A^* A_S (A_S^* A_S)^{-1} \operatorname{sign}(x_S),$$

which trivially lies in the range of A^* and satisfies $v_S = \operatorname{sign}(x_S)$. The second condition $\|v_{S^c}\|_\infty < 1$ remains to be satisfied, as required by assumption (ii) of the following proposition.

PROPOSITION A.1 ([14, Corollary 4.28]) For $x \in \mathbb{C}^n$ with support S, if

(i) A_S is injective,

(ii) $\left|\left\langle A_S^\dagger A e_\ell, \operatorname{sign}(x_S)\right\rangle\right| < 1$ for all $\ell \in S^c$,

then the vector x is the unique solution of (BP) with $y = Ax$.

By Lemma D.1, one gets the injectivity of A_S with high probability. The rest of the proof is then dedicated to ensure (ii) of Proposition A.1. By a union bound, one can control the probability that (BP) fails to recover x as follows:

$$\begin{aligned}
&\mathbb{P}(\text{failure of BP}) \\
&\leqslant \mathbb{P}\left(\max_{\ell\in S^c}\left|\left\langle A_S^\dagger A e_\ell, \operatorname{sign}(x_S)\right\rangle\right| \geqslant 1\right) \\
&\leqslant \mathbb{P}\left(\max_{\ell\in S^c}\left|\left\langle A_S^\dagger A e_\ell, \operatorname{sign}(x_S)\right\rangle\right| \geqslant 1 \,\middle|\, \max_{\ell\in S^c}\|A_S^\dagger A e_\ell\|_2 \leqslant \alpha\right) + \mathbb{P}\left(\max_{\ell\in S^c}\|A_S^\dagger A e_\ell\|_2 \geqslant \alpha\right) \\
&\leqslant \sum_{\ell\in S^c}\mathbb{P}\left(\left|\left\langle A_S^\dagger A e_\ell, \operatorname{sign}(x_S)\right\rangle\right| \geqslant \|A_S^\dagger A e_\ell\|_2 \alpha^{-1} \,\middle|\, \max_{\ell\in S^c}\|A_S^\dagger A e_\ell\|_2 \leqslant \alpha\right) + \mathbb{P}\left(\max_{\ell\in S^c}\|A_S^\dagger A e_\ell\|_2 \geqslant \alpha\right) \\
&\leqslant 2n\exp\left(-\frac{1}{2\alpha^2}\right) + \mathbb{P}\left(\max_{\ell\in S^c}\|A_S^\dagger A e_\ell\|_2 \geqslant \alpha\right),
\end{aligned}$$

where the last inequality is obtained using a Hoeffding-type bound (see [14, Corollary 7.21 and Corollary 8.10] for Rademacher and Steinhaus sequences, respectively). To control the second term, one can remark that for all $\ell \in S^c$

$$\|A_S^\dagger A e_\ell\|_2 = \|(A_S^* A_S)^{-1} A_S^* A e_\ell\|_2 \leqslant \|(A_S^* A_S)^{-1}\|_{2\to 2}\|A_S^* A e_\ell\|_2.$$

Using Lemma D.1, $\|(A_S^* A_S)^{-1}\|_{2\to 2}$ is bounded by $1/(1-\delta)$ for some $\delta > 0$ with high probability. Using Lemma D.2, $\max_{\ell\in S^c}\|A_S^* A e_\ell\|_2 \leqslant t$ with high probability. Then, we set

$$\alpha := t/(1-\delta).$$

The probability that (BP) fails is then bounded by

$$\begin{aligned}&\mathbb{P}\left(\max_{\ell\in S^c}\left|\left\langle A_S^\dagger Ae_\ell, \operatorname{sign}(x_S)\right\rangle\right| \geqslant 1\right)\\ &\leqslant \underbrace{2n\exp\left(-\frac{1}{2\alpha^2}\right)}_{P_1} + \underbrace{\mathbb{P}\left(\|A_S^*A_S - \mathrm{Id}\|_{2\to 2} \geqslant \delta\right)}_{P_2} + \underbrace{\mathbb{P}\left(\max_{\ell\in S^c}\|A_S^*Ae_\ell\|_2 \geqslant t\right)}_{P_3}.\end{aligned}$$

Note that $P_2 \leqslant \varepsilon/3$ by Lemma D.1 if

$$m \geqslant \frac{1+2\delta/3}{\delta^2/2}\cdot \Lambda(S,F)\cdot \ln\left(\frac{6s}{\varepsilon}\right). \tag{A.1}$$

As for the last term P_3, using Lemma D.2, choosing $t = \sqrt{\Theta/m} + t'$ for some $t' > 0$ to be fixed later, one has $P_3 \leqslant \varepsilon/3$ if

$$m \geqslant \frac{2}{(t')^2}\cdot \Theta(S,F)\cdot\left(3 + 2t'/3\right)\ln\left(\frac{3n}{\varepsilon}\right), \tag{A.2}$$

where we have used that $m \geqslant 4\Theta(S,F)$. Finally by setting $t' = \delta$, one has

$$\begin{aligned}P_1 \leqslant \varepsilon/3 &\iff 2n\exp\left(-\frac{1}{2\alpha^2}\right) \leqslant \varepsilon/3 \iff 2n\exp\left(-\frac{(1-\delta)^2}{2t^2}\right) \leqslant \varepsilon/3\\ &\iff 2n\exp\left(-\frac{(1-\delta)^2}{2\left(\sqrt{\Theta/m}+\delta\right)^2}\right) \leqslant \varepsilon/3 \iff \frac{(1-\delta)^2}{2\left(\sqrt{\Theta/m}+\delta\right)^2} \geqslant \ln\left(\frac{6n}{\varepsilon}\right).\end{aligned}$$

Assuming that $m \geqslant c\cdot\Theta(S,F)\ln(6n/\varepsilon)$ and choosing $\delta = \sqrt{\frac{1}{c\ln(6n/\varepsilon)}}$ with $c \geqslant 13$ leads to the previous inequality (notice that the constant 13 could be optimized in principle). Consequently, if

$$\delta = \frac{1}{\sqrt{13\ln\left(\frac{6n}{\varepsilon}\right)}} \qquad \text{and} \qquad m \geqslant 13\cdot\Theta(S,F)\cdot\ln\left(\frac{6n}{\varepsilon}\right),$$

then $P_1 \leqslant \varepsilon/3$.

Plugging this value of δ into (A.1) and (A.2), one gets the following conditions on the number of measurements:

$$\begin{cases} m & \geqslant 30\cdot\Lambda(S,F)\cdot\ln\left(\frac{6n}{\varepsilon}\right)\ln\left(\frac{6s}{\varepsilon}\right),\\ m & \geqslant 82\cdot\Theta(S,F)\cdot\ln^2\left(\frac{6n}{\varepsilon}\right).\end{cases}$$

Finally, the probability of recovery failure from (BP) is bounded by ε, if the total number of blocks of measurements satisfy

$$m \geqslant 82\cdot\Theta(S,F)\cdot\ln^2\left(\frac{6n}{\varepsilon}\right).$$

This concludes the proof of Theorem 3.3.

A.2 *Proof of Theorem 3.6*

In order to ensure stable and robust recovery via (qBP), we take advantage of a result analogous to Proposition A.1, which hinges again on the concept of dual certificate. (Notice that in [14, Theorem 4.33] S is assumed to be the set of s largest absolute entries of x, but an inspection of the proof reveals that S can be an arbitrary subset of $\{1,\dots,n\}$).

PROPOSITION A.2 ([14, Theorem 4.33]) Let $x \in \mathbb{C}^n$, $S \subseteq \{1,\dots,n\}$, and $y = Ax + \varepsilon$ with $\|\varepsilon\|_2 \leqslant \eta$. For $\delta, t, \gamma, \theta, \tau \geqslant 0$, with $\delta < 1$, assume that

$$\|A_S^* A_S - \mathrm{Id}\|_{2\to 2} \leqslant \delta, \quad \max_{\ell \in S^c} \|A_S^* A e_\ell\|_2 \leqslant t, \tag{A.3}$$

and that there exists a vector $u = A^* h \in \mathbb{C}^n$ with h such that

$$\|u_S - \mathrm{sign}(x_S)\|_2 \leqslant \gamma, \quad \|u_S\|_\infty \leqslant \theta, \quad \|h\|_2 \leqslant \tau\sqrt{s}.$$

If $\rho := \theta + t\gamma/(1-\delta) < 1$, then a minimizer $x^\sharp$ of (qBP) satistfies

$$\|x - x^\sharp\|_2 \leqslant C_1 \|x - x_S\|_1 + (C_2 + C_3\sqrt{s})\eta$$

for some constants $C_1, C_2, C_3 > 0$ depending only on $\delta, t, \gamma, \theta, \tau$.

Using arguments analogous to the proof of Theorem 3.3, one can show that (A.3) holds with high probability. Therefore, using Proposition A.2, we choose the dual certificate to be $u = A^* h$ with $h = A_S (A_S^* A_S)^{-1} \mathrm{sign}(x_S)$.

Since $u_S = \mathrm{sign}(x_S)$, for any $\gamma \geqslant 0$, the condition $\|u_S - \mathrm{sign}(x_S)\|_2 \leqslant \gamma$ is trivially satisfied (in particular, we can set $\gamma = 0$).

As for ensuring that $\|u_{S^c}\|_\infty \leqslant \theta$ with high probability, we have to slightly modify a part of the proof in the noiseless setting. First, let us observe that

$$\|u_{S^c}\|_\infty = \max_{\ell \in S^c} \left| \left\langle A^* A_S (A_S^* A_S)^{-1} \mathrm{sign}(x_S), e_\ell \right\rangle \right| = \max_{\ell \in S^c} \left| \left\langle \mathrm{sign}(x_S), A_S^\dagger A e_\ell \right\rangle \right|.$$

For $0 < \theta = 1/4 < 1$, using again a union bound and a Hoeffding-type inequality (see [14, Corollary 7.21 and Corollary 8.10] for Rademacher and Steinhaus sequences, respectively), one has

$$\begin{aligned}
&\mathbb{P}\left(\max_{\ell \in S^c} \left| \left\langle A_S^\dagger A e_\ell, \mathrm{sign}(x_S) \right\rangle \right| \geqslant \theta \right) \\
&\leqslant \mathbb{P}\left(\max_{\ell \in S^c} \left| \left\langle A_S^\dagger A e_\ell, \mathrm{sign}(x_S) \right\rangle \right| \geqslant \theta \,\middle|\, \max_{\ell \in S^c} \|A_S^\dagger A e_\ell\|_2 \leqslant \alpha \right) + \mathbb{P}\left(\max_{\ell \in S^c} \|A_S^\dagger A e_\ell\|_2 \geqslant \alpha \right) \\
&\leqslant \sum_{\ell \in S^c} \mathbb{P}\left(\left| \left\langle A_S^\dagger A e_\ell, \mathrm{sign}(x_S) \right\rangle \right| \geqslant \theta \|A_S^\dagger A e_\ell\|_2 \alpha^{-1} \,\middle|\, \max_{\ell \in S^c} \|A_S^\dagger A e_\ell\|_2 \leqslant \alpha \right) + \mathbb{P}\left(\max_{\ell \in S^c} \|A_S^\dagger A e_\ell\|_2 \geqslant \alpha \right) \\
&\leqslant 2n \exp\left(-\frac{\theta^2}{2\alpha^2} \right) + \mathbb{P}\left(\max_{\ell \in S^c} \|A_S^\dagger A e_\ell\|_2 \geqslant \alpha \right) \\
&\leqslant \underbrace{2n \exp\left(-\frac{\theta^2}{2\alpha^2} \right)}_{P_1} + \underbrace{\mathbb{P}\left(\|A_S^* A_S - \mathrm{Id}\|_{2\to 2} \geqslant \delta \right)}_{P_2} + \underbrace{\mathbb{P}\left(\max_{\ell \in S^c} \|A_S^* A e_\ell\|_2 \geqslant t \right)}_{P_3}.
\end{aligned}$$

Now, analogously to the proof of Theorem 3.3, we fix

$$\alpha = \frac{t}{1-\delta}, \quad t = \sqrt{\frac{\Theta}{m}} + t', \quad t' = \delta.$$

With these choices of parameters, conditions (A.1) and (A.2) suffice to guarantee that $P_2 \leqslant \varepsilon/3$ and $P_3 \leqslant \varepsilon/3$, respectively. Moreover, since $\theta = 1/4$, if

$$\delta = \frac{1}{\sqrt{146 \ln\left(\frac{6n}{\varepsilon}\right)}} \quad \text{and} \quad m \geqslant 146 \cdot \Theta(S,F) \cdot \ln\left(\frac{6n}{\varepsilon}\right),$$

then $P_1 \leqslant \varepsilon/3$. Plugging this value of δ in (A.1) and (A.2), one gets the following conditions on the number of measurements:

$$\begin{cases} m & \geqslant 305 \cdot \Lambda(S,F) \cdot \ln\left(\frac{6n}{\varepsilon}\right) \ln\left(\frac{6s}{\varepsilon}\right) \\ m & \geqslant 889 \cdot \Theta(S,F) \cdot \ln^2\left(\frac{6n}{\varepsilon}\right). \end{cases}$$

Now, we have to find a suitable constant $\tau > 0$ such that $\|h\|_2 \leqslant \tau\sqrt{s}$. Note that

$$\|h\|_2 = \|A_S(A_S^* A_S)^{-1}\mathrm{sign}(x_S)\|_2 \leqslant \|A_S(A_S^* A_S)^{-1}\|_{2\to 2}\|\mathrm{sign}(x_S)\|_2 = \frac{1}{\sigma_{\min}(A_S)}\sqrt{s},$$

since A_S has full column rank. Moreover, observe that

$$\|A_S^* A_S - \mathrm{Id}\|_{2\to 2} \leqslant \delta \Rightarrow \sigma_{\min}(A_S) \geqslant \sqrt{1-\delta}.$$

Hence, one can choose $\tau = \frac{1}{\sqrt{1-\delta}}$, so that

$$\mathbb{P}\left(\|h\|_2 \geqslant \sqrt{\frac{s}{1-\delta}}\right) \leqslant \mathbb{P}\left(\frac{1}{\sigma_{\min}(A_S)} \geqslant \sqrt{\frac{1}{1-\delta}}\right) \leqslant \mathbb{P}\left(\|A_S^* A_S - \mathrm{Id}\|_{2\to 2} \geqslant \delta\right),$$

which has been already controlled as $P_2 \leqslant \varepsilon/3$.

Finally, in order to apply Proposition A.2, let us notice that $\theta + \gamma t/(1-\delta) < 1$ since $\gamma = 0$ and $\theta = 1/4$.

This leads to the desired result.

A.3 *Proof of Theorem 3.8*

We will show a more general result, which will imply Theorem 3.8 as a corollary. In particular, in order to obtain Theorem 3.8 it will suffice to consider the partition S^c formed by its singletons in Theorem A.3, i.e. $S^c = \sqcup_{j\in S^c}\{j\}$.

THEOREM A.3 Let $x \in \mathbb{R}^n$ or $\mathbb{C}^n$ be a vector supported on S, such that $\mathrm{sign}(x_S)$ forms a Rademacher or Steinhaus sequence. Let A be the random sensing matrix defined in (2.3) with parameter $\Lambda(S,F)$ and let $y = Ax + \varepsilon$, with $\|\varepsilon\|_2 \leqslant \eta$. Then, there exist constants $c_1, C_1, C_2 > 0$ such that the following holds. For every $0 < \varepsilon < 1$, if

$$m \geqslant c_1 \min_{\substack{\{S_r\} \\ \text{partition of } S^c}} \max_r \Lambda(S \cup S_r, F) \ln\left(\frac{6|\{S_r\}|\max_r|S \cup S_r|}{\varepsilon}\right) \ln\left(\frac{6n}{\varepsilon}\right),$$

then, with probability at least $1-\varepsilon$, a minimizer $x^\sharp$ of (qBP) satisfies

$$\|x-x^\sharp\|_2 \leqslant (C_1+C_2\sqrt{s})\eta.$$

In particular, in the noiseless case (i.e., $\eta=0$), x is exactly recovered via (BP) with probability at least $1-\varepsilon$ and with constant $c_1=19$.

For the sake of simplicity, let us address the case $\eta=0$ (the argument can be generalized to the case $\eta>0$ in the same spirit of Appendix A.2). Analogously to the proof of Theorem 3.3 in Appendix A.1, our aim is to find sufficient conditions such that

$$\underbrace{2n\exp\left(-\frac{1}{2\alpha^2}\right)}_{P_1}+\underbrace{\mathbb{P}\left(\|A_S^*A_S-\mathrm{Id}\|_{2\to 2}\geqslant\delta\right)}_{P_2}+\underbrace{\mathbb{P}\left(\max_{\ell\in S^c}\|A_S^*Ae_\ell\|_2\geqslant t\right)}_{P_3}\leqslant\varepsilon,$$

where $\alpha=t/(1-\delta)$. In the following, we show how to ensure that $P_i\leqslant\varepsilon/3$, for $i=1,2,3$.

By using Lemma D.1, we have $P_2\leqslant\varepsilon/3$ under condition (A.1).

In order to control P_3, let us consider a partition $\{S_r\}_r$ of S^c and note that

$$\begin{aligned}\max_{\ell\in S^c}\|A_S^*Ae_\ell\|_2 &\leqslant \max_r\|A_S^*A_{S_r}\|_{2\to 2}=\max_r\|P_SA^*AP_{S_r}^*\|_{2\to 2}=\max_r\|P_S(A^*A-\mathrm{Id})P_{S_r}^*\|_{2\to 2}\\ &\leqslant \max_r\|P_{S\cup S_r}(A^*A-\mathrm{Id})P_{S\cup S_r}^*\|_{2\to 2}=\max_r\|A_{T_r}^*A_{T_r}-\mathrm{Id}\|_{2\to 2},\end{aligned}$$

where we have used that $P_SP_{S_r}^*=0$ and where we have defined $T_r=S\cup S_r$ for each index r of the partition. Now, combining the above inequality with Lemma D.1 and with a union bound over the elements S_r of the partition $\{S_r\}$, we obtain

$$P_3\leqslant\mathbb{P}\left(\max_r\|A_{T_r}^*A_{T_r}-\mathrm{Id}\|_{2\to 2}\geqslant t\right)\leqslant 2|\{S_r\}|\left(\max_r|T_r|\right)\exp\left(-\frac{mt^2/2}{\left(\max_r\Lambda(T_r,F)\right)(1+2t/3)}\right).$$

Notice that introducing the partition $\{S_r\}$ allowed us to control P_3 by the quantity $\max_r\Lambda(T_r,F)$ rather than $\Theta(S,F)$. Simple algebraic manipulations show that the condition

$$m\geqslant\frac{2}{t^2}\left(1+\frac{2t}{3}\right)\max_r\Lambda(T_r,F)\ln\left(\frac{6|\{S_r\}|\max_r|T_r|}{\varepsilon}\right),\tag{A.4}$$

is sufficient to have $P_3\leqslant\varepsilon/3$.

Now, by choosing

$$t=\delta\quad\text{and}\quad\delta=\frac{1}{\sqrt{8\ln(6n/\varepsilon)}},$$

it is not difficult to show that $P_1\leqslant\varepsilon/3$ is satisfied. Moreover, with this choice, condition (A.4) implies (A.1).

Finally, observing that

$$\frac{2}{\delta^2}\left(1+\frac{2\delta}{3}\right)\leqslant 19\ln\left(\frac{6n}{\varepsilon}\right)$$

and taking the minimum over all partitions $\{S_r\}$ of S^c in (A.4) concludes the proof.

A.4 *Proof of Theorem 3.10*

The proof is analogous to that of Theorem 3.3. Therefore, we will employ the same notation as in Appendix A.1. By similar arguments, one has

$$\mathbb{P}(\text{failure of BP}) \leqslant \underbrace{2n\exp\left(-\frac{1}{2\alpha^2}\right)}_{P_1} + \underbrace{\mathbb{P}\left(\|A_S^*A_S - \mathrm{Id}\|_{2\to 2} \geqslant \delta\right)}_{P_2} + \underbrace{\mathbb{P}\left(\max_{\ell\in S^c}\|A_S^*Ae_\ell\|_2 \geqslant t\right)}_{P_3},$$

for some $t, \delta > 0$ and $\alpha = t/(1-\delta)$. Note that P_2 is controlled as before using Lemma D.1. In particular, $P_2 \leqslant \varepsilon/3$ if

$$m \geqslant \frac{1+\delta/3}{\delta^2/2}\cdot\Lambda(S,F)\cdot\ln\left(\frac{6s}{\varepsilon}\right). \tag{A.5}$$

Thus, let us suppose for the rest of the proof to choose

$$m \geqslant c_2\cdot\Lambda(S,F)\cdot\ln\left(\frac{3n}{\varepsilon}\right), \tag{A.6}$$

where $c_2 \geqslant (1+\delta/3)/(\delta^2/2)$ is a constant that will be fixed later and where we are using $s \leqslant n/2$. The slight modification in the proof compared to the one in Appendix A.1 appears in the control of P_3. Indeed, using Lemma D.3, one has for some $t' > 0$

$$\mathbb{P}\left(\max_{i\in S^c}\|A_S^*Ae_i\|_2 \geqslant \sqrt{\frac{\Lambda}{m}} + t'\right) \leqslant n\exp\left(-\frac{m(t')^2/2}{\Gamma + 4\Lambda\sqrt{\Gamma/m} + 2t'\sqrt{\Lambda\Gamma}/3}\right).$$

Therefore, combining the above inequality with (A.6), fixing $t' = \frac{1}{2}\sqrt{\frac{\Lambda}{m}}$, and assuming that $\Lambda \geqslant c_1\Gamma\ln(3n/\varepsilon)$, one has

$$P_3 \leqslant n\exp\left(-\frac{\Lambda/8}{\Gamma + 13/3\Lambda\sqrt{\Gamma/m}}\right) \leqslant n\exp\left(-\frac{1/8}{1/c_1 + 13/(3\sqrt{c_1c_2})}\ln(3n/\varepsilon)\right) \leqslant \varepsilon/3,$$

provided that

$$\frac{1}{8} \geqslant \frac{1}{c_1} + \frac{13}{3\sqrt{c_1c_2}}. \tag{A.7}$$

Condition on $P_1 \leqslant \varepsilon/3$ remains to be checked. In accordance with the previous computation, we set $t = \frac{3}{2}\sqrt{\frac{\Lambda}{m}}$ and if we choose $\delta = 1/2$, with $\alpha = t/(1-\delta)$, we obtain

$$P_1 \leqslant \varepsilon/3 \qquad \Longleftarrow \qquad m \geqslant 18\cdot\Lambda(S,F)\cdot\ln\left(\frac{3n}{\varepsilon}\right).$$

Finally, we note that (A.7) holds true for, e.g., $c_1 = 50$ and $c_2 = 100$. Therefore, conditions

$$\begin{cases} \Lambda(S,F) \geqslant 50\cdot\Gamma(F)\cdot\ln(3n/\varepsilon) \\ m \geqslant 100\cdot\Lambda(S,F)\cdot\ln\left(\frac{3n}{\varepsilon}\right), \end{cases}$$

ensure that (BP) exactly recovers x with probability larger than $1-\varepsilon$.

Note that no effort was made in order to optimize the constants c_1 and c_2, and the result in the noisy setting can be easily deduced from the noiseless one as in Appendix A.2.

B. A discussion on stability

In Theorems 3.8, 3.9, and 3.10, we assume that the vector x is exactly s-sparse. However, it is possible to extend these results to the case of an arbitrary vector $x \in \mathbb{R}^n$ or $\mathbb{C}^n$ such that $\text{sign}(x_S)$ is a random Rademacher or Steinhaus sequence. Indeed, in this case, if $|S| = s$ and the measurements are corrupted by noise ε such that $\|\varepsilon\|_2 \leqslant \eta$, Proposition A.2 ensures a recovery guarantee of the form

$$\|x - x^\sharp\|_2 \leqslant (C_1 + C_2\sqrt{s})\eta + C_3\|x_{S^c}\|_1.$$

This implies the stability of the recovery guarantees with respect to the standard sparsity model, in addition to its robustness to bounded noise. We now clarify in what sense this generalization of Theorems 3.8, 3.9, and 3.10 would still be of "oracle type".

Let us go back to the proof of Proposition 3.1. We showed that the definition (3.4) of the oracle least-squares estimator $x^\star$ is sufficient to have

$$\|x^\star - x\|_2 \leqslant \frac{1}{\sigma_{\min}(A_S)}\|\varepsilon\|_2 + \|(A_S^*A_S)^{-1}\|_{2\to 2}\|A_S^*A_{S^c}\|_{1\to 2}\|x_{S^c}\|_1. \tag{A.1}$$

Moreover, in view of Lemma D.1 condition (3.3) implies $\sigma_{\min}(A_S) \geqslant \sqrt{1-\delta}$ and $\|(A_S^*A_S)^{-1}\|_{2\to 2} \leqslant 1/(1-\delta)$. In order to make (A.1) a stable and robust recovery guarantee, we need to control the quantity $\|A_S^*A_{S^c}\|_{1\to 2}$. To achieve this, we observe that $\|A_S^*A_{S^c}\|_{1\to 2} = \max_{j\in S^c}\|A_S^*Ae_j\|_2$. Therefore, the condition

$$\max_{j\in S^c}\|A_S^*Ae_j\|_2 \leqslant t, \tag{A.2}$$

combined with (A.1), implies the following stable and robust recovery guarantee for the oracle least-squares estimator:

$$\|x^\star - x\|_2 \leqslant \frac{1}{\sqrt{1-\delta}}\|\varepsilon\|_2 + \frac{t}{1-\delta}\|x_{S^c}\|_1.$$

We conclude by observing that (A.2) is one of the hypotheses of Proposition A.2 (see (A.3)) and that it holds with probability at least $1-\varepsilon/3$ under the assumptions on m stated in any of the Theorems 3.8, 3.9, or 3.10 (recall the definition of P_3 in the corresponding proofs).

C. Bernstein's inequalities

In this section, we present two Bernstein-type inequalities employed in the proofs of Appendix D.

First, we present an extension of the vector Bernstein inequality [14, Theorem 8.45] in order to handle random vectors Y_i that are independent but not necessarily identically distributed. Then, we recall a Bernstein inequality for self-adjoint matrices, corresponding to [14, Corollary 8.15]. Note that both results are independent of the dimension of the random objects (vectors or matrices) involved.

THEOREM C.1 (Vector Bernstein Inequality) Consider a set of independent random vectors $Y_1, \dots, Y_m$ such that

$$\mathbb{E}Y_i = 0, \quad \|Y_i\|_2 \leqslant K, \text{ a.s.}, \quad \forall i = 1, \dots, m,$$

and let $\sigma, \mu > 0$ such that

$$\sup_{\|x\|_2\leqslant 1}\sum_{i=1}^m \mathbb{E}|\langle x, Y_i\rangle|^2 \leqslant \sigma^2, \quad \mathbb{E}Z \leqslant \mu, \quad \text{where } Z := \left\|\sum_{i=1}^m Y_i\right\|_2.$$

Then, for every $t > 0$, the following holds:

$$\mathbb{P}\left(Z \geqslant \mu + t\right) \leqslant \exp\left(-\frac{t^2/2}{\sigma^2 + 2K\mu + tK/3}\right).$$

Proof. Assume $Y_i \in \mathbb{C}^n$, consider $B := \{x \in \mathbb{C}^n : \|x\|_2 \leqslant 1\}$, and let $\tilde{B}$ be a dense countable subset of B. Define $F_x(Y) := \Re\langle x, Y\rangle$. Then, Z can be expressed as a supremum of an empirical process. Indeed,

$$Z = \left\|\sum_{i=1}^m Y_i\right\|_2 = \sup_{x\in B} \Re\langle x, \sum_{i=1}^m Y_i\rangle = \sup_{x\in\tilde{B}} \sum_{i=1}^m F_x(Y_i) = \sup_{F\in\mathscr{F}} \sum_{i=1}^m F(Y_i),$$

where we have defined $\mathscr{F} := \{F_x : x \in \tilde{B}\}$ in the last step. Now, we verify the hypotheses needed to apply Talagrand's inequality (see [14, Theorem 8.42]).

First, $\mathbb{E}F_x(Y_i) = \mathbb{E}\Re\langle x, Y_i\rangle = 0$ because the Y_i's are centered. Moreover, for every $x \in \tilde{B}$, $F_x(Y_i) = \mathbb{E}\Re\langle x, Y_i\rangle \leqslant \|Y_i\|_2 \leqslant K$ almost surely and

$$\sum_{i=1}^m \mathbb{E}F_x^2(Y_i) = \sum_{i=1}^m \mathbb{E}(\Re\langle x, Y_i\rangle)^2 \leqslant \sum_{i=1}^m \mathbb{E}|\langle x, Y_i\rangle|^2 \leqslant \sigma^2.$$

Finally, Talagrand's inequality yields

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

which is the desired result. □

THEOREM C.2 (Bernstein Inequality for self-adjoint matrices, [14, Corollary 8.15]) Let $(Z_k)_{1\leqslant k\leqslant n}$ be a finite sequence of independent, random, self-adjoint matrices such that $\mathbb{E}Z_k = 0$ and that $\|Z_k\|_{2\to 2} \leqslant K$ a.s. for some constant $K > 0$ independent of k. Define

$$\sigma^2 = \left\|\sum_{k=1}^n \mathbb{E}Z_k^2\right\|_{2\to 2}$$

Then, for any $t > 0$, we have that

$$\mathbb{P}\left(\left\|\sum_{k=1}^n Z_k\right\|_{2\to 2} \geqslant t\right) \leqslant 2d\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).$$

D. Auxiliary lemmas

In this section, we show some deviation inequalities involving submatrices of the sensing matrix A corresponding to the setting described in Section 2.3. In particular, the tail probabilities will be controlled by using the quantities $\Lambda = \Lambda(S,F)$, $\Theta = \Theta(S,F)$, and $\Gamma = \Gamma(F)$ introduced in Definition 2.3, where $S \subseteq \{1,\ldots,n\}$.

The first auxiliary lemma controls the deviation of $A_S^* A_S$ from the identity by means of the quantity $\Lambda(S,F)$.

LEMMA D.1 For every $S \subseteq \{1,\dots,n\}$ with $|S| = s$ and for every $\delta > 0$, the following holds

$$\mathbb{P}(\|A_S^* A_S - \mathrm{Id}\|_{2\to 2} \geqslant \delta) \leqslant 2s \exp\left(-\frac{m\delta^2/2}{\Lambda(1+2\delta/3)}\right).$$

Proof. Our objective is to apply Theorem C.2. Consider the splitting

$$A_S^* A_S - \mathrm{Id} = \sum_{i=1}^m \underbrace{\frac{1}{m}(B_{i,S}^* B_{i,S} - \mathbb{E}[B_{i,S}^* B_{i,S}])}_{=:X_i} = \sum_{i=1}^m X_i \in \mathbb{C}^{s\times s}.$$

Of course, $\mathbb{E}X_i = 0$. Moreover,

$$\|X_i\|_{2\to 2} = \frac{1}{m} \sup_{\|x\|_2 \leqslant 1} |\langle x, X_i x\rangle| = \frac{1}{m} \sup_{\|x\|_2\leqslant 1} \left| \|B_{i,S}x\|_2^2 - \mathbb{E}\|B_{i,S}x\|_2^2 \right|.$$

Now, observing that $\|B_{i,S}x\|_2^2 \leqslant \Lambda \|x\|_2^2$, we have

$$\|X_i\|_{2\to 2} \leqslant 2\Lambda/m =: K.$$

Being $\sum_{i=1}^m \mathbb{E}X_i^2$ self-adjoint, we have

$$\sigma^2 := \left\| \sum_{i=1}^m \mathbb{E}X_i^2 \right\|_{2\to 2} = \sup_{\|x\|_2\leqslant 1} \sum_{i=1}^m \langle x, \mathbb{E}X_i^2 x\rangle.$$

In order to estimate this term, we notice that

$$\mathbb{E}X_i^2 = \frac{1}{m^2}\mathbb{E}(B_{i,S}^* B_{i,S} - \mathbb{E}[B_{i,S}^* B_{i,S}])^2 = \frac{1}{m^2}\left[\mathbb{E}(B_{i,S}^* B_{i,S})^2 - (\mathbb{E}[B_{i,S}^* B_{i,S}])^2\right].$$

Now, observing that $\langle x, (\mathbb{E}[B_{i,S}^* B_{i,S}])^2 x\rangle = \|\mathbb{E}[B_{i,S}^* B_{i,S}]x\|_2^2 \geqslant 0$ for every x, we have

$$\begin{aligned}\sum_{i=1}^m \langle x, \mathbb{E}X_i^2 x\rangle &\leqslant \frac{1}{m^2}\sum_{i=1}^m \langle x, \mathbb{E}(B_{i,S}^* B_{i,S})^2 x\rangle = \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}\|B_{i,S}^* B_{i,S} x\|_2^2 \\ &\leqslant \frac{1}{m^2}\Lambda \sum_{i=1}^m \mathbb{E}\|B_{i,S}x\|_2^2 = \frac{\Lambda}{m^2}\|x\|_2^2\end{aligned}$$

Therefore, $\sigma^2 \leqslant \Lambda/m^2$. Now, we apply Theorem C.2 and we obtain

$$\begin{aligned}\mathbb{P}(\|A_S^* A_S - I\|_{2\to 2} \leqslant \delta) &\leqslant 2s\exp\left(-\frac{\delta^2/2}{\sigma^2 + K\delta/3}\right) \leqslant 2s\exp\left(-\frac{m\delta^2/2}{\Lambda(1/m+2\delta/3)}\right), \\ &\leqslant 2s\exp\left(-\frac{m\delta^2/2}{\Lambda(1+2\delta/3)}\right),\end{aligned}$$

since $m \geqslant 1$, which concludes the proof. □

The next two lemmas aim at controlling the growth of the absolute entries of $A_S^* A_{S^c}$ in terms of the quantity $\Theta(S,F)$.

LEMMA D.2 Let $S \subseteq \{1,\dots,n\}$. Then, for every $t > 0$

$$\mathbb{P}\left(\max_{i\in S^c} \|A_S^* A e_i\|_2 \geqslant \sqrt{\Theta/m}+t\right) \leqslant n\exp\left(-\frac{mt^2/2}{\Theta(1+4\sqrt{\Theta/m}+2t/3)}\right).$$

Proof.

This proof is relies on Theorem C.1. Fix $i \in S^c$ and consider $\|A_S^* A e_i\|_2$. Note that $\mathbb{E}A_S^* A e_i = 0$ since $i \in S^c$. Moreover,

$$\|A_S^* A e_i\|_2 = \left\| \sum_{j=1}^m \underbrace{\frac{1}{m}(B_{j,S}^* B_j e_i - \mathbb{E}(B_{j,S}^* B_j e_i))}_{=:Y_j} \right\|_2 = \left\| \sum_{j=1}^m Y_j \right\|_2 =: Z.$$

Now, we verify the hypotheses of Theorem C.1 for the random vectors $Y_1,\dots,Y_m$. Of course, $\mathbb{E}Y_j = 0$. Moreover, recalling the definition of Θ and of $\|\cdot\|_{\infty\to\infty}$, we have

$$\|B_{j,S}^* B_j e_i\|_2 \leqslant \|B_{j,S}^* B_j e_i\|_1 \leqslant \Theta.$$

Therefore, we obtain

$$\|Y_j\|_2 = \frac{1}{m}\|B_{j,S}^* B_j e_i - \mathbb{E}(B_{j,S}^* B_j e_i)\|_2 \leqslant 2\Theta/m =: K.$$

Then, using the Cauchy-Schwarz inequality, we estimate

$$\begin{aligned}\sup_{\|x\|_2\leqslant 1} \sum_{j=1}^m \mathbb{E}|\langle x, Y_j\rangle|^2 &= \frac{1}{m^2} \sup_{\|x\|_2\leqslant 1} \sum_{j=1}^m \mathbb{E}|\langle x, B_{j,S}^* B_j e_i\rangle|^2 \leqslant \frac{1}{m^2} \sup_{\|x\|_2\leqslant 1} \sum_{j=1}^m \mathbb{E}[\|B_{j,S}\|_{2\to 2}^2 \|x\|_2^2 \|B_j e_i\|_2^2] \\ &\leqslant \frac{\Lambda}{m^2} \sum_{j=1}^m \mathbb{E}\|B_j e_i\|_2^2 = \frac{\Lambda}{m}\|e_i\|_2^2 = \frac{\Lambda}{m} \leqslant \frac{\Theta}{m} =: \sigma^2,\end{aligned}$$

where in the last step we used $\frac{1}{m}\sum_{j=1}^m \mathbb{E}(B_j^* B_j) = \mathrm{Id}$. Furthermore, using the independence of the Y_i's and their zero-mean property, we see that

$$(\mathbb{E}Z)^2 \leqslant \mathbb{E}Z^2 = \mathbb{E}\left\|\sum_{j=1}^m Y_j\right\|_2^2 = \mathbb{E}\sum_{j=1}^m \sum_{k=1}^m \langle Y_j, Y_k\rangle = \sum_{j=1}^m \mathbb{E}\|Y_j\|^2 + \underbrace{\sum_{j=1}^m \sum_{k\neq j} \langle \mathbb{E}Y_j, \mathbb{E}Y_k\rangle}_{=0}.$$

Now, combining the above inequality with the following estimate

$$\begin{aligned}\sum_{j=1}^m \mathbb{E}\|Y_j\|_2^2 &= \frac{1}{m^2}\sum_{j=1}^m \mathbb{E}\|B_{j,S}^* B_j e_i - \mathbb{E}(B_{j,S}^* B_j e_i)\|_2^2 = \frac{1}{m^2}\sum_{j=1}^m (\mathbb{E}\|B_{j,S}^* B_j e_i\|_2^2 - \|\mathbb{E}(B_{j,S}^* B_j e_i)\|_2^2) \\ &\leqslant \frac{1}{m^2}\sum_{j=1}^m \mathbb{E}\|B_{j,S}^* B_j e_i\|_2^2 \leqslant \frac{\Lambda}{m^2}\sum_{j=1}^m \mathbb{E}\|B_j e_i\|_2^2 = \frac{\Lambda}{m} \leqslant \frac{\Theta}{m}.\end{aligned}$$

we obtain

$$\mathbb{E}Z \leqslant \sqrt{\Theta/m} =: \mu.$$

We are now in a position to apply Theorem C.1. We have

$$\mathbb{P}\left(\|A_S^*Ae_i\|_2 \geqslant \sqrt{\Theta/m}+t\right) \leqslant \exp\left(-\frac{mt^2/2}{\Theta(1+4\sqrt{\Theta/m}+2t/3)}\right).$$

A union bound over $i \in S^c$ concludes the proof. □

The next lemma is a refined version of Lemma D.2, involving the quantities $\Lambda(S,F)$ and $\Gamma(F)$ instead of $\Theta(S,F)$. Indeed, it is worth recalling that $\Theta(S,F) \geqslant \Lambda(S,F)$.

LEMMA D.3 Let $S \subseteq \{1,\dots,n\}$. Then, for every $t > 0$

$$\mathbb{P}\left(\max_{i\in S^c}\|A_S^*Ae_i\|_2 \geqslant \sqrt{\Lambda/m}+t\right) \leqslant n\exp\left(-\frac{mt^2/2}{\Gamma+4\Lambda\sqrt{\Gamma/m}+2t\sqrt{\Lambda\Gamma}/3}\right).$$

Proof. The proof relies on the Bernstein-type inequality in Theorem C.1. Fix $i \in S^c$ and consider $\|A_S^*Ae_i\|_2$. Note that $\mathbb{E}A_S^*Ae_i = 0$ since $i \in S^c$. Moreover,

$$\|A_S^*Ae_i\|_2 = \left\|\sum_{j=1}^m \underbrace{\frac{1}{m}(B_{j,S}^*B_je_i - \mathbb{E}(B_{j,S}^*B_je_i))}_{=:Y_j}\right\|_2 = \left\|\sum_{j=1}^m Y_j\right\|_2 =: Z.$$

Now, we verify the hypotheses of Theorem C.1 for the random vectors $Y_1,\dots,Y_m$. Of course, $\mathbb{E}Y_j = 0$. Moreover, since

$$\|B_{j,S}^*B_je_i\|_2 \leqslant \sqrt{\Lambda}\|B_je_i\|_2 \leqslant \sqrt{\Lambda\Gamma},$$

we obtain

$$\|Y_j\|_2 = \frac{1}{m}\|B_{j,S}^*B_je_i - \mathbb{E}(B_{j,S}^*B_je_i)\|_2 \leqslant 2\sqrt{\Lambda\Gamma}/m =: K.$$

Then, using the Cauchy-Schwarz inequality, we estimate

$$\begin{aligned}\sup_{\|x\|_2\leqslant 1}\sum_{j=1}^m \mathbb{E}|\langle x,Y_j\rangle|^2 &= \frac{1}{m^2}\sup_{\|x\|_2\leqslant 1}\sum_{j=1}^m \mathbb{E}|\langle x,B_{j,S}^*B_je_i\rangle|^2 \leqslant \frac{1}{m^2}\sup_{\|x\|_2\leqslant 1}\sum_{j=1}^m \mathbb{E}[\|B_{j,S}x\|_2^2\|B_je_i\|_2^2]\\ &\leqslant \frac{\Gamma}{m^2}\sup_{\|x\|_2\leqslant 1}\sum_{j=1}^m \mathbb{E}\|B_{j,S}x\|_2^2 = \frac{\Gamma}{m}\sup_{\|x\|_2\leqslant 1}\|x_S\|_2^2 = \frac{\Gamma}{m} =: \sigma^2,\end{aligned}$$

where in the last step we used $\frac{1}{m}\sum_{j=1}^m \mathbb{E}(B_j^*B_j) = \mathrm{Id}$. Furthermore, using the independence of the Y_i's, we see that

$$(\mathbb{E}Z)^2 \leqslant \mathbb{E}Z^2 = \mathbb{E}\left\|\sum_{j=1}^m Y_j\right\|_2^2 = \mathbb{E}\sum_{j=1}^m\sum_{k=1}^m\langle Y_j,Y_k\rangle = \sum_{j=1}^m \mathbb{E}\|Y_j\|^2 + \underbrace{\sum_{j=1}^m\sum_{k\neq j}\langle \mathbb{E}Y_j,\mathbb{E}Y_k\rangle}_{=0}.$$

Now, combining the above inequality with the following estimate

$$\begin{aligned}\sum_{j=1}^m \mathbb{E}\|Y_j\|_2^2 &= \frac{1}{m^2}\sum_{j=1}^m \mathbb{E}\|B_{j,S}^*B_je_i - \mathbb{E}(B_{j,S}^*B_je_i)\|_2^2 = \frac{1}{m^2}\sum_{j=1}^m(\mathbb{E}\|B_{j,S}^*B_je_i\|_2^2 - \|\mathbb{E}(B_{j,S}^*B_je_i)\|_2^2)\\ &\leqslant \frac{1}{m^2}\sum_{j=1}^m \mathbb{E}\|B_{j,S}^*B_je_i\|_2^2 \leqslant \frac{\Lambda}{m^2}\sum_{j=1}^m \mathbb{E}\|B_je_i\|_2^2 = \frac{\Lambda}{m},\end{aligned}$$

we obtain

$$\mathbb{E}Z \leqslant \sqrt{\Lambda/m} =: \mu.$$

We are now in a position to apply Theorem C.1. We have

$$\mathbb{P}\left(\|A_S^* A e_i\|_2 \geqslant \sqrt{\Lambda/m} + t\right) \leqslant \exp\left(-\frac{mt^2/2}{\Gamma + 4\Lambda\sqrt{\Gamma/m} + 2t\sqrt{\Lambda\Gamma}/3}\right).$$

A union bound over $i \in S^c$ concludes the proof. □

E. Proofs of Section 4

Propositions in Section 4 are direct consequences of the minimization of the bounds obtained on the number of measurements in Theorems of Section 3, given the following lemma.

LEMMA E.1 Given $(\gamma_k)_{1\leqslant k\leqslant p}$ such that for all $1 \leqslant k \leqslant p, \gamma_k \geqslant 0$. Define for π a probability distribution on $\{1,\dots,p\}$, the following function:

$$K(\pi) := \max_{1\leqslant k\leqslant p} \frac{\gamma_k}{\pi_k}.$$

Then, it holds that

$$\min_{\substack{\pi_k\geqslant 0 \\ \sum_k \pi_k = 1}} K(\pi) = \sum_{j=1}^{p} \gamma_j,$$

and the unique minimizer $\pi^\star$ of K over discrete probability distributions is for all $1 \leqslant k \leqslant p$

$$\pi_k^\star = \frac{\gamma_k}{\sum_{j=1}^p \gamma_j}.$$

Proof. The proof is inspired by an intermediary result in [11]. We give the proof here for completeness. Note that for $\pi^\star$, $K(\pi^\star) = \sum_{j=1}^p \gamma_j$. For a probability distribution $q \neq \pi^\star$, there exists an index j_0 such that $q_{j_0} < \pi^\star_{j_0}$ since both sum to 1. Then, $K(q) \geqslant \gamma_{j_0}/q_{j_0} > \gamma_{j_0}/\pi^\star_{j_0} = \sum_{j=1}^p \gamma_j = K(\pi^\star)$. □

Therefore, the proofs of Propositions 4.1, 4.3, 4.2, 4.4 easily follow Lemma E.1 combined respectively with Theorems 3.6, 3.10.

F. Proofs of Section 5

The appendix contains the proofs of the results stated in Section 5, regarding the application of the proposed analysis to the case of one-dimensional and two-dimensional Fourier-Haar setting.

F.1 *Proof of Corollary 5.1*

Let $A_0 = \mathscr{F}H^* = (d_1|d_2|\dots|d_n)^* \in \mathbb{C}^{n\times n}$ with $n = 2^{J+1}$ be the Fourier-Haar transform and consider wavelet and frequency subbands $(\Omega_j)_{0\leqslant j\leqslant J}$ and $(W_j)_{0\leqslant j\leqslant J}$ as illustrated in Section 5.1.

Recalling the local coherence estimate (5.1) from [4, Lemma 1], one has for $k \in \{1,\dots,n\}$

$$\|d_{k,S}\|_2^2 = \sum_{j=0}^{J} \|P_S P_{\Omega_j} d_k\|_2^2 \leqslant \sum_{j=0}^{J} s_j \|P_{\Omega_j} d_k\|_\infty^2 \leqslant C\cdot 2^{-j(k)} \sum_{j=0}^{J} s_j 2^{-|j(k)-j|},$$

where $j(k)$ is the frequency level corresponding to index k and where $C > 0$ is a universal constant. Consequently, an estimate for $\Lambda(S,\pi)$ is

$$\Lambda(S,\pi) := \max_{1\leqslant k\leqslant n} \frac{C\cdot 2^{-j(k)}\sum_{j=0}^{J} s_j 2^{-|j(k)-j|}}{\pi_k}.$$

Recalling Lemma E.1, the probability distribution π^{Λ} minimizing the previous bound is such that for all $k\in\{1,\dots,n\}$

$$\pi_k^{\Lambda} = \frac{2^{-j(k)}\sum_{j=0}^{J} s_j 2^{-|j(k)-j|}}{\sum_{p=1}^{n} 2^{-j(p)}\sum_{j=0}^{J} s_j 2^{-|j(p)-j|}}.$$

Since the size of the frequency subbands is $|W_j| = 2^{\max(j,1)}$, one can rewrite, for all $k\in\{1,\dots,n\}$

$$\pi_k^{\Lambda} = \frac{2^{-j(k)}\sum_{j=0}^{J} s_j 2^{-|j(k)-j|}}{\sum_{j'=0}^{J} 2^{\max(j',1)-j'}\sum_{j=0}^{J} s_j 2^{-|j'-j|}}.$$

Therefore, since $2^j \leqslant 2^{\max(j,1)} \leqslant 2^{j+1}$, we have

$$\Lambda(S,\pi^{\Lambda}) = \sum_{j'=0}^{J} 2^{\max(j',1)-j'}\sum_{j=0}^{J} s_j 2^{-|j'-j|} = D\cdot\sum_{j=0}^{J}\left(s_j + \sum_{\substack{j'=0\\ j'\neq j}}^{J} s_{j'} 2^{-|j-j'|}\right),$$

with $1\leqslant D\leqslant 2$. With such a choice of probability distribution π^{Λ}, an estimate for Γ can be derived as follows:

$$\Gamma(\pi^{\Lambda}) := \max_{1\leqslant k\leqslant n}\frac{C\cdot 2^{-j(k)}}{\pi_k^{\Lambda}} = \max_{1\leqslant k\leqslant n}\frac{C\cdot 2^{-j(k)}\Lambda(S,\pi^{\Lambda})}{2^{-j(k)}\sum_{j'=0}^{J} s_{j'}2^{-|j(k)-j'|}} = \frac{C\cdot\Lambda(S,\pi^{\Lambda})}{\min\limits_{1\leqslant k\leqslant n}\sum\limits_{j'=0}^{J} s_{j'}2^{-|j(k)-j'|}},$$

where we used that $\|d_k\|_{\infty}^2 \leqslant C\cdot 2^{-j(k)}$ thanks to (5.1). Therefore Condition (3.9) is satisfied if

$$\min_{1\leqslant k\leqslant n} s_{j(k)} + \sum_{j'\neq j(k)} s_{j'}2^{-|j(k)-j'|} \gtrsim \ln(3n/\varepsilon),$$

which, in turn, is equivalent to

$$\min_{0\leqslant j\leqslant J} s_j + \sum_{j'\neq j} s_{j'}2^{-|j-j'|} \gtrsim \ln(3n/\varepsilon).$$

Theorem 3.10 can be now applied, ensuring stable and robust recovery with probability $1-\varepsilon$ with a required number of measurements

$$m \gtrsim \sum_{j=0}^{J}\left(s_j + \sum_{\substack{j'=0\\ j'\neq j}}^{J} s'_j 2^{-|j-j'|}\right)\cdot\ln(3n/\varepsilon).$$

This concludes the proof of the corollary.

F.2 *Proof of Corollary 5.3*

The proof of Corollary 5.3 relies on the estimate of $\Theta(S,\pi)$ and on the choice of the corresponding optimal measure π^Θ on the space of vertical lines. With a slight abuse of notation, here S can be interpreted as a subset of $\{1,\dots,\sqrt{n}\}^2$ or as a subset of $\{1,\dots,n\}$, depending on whether vectorization is used or not. In order to estimate $\Theta(S,\pi)$, we consider the term

$$\|D_k^* D_{k,S}\|_{\infty\to\infty} = \max_{1\leqslant i\leqslant n} \|e_i^* D_k^* D_{k,S}\|_1.$$

Recalling that $D_k = \phi_{k,:}\otimes\phi$ and that ϕ is an isometry, we obtain

$$D_k^* D_{k,S} = (\phi_{k,:}\otimes\phi)^*(\phi_{k,:}\otimes\phi)P_S^* = ((\phi_{k,:})^*\phi_{k,:}\otimes\phi^*\phi)P_S^* = ((\phi_{k,:})^*\phi_{k,:}\otimes I)P_S^*.$$

Let us explicitly write

$$(\phi_{k,:})^*\phi_{k,:}\otimes I = \begin{pmatrix} \overline{\phi_{k,1}}\phi_{k,1}I & \cdots & \overline{\phi_{k,1}}\phi_{k,\sqrt{n}}I \\ \vdots & \ddots & \vdots \\ \overline{\phi_{k,\sqrt{n}}}\phi_{k,1}I & \cdots & \overline{\phi_{k,\sqrt{n}}}\phi_{k,\sqrt{n}}I \end{pmatrix}.$$

Then, we have

$$\begin{aligned}\max_{1\leqslant i\leqslant n} \|e_i^* D_k^* D_{k,S}\|_1 &= \max_{1\leqslant q\leqslant\sqrt{n}} \max_{1\leqslant t\leqslant\sqrt{n}} \|e_t^*\left(\overline{\phi_{k,q}}\phi_{k,1}I|\cdots|\overline{\phi_{k,q}}\phi_{k,\sqrt{n}}I\right)P_S^*\|_1 \\ &= \max_{1\leqslant q\leqslant\sqrt{n}} |\phi_{k,q}| \max_{1\leqslant t\leqslant\sqrt{n}} \|e_t^*\left(\phi_{k,1}I|\cdots|\phi_{k,\sqrt{n}}I\right)P_S^*\|_1.\end{aligned}$$

The local coherence bound (5.1) yields $\max\limits_{1\leqslant q\leqslant\sqrt{n}} |\phi_{k,q}| \lesssim 2^{-j(k)/2}$. We estimate

$$\begin{aligned}\max_{1\leqslant t\leqslant\sqrt{n}} \|e_t^*\left(\phi_{k,1}I|\cdots|\phi_{k,\sqrt{n}}I\right)P_S^*\|_1 &= \max_{1\leqslant t\leqslant\sqrt{n}} \left\|\left(\phi_{k,1}(P_{S\cap C_1})_{t,:}|\cdots|\phi_{k,\sqrt{n}}(P_{S\cap C_{\sqrt{n}}})_{t,:}\right)\right\|_1 \\ &= \max_{1\leqslant t\leqslant\sqrt{n}} \sum_{q=1}^{\sqrt{n}} |\phi_{k,q}| \left\|(P_{S\cap C_q})_{t,:}\right\|_1,\end{aligned}$$

where $C_q = \{1,\dots,\sqrt{n}\}\times\{q\}$ is the q-th vertical line of $\{1,\dots,n\}^2$. Now, we observe that

$$\left\|(P_{S\cap C_q})_{t,:}\right\|_1 = \mathbb{1}_{S\cap C_q}(t) := \begin{cases} 1 & \text{if } t\in S\cap C_q \\ 0 & \text{otherwise.}\end{cases}$$

Finally, using the above relation, the decomposition of $\{1,\dots,n\}^2$ into subbands $(\Omega_{\ell,j})_{0\leqslant\ell,j\leqslant J}$, and resorting again to (5.1), we see that

$$\begin{aligned}\max_{1\leqslant t\leqslant\sqrt{n}}\sum_{q=1}^{\sqrt{n}}|\phi_{k,q}|\left\|(P_{S\cap C_q})_{t,:}\right\|_1 &= \max_{1\leqslant t\leqslant\sqrt{n}}\sum_{q=1}^{\sqrt{n}}|\phi_{k,q}|\mathbb{1}_{S\cap C_q}(t) = \max_{0\leqslant\ell\leqslant J}\max_{t\in\Omega_\ell}\sum_{j=0}^{J}\sum_{q\in\Omega_j}|\phi_{k,q}|\mathbb{1}_{S\cap C_q}(t)\\ &\leqslant \max_{0\leqslant\ell\leqslant J}\max_{t\in\Omega_\ell}\sum_{j=0}^{J}\max_{q\in\Omega_j}|\phi_{k,q}|\sum_{q\in\Omega_j}\mathbb{1}_{S\cap C_q}(t)\\ &= \max_{0\leqslant\ell\leqslant J}\max_{t\in\Omega_\ell}\sum_{j=0}^{J}\max_{q\in\Omega_j}|\phi_{k,q}||S\cap\Omega_{\ell,j}\cap R_t|\\ &\lesssim \sum_{j=0}^{J}2^{-j(k)/2}2^{-|j(k)-j|/2}s_j^r.\end{aligned}$$

This leads to the desired estimate of $\Theta(S,\pi)$. Using Lemma E.1 we can then derive the corresponding optimal measure π^Θ and conclude the proof.

F.3 *Proof of Corollary 5.4*

We divide the proof of Corollary 5.4 into two steps. First, we find the optimal sampling measure π^Λ that minimizes $\Lambda(S,\pi)$ and compute the corresponding $\Lambda(S,\pi^\Lambda)$. In the second step, we estimate $\Gamma(\pi^\Lambda)$ and derive the extra condition on the structured sparsity of the signal.

STEP 1: ESTIMATE OF Λ AND DERIVATION OF π^Λ As in the previous section, here S can be interpreted as a subset of $\{1,\dots,\sqrt{n}\}^2$ or as a subset of $\{1,\dots,n\}$, depending on whether vectorization is used or not. Let us fix a frequency $k\in W_j$. Our first goal is to estimate the quantity $\|D_{k,S}^*D_{k,S}\|_{2\to2}$. This will lead us to the optimal choice of $\pi=\pi^\Lambda$.

First, we observe that $\|D_{k,S}^*D_{k,S}\|_{2\to2}=\|D_{k,S}\|_{2\to2}^2$. Now, let $v\in\mathbb{C}^n$ and $V\in\mathbb{C}^{\sqrt{n}\times\sqrt{n}}$ be such that $v=\mathrm{Vect}(V)$ (recall that the vectorization operator $\mathrm{Vect}(\cdot)$ stacks the columns of a matrix on top of each other). Taking into account the structure of $D_{k,S}$ and v, and recalling that ϕ is an isometry, we have

$$\|D_{k,S}v\|_2^2=\left\|\sum_{i=1}^{\sqrt{n}}\phi_{k,i}\phi(V_S)_{:,i}\right\|_2^2=\left\|\phi\sum_{i=1}^{\sqrt{n}}\phi_{k,i}(V_S)_{:,i}\right\|_2^2=\left\|\sum_{i=1}^{\sqrt{n}}\phi_{k,i}(V_S)_{:,i}\right\|_2^2=\sum_{q=1}^{\sqrt{n}}\left|\sum_{i=1}^{\sqrt{n}}\phi_{k,i}(V_S)_{q,i}\right|^2.$$

Taking advantage of the decomposition of the wavelet multi-index space into tensor product subbands $(\Omega_{\ell,j})_{0\leqslant\ell,j\leqslant J}$ and recalling (5.1), we extend the previous chain of equalities as follows:

$$\begin{aligned}\|D_{k,S}v\|_2^2=\sum_{\ell=0}^{J}\sum_{q\in\Omega_\ell}\left|\sum_{j=0}^{J}\sum_{i\in\Omega_j}\phi_{k,i}(V_S)_{q,i}\right|^2 &\leqslant \sum_{\ell=0}^{J}\sum_{q\in\Omega_\ell}\left(\sum_{j=0}^{J}\max_{i\in\Omega_j}|\phi_{k,i}|\sum_{i\in\Omega_j}|(V_S)_{q,i}|\right)^2\\ &= \sum_{\ell=0}^{J}\sum_{q\in\Omega_\ell}\left(\sum_{j=0}^{J}\sqrt{\mu_{j(k),j}}\|V_{S\cap\Omega_{\ell,j}\cap R_q}\|_1\right)^2,\end{aligned}$$

where $j(k)$ is such that $k\in W_{j(k)}$ and $\mu_{\ell,j}$ is the local coherence of the Fourier-Haar transform defined in (5.1). Using the Cauchy-Schwarz inequality, rearranging the summation, and recalling Definition 5.2,

we see that

$$\|D_{k,S}v\|_2^2 \leqslant \sum_{\ell=0}^{J}\sum_{q\in\Omega_\ell}\left(\sum_{j=0}^{J}\sqrt{\mu_{j(k),j}}\sqrt{s_j^r}\|V_{S\cap\Omega_{\ell,j}\cap R_q}\|_2\right)^2 \leqslant J\sum_{\ell=0}^{J}\sum_{q\in\Omega_\ell}\sum_{j=0}^{J}\mu_{j(k),j}s_j^r\|V_{S\cap\Omega_{\ell,j}\cap R_q}\|_2^2$$
$$= J\left(\sum_{j=0}^{J}\mu_{j(k),j}s_j^r\right)\sum_{j=0}^{J}\sum_{\ell=0}^{J}\sum_{q\in\Omega_\ell}\|V_{S\cap\Omega_{\ell,j}\cap R_q}\|_2^2 = J\left(\sum_{\ell=0}^{J}\mu_{j(k),j}s_j^r\right)\|v\|_2^2.$$

Combining the above relations an using again (5.1) yields

$$\|D_{k,S}^*D_{k,S}\|_{2\to2} \leqslant J\sum_{j=0}^{J}\mu_{j(k),j}s_j^r \leqslant C\cdot J2^{-j(k)}\sum_{j=0}^{J}2^{-|j(k)-j|}s_j^r,$$

where $C>0$ is the universal constant in (5.1). Using Lemma E.1 and recalling that $|W_j| = 2^{\max(1,j)}$, the optimal probability π^Λ is, for all $k=1,\dots,\sqrt{n}$,

$$\pi_k^\Lambda = \frac{2^{-j(k)}\sum_{j=0}^{J}2^{-|j(k)-j|}s_j^r}{\sum_{t=0}^{\sqrt{n}}2^{-j(t)}\sum_{j=0}^{J}2^{-|j(t)-j|}s_j^r} = \frac{2^{-j(k)}\sum_{j=0}^{J}2^{-|j(k)-j|}s_j^r}{\sum_{\ell=0}^{J}2^{\max(1,\ell)-\ell}\sum_{j=0}^{J}2^{-|\ell-j|}s_j^r}.$$

This leads to

$$\max_{0\leqslant k\leqslant\sqrt{n}}\frac{\|D_{k,S}^*D_{k,S}\|_{2\to2}}{\pi_k^\Lambda} \leqslant CJ\cdot\sum_{\ell=0}^{J}2^{\max(1,\ell)-\ell}\sum_{j=0}^{J}2^{-|j-\ell|}s_j^r =: \Lambda(S,\pi^\Lambda).$$

STEP 2: ESTIMATE OF Γ AND CONCLUSION Employing (5.1) and recalling that ϕ is an isometry, we have

$$\|D_k\|_{1\to2}^2 = \max_{1\leqslant i\leqslant\sqrt{n}}\max_{1\leqslant q\leqslant\sqrt{n}}\|\phi_{k,i}\phi_{:,q}\|_2^2 = \max_{1\leqslant i\leqslant\sqrt{n}}\max_{1\leqslant q\leqslant\sqrt{n}}|\phi_{k,i}|^2\underbrace{\|\phi_{:,q}\|_2^2}_{=1} \leqslant C\cdot2^{-j(k)}.$$

Therefore, we estimate

$$\max_{1\leqslant k\leqslant\sqrt{n}}\frac{\|D_k\|_{1\to2}}{\pi_k^\Lambda} \leqslant \max_{1\leqslant k\leqslant\sqrt{n}}\frac{C\cdot2^{-j(k)}}{2^{-j(k)}\sum_{j=0}^{J}2^{-|j-j(k)|}s_j^r}\Big(\sum_{\ell=0}^{J}2^{\max(1,\ell)-\ell}\sum_{j=0}^{J}2^{-|j-\ell|}s_j^r\Big)$$
$$= \frac{\Lambda(S,\pi^\Lambda)}{J\cdot\min\limits_{0\leqslant\ell\leqslant J}\sum\limits_{j=0}^{J}2^{-|j-\ell|}s_j^r} =: \Gamma(\pi^\Lambda).$$

As a consequence, the condition (3.9) of Theorem 3.10 reads

$$J\cdot\min_{0\leqslant\ell\leqslant J}\sum_{j=0}^{J}2^{-|j-\ell|}s_j^r \gtrsim \ln(3n/\varepsilon)$$

and the minimum sampling complexity is

$$m \gtrsim \Lambda(S,\pi^\Lambda)\ln(3n/\varepsilon),$$

which, in turn, is implied by

$$m \gtrsim \left(\sum_{j=0}^{J} s_j^r + \sum_{\substack{\ell=0 \\ \ell \neq j}}^{J} 2^{-|\ell-j|} s_\ell^r \right) \log_2(\sqrt{n}) \ln(3n/\varepsilon).$$

This concludes the proof.

References

[1] Ben Adcock, Simone Brugiapaglia, and Clayton G. Webster. Compressed sensing approaches for polynomial approximation of high-dimensional functions. In *Compressed Sensing and its Applications*, pages 93–124. Springer, 2017.

[2] Ben Adcock, Anders C. Hansen, Clarice Poon, and Bogdan Roman. Breaking the coherence barrier: A new theory for compressed sensing. In *Forum of Mathematics, Sigma*, volume 5. Cambridge University Press, 2017.

[3] Ben Adcock, Anders C. Hansen, and Bogdan Roman. The quest for optimal sampling: Computationally efficient, structure-exploiting measurements for compressed sensing. *Book Chapter, Compressed Sensing and its Applications, Springer (to appear), arXiv preprint arXiv:1403.6540*, 2014.

[4] Ben Adcock, Anders C. Hansen, and Bogdan Roman. A note on compressed sensing of structured sparse wavelet coefficients from subsampled fourier measurements. *IEEE Signal Processing Letters*, 23(5):732–736, 2016.

[5] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

[6] Jérémie Bigot, Claire Boyer, and Pierre Weiss. An analysis of block sampling strategies in compressed sensing. *IEEE transactions on information theory*, 62(4):2125–2139, 2016.

[7] Claire Boyer, Jérémie Bigot, and Pierre Weiss. Compressed sensing with structured sparsity and structured acquisition. *Applied and Computational Harmonic Analysis*, 2017.

[8] Emmanuel Candès and Yaniv Plan. A probabilistic and ripless theory of compressed sensing. *Information Theory, IEEE Transactions on*, 57(11):7235–7254, 2011.

[9] Emmanuel Candès and Justin Romberg. Sparsity and incoherence in compressive sampling. *Inverse problems*, 23(3):969, 2007.

[10] Nicolas Chauffert, Philippe Ciuciu, Jonas Kahn, and Pierre Weiss. Variable density sampling with continuous trajectories. *SIAM Journal on Imaging Sciences*, 7(4):1962–1992, 2014.

[11] Nicolas Chauffert, Philippe Ciuciu, and Pierre Weiss. Variable density compressed sensing in MRI. Theoretical vs heuristic sampling strategies. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pages 298–301. IEEE, 2013.

[12] Abdellah Chkifa, Nick Dexter, Hoang Tran, and Clayton G. Webster. Polynomial approximation via compressed sensing of high-dimensional functions on lower sets. *Mathematics of Computation*, 2017.

[13] Il Yong Chun and Ben Adcock. Compressed sensing and parallel acquisition. *IEEE Transactions on Information Theory*, 2017.

[14] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Springer, 2013.

[15] Felix Krahmer and Rachel Ward. Stable and robust sampling strategies for compressive imaging. *IEEE transactions on image processing*, 23(2):612–622, 2014.

[16] Chen Li and Ben Adcock. Compressed sensing with local structure: uniform recovery guarantees for the sparsity in levels class. *Appl. Comput. Harmon. Anal. (to appear)*, 2017.

[17] Li Liu, Yuntao He, Jianguo Zhang, Huayu Jia, and Jun Ma. Optimum linear array for aperture synthesis imaging based on redundant spacing calibration. *Optical Engineering*, 53(5):053109, 2014.

[18] Gilles Puy, Pierre Vandergheynst, and Yves Wiaux. On variable density compressive sampling. *Signal Processing Letters, IEEE*, 18(10):595–598, 2011.

[19] Céline Quinsac, Adrian Basarab, Jean-Marc Girault, and Denis Kouamé. Compressed sensing of ultrasound images: Sampling of spatial and frequency domains. In *Signal Processing Systems (SIPS), 2010 IEEE Workshop on*, pages 231–236. IEEE, 2010.

[20] Holger Rauhut and Rachel Ward. Sparse legendre expansions via ℓ_1-minimization. *Journal of approximation theory*, 164(5):517–533, 2012.

[21] Joel A. Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, pages 67–101. Springer, 2015.

[22] Ewout van den Berg and Michael P. Friedlander. SPGL1: A solver for large-scale sparse reconstruction, June 2007. http://www.cs.ubc.ca/labs/scl/spgl1.

[23] Ewout van den Berg and Michael P. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2008.

[24] Yi Wang. Description of parallel imaging in mri using multiple coils. *Magnetic Resonance in Medicine*, 44(3):495–499, 2000.