

Improved recovery guarantees and sampling strategies for TV minimization in compressive imaging *

Ben Adcock, Nick Dexter and Qinghong Xu[†]

Abstract. In this paper, we consider the use of Total Variation (TV) minimization for compressive imaging; that is, image reconstruction from subsampled measurements. Focusing on two important imaging modalities – namely, Fourier imaging and structured binary imaging via the Walsh–Hadamard transform – we derive uniform recovery guarantees asserting stable and robust recovery for arbitrary random sampling strategies. Using this, we then derive a class of theoretically-optimal sampling strategies. For Fourier sampling, we show recovery of an image with approximately s -sparse gradient from $m \gtrsim_d s \cdot \log^2(s) \cdot \log^4(N)$ measurements, in $d \geq 1$ dimensions. When $d = 2$, this improves the current state-of-the-art result by a factor of $\log(s) \cdot \log(N)$. It also extends it to arbitrary dimensions $d \geq 2$. For Walsh sampling, we prove that $m \gtrsim_d s \cdot \log^2(s) \cdot \log^2(N/s) \cdot \log^3(N)$ measurements suffice in $d \geq 2$ dimensions. To the best of our knowledge, this is the first recovery guarantee for structured binary sampling with TV minimization.

Key words. compressive imaging, TV minimization, Fourier imaging, binary imaging, sampling strategies

AMS subject classifications. 94A08, 94A20, 68U10, 68Q25

1. Introduction. Total Variation (TV) minimization is an important technique in modern image processing [12, 13], with a wide range of applications including denoising, deblurring and reconstruction. In this paper, we consider the latter problem. Specifically, given noisy, linear measurements $y = Ax + e \in \mathbb{C}^m$ of an unknown d -dimensional image $x \in \mathbb{C}^{N^d}$, we study its reconstruction via the constrained TV minimization problem

$$(1.1) \quad \min_{z \in \mathbb{C}^{N^d}} \|z\|_{\text{TV}} \text{ subject to } \|Az - y\|_{\ell^2} \leq \eta,$$

where $\|\cdot\|_{\text{TV}}$ is the TV semi-norm. Natural images have approximately sparse gradients. As is now well known, minimizing the TV semi-norm promotes this structure, often leading to high-quality reconstructions from a relatively small number of measurements. TV minimization has proved an extremely effective tool for image reconstruction, with many applications in medical, scientific and industrial modalities.

A fundamental issue in image reconstruction is choosing a measurement matrix A . The main goal of so-called *compressive imaging* is to choose A so as to deliver high-quality reconstructions from as few measurements m as possible. Generally speaking, the possible choices are dictated by the physical sensing apparatus. In this paper, we consider two important image acquisition protocols: namely, Fourier sampling with the discrete Fourier transform and binary sampling via the Walsh–Hadamard transform. Arguably, these are two out of the

*Submitted to the editors DATE.

Funding: ND acknowledges the support of the PIMS Postdoctoral Fellowship program. This work was supported by the PIMS CRG “High-dimensional Data Analysis”, SFU’s Big Data Initiative “Next Big Question” Fund and NSERC through grant R611675.

[†]Simon Fraser University, 8888 University Drive Burnaby, BC V5A 1S6, Canada (ben_adcock@sfu.ca, nicholas_dexter@sfu.ca, qinghong_xu@sfu.ca)

three most important types of sampling encountered in imaging – the other being the Radon transform. Fourier sampling arises in numerous applications, including Magnetic Resonance Imaging (MRI), Nuclear Magnetic Resonance (NMR) and radio interferometry, while binary sampling arises in numerous optical imaging modalities, such as lensless imaging, infrared imaging holography, fluorescence microscopy and so forth.

Once the acquisition protocol has been fixed, the task of selecting measurements reduces to that of designing a *sampling strategy*, i.e. a specific choice of m Fourier or Walsh frequencies to sample. The main objective of this paper is to develop sampling strategies for TV minimization in these scenarios. In tandem, we also derive sufficient conditions on the number of measurements m under which the underlying image is accurately recovered via (1.1). We do this by leveraging the theory of compressed sensing [16] to prove new recovery guarantees for TV minimization which relate the number of measurements m to the approximate gradient sparsity s of the underlying image.

1.1. Previous work. TV minimization was studied in some of the first papers on compressed sensing. In [10], Candès, Romberg & Tao considered the recovery of a one-dimensional image $x \in \mathbb{C}^N$ with exactly s -sparse gradient from m noiseless Fourier measurements taken uniformly and randomly. They showed that x could be recovered exactly by solving (1.1) with $\eta = 0$ with high probability, provided the number of measurements $m \gtrsim s \cdot \log(N)$. The first results asserting recovery for approximately sparse images from noisy measurements were shown by Needell & Ward for the two-dimensional case in [25], and later for the d -dimensional case in [24]. In particular, these works were the first to exploit (in the compressed sensing context) the important connection between the TV semi-norm and Haar wavelet coefficients. Neither of these works pertained directly to Fourier sampling. The first results on Fourier sampling were shown by Krahmer & Ward [20] and Poon [26]. In the former, uniform recovery guarantees¹ were shown for two-dimensional images from noisy Fourier measurements, with frequencies chosen randomly according to an *inverse square law* density. Specifically, if

$$(1.2) \quad m \gtrsim s \cdot \log^3(s) \cdot \log^5(N),$$

then with high probability, the recovered vector \hat{x} satisfies

$$(1.3) \quad \|x - \hat{x}\|_{\ell^2} \lesssim \frac{\sigma_s(\nabla x)_{\ell^1}}{\sqrt{s}} + \eta, \quad \sigma_s(\nabla x)_{\ell^1} = \min\{\|\nabla x - z\|_{\ell^1} : z \in \mathbb{C}^{N^2} \text{ is } s\text{-sparse}\},$$

where η is an upper bound on a certain weighted ℓ^2 -norm of the noise term e . Conversely, [26] established nonuniform recovery guarantees in the one- and two-dimensional cases for both uniform random sampling and variable density sampling. Amongst other features, [26] was the first to prove results demonstrating the benefits of variable density sampling: namely, while both uniform random and variable density sampling recover the image gradient accurately, the latter leads to better recovery of the image itself. In comparison with (1.2)–(1.3), in the two-dimensional case [26] showed that if

$$(1.4) \quad m \gtrsim s \cdot \log(N),$$

¹In compressed sensing, a *uniform* recovery guarantee states that a single random draw of a given measurement matrix suffices for recovery of all (approximately) sparse vectors. This is stronger than a nonuniform recovery guarantee, which asserts that a single random draw is sufficient for recovery of a fixed vector.

Fourier samples were drawn using a combination of uniform random and inverse square law sampling, then, with high probability,

$$(1.5) \quad \|x - \hat{x}\|_{\ell^2} \lesssim \log(s) \cdot \log(N^2/s) \log^{1/2}(N) \log^{1/2}(m) \left(\log^{1/2}(m) \log(s) \frac{\sigma_s(\nabla x)_{\ell^{2,1}}}{\sqrt{s}} + \eta \right),$$

where η is an upper bound for the (unweighted) ℓ^2 -norm of the noise (the appearance of $\sigma(\cdot)_{\ell^{2,1}}$ here indicates that [26] considered the isotropic TV norm, whereas [20] considered the anisotropic TV norm – see later). In particular, this approach leads to a better measurement condition (1.4) than the measurement condition (1.2), but a correspondingly worse error bound (1.5) over (1.3).

1.2. Contributions. The above results of [20] and [26] represent the state-of-the-art recovery guarantees for TV minimization in compressed sensing with Fourier sampling. In this paper we improve and generalize these results in the following ways:

1. We derive recovery guarantees in $d \geq 1$ dimensions, as opposed to $d = 2$ in [20] and $d = 1, 2$ in [26]. We consider both the isotropic (like in [26]) and anisotropic (like in [20]) TV semi-norms. Also as in [26] we examine both uniform random and variable density sampling.

2. As in [20], our recovery guarantees are uniform, and when $d \geq 2$ they take the form

$$(1.6) \quad \|x - \hat{x}\|_{\ell^2} \lesssim_d \frac{\sigma_s(\nabla x)_{\ell^1}}{\sqrt{s}} + \sqrt{\log(N)}\eta,$$

for variable density sampling. Unlike [20], we do not impose a weighted norm on the noise vector. This gives rise to the $\sqrt{\log(N)}$ factor in (1.6). As in [26], we also derive error bounds for the recovery of the image gradient ∇x .

3. Unlike [20, 26] we derive a recovery guarantee for arbitrary variable density sampling schemes in order to examine the effect of the sampling scheme on the measurement condition.

4. We derive theoretically-optimal variable density sampling schemes in $d \geq 1$ dimensions. For such schemes, our measurement condition is

$$m \gtrsim_d s \cdot \log^2(s) \cdot \log^4(N).$$

In particular, for the $d = 2$ case, we improve the current state-of-the-art measurement condition (1.2) for uniform recovery by a factor of $\log(s) \cdot \log(N)$. When $d = 2$ we show that the inverse square law scheme of [20, 26] is an instance of a theoretically-optimal scheme.

5. Interestingly, we show that the theoretically-optimal Fourier sampling scheme ceases to be radially-symmetric in $d \geq 3$ dimensions. We also derive a near-optimal sampling scheme based on so-called *hyperbolic cross* sampling densities.

6. Finally, unlike [20, 26] we also consider binary sampling with the Walsh–Hadamard transform. In this case, we prove a recovery guarantee of the form

$$\|x - \hat{x}\|_{\ell^2} \lesssim_d \frac{\sigma_s(\nabla x)_{\ell^1}}{\sqrt{s \log(N)}} + \sqrt{\log(N)}\eta,$$

and derive theoretically-optimal variable density sampling strategies for which the measurement condition reads

$$m \gtrsim_d s \cdot \log^2(s) \cdot \log^2(N/s) \cdot \log^3(N).$$

Unlike in the Fourier case, we show that certain radially-symmetric sampling schemes are theoretically optimal in any dimension for Walsh sampling. To the best of our knowledge, this is the first recovery guarantee for TV minimization with structured binary sampling. For results on binary sampling with wavelet sparsifying transforms, see [2, 22, 23].

Note that our focus in this paper is on Fourier and Walsh sampling, since these acquisition protocols arise in many practical imaging settings. Although common in compressed sensing, we do not consider sampling with Gaussian or Bernoulli random matrices. These are generally infeasible for imaging, since they lead to dense, unstructured matrices. Moreover, even if they were, it is well known that they are highly suboptimal for imaging, being significantly outperformed by structured Fourier and Walsh sampling [4, 3, 27]. For recovery guarantees for TV minimization from Gaussian or Bernoulli measurements, see [9, 19].

1.3. Structure dependence. Similar to [20, 26], the sampling schemes we develop in this paper are independent of the image (or class of images) being recovered. In particular, they exploit only the sparsity of ∇x and no further local, or geometric, properties of the edges of x . As has been well documented [4, 26, 27], optimal sampling strategies in practice should also take local properties into account: roughly speaking, an image with well separated edges should be sampled more densely at low frequencies than an image with edges that lie close to each other, even when the two images possess the same gradient sparsity. In the case of sparsity in orthonormal wavelets, it is well understood (from a theoretical and practical perspective) how to design sampling strategies that exploit such local structure [2, 3, 4, 7, 21]. Yet, this is not well understood for gradient sparsity. We shall not attempt to tackle this problem, although we do discuss it in the context of our numerical examples. We refer to [11, 26] for some further discussion on this topic. Nonetheless, as we show in our examples, good all-round performance across a range of images, resolutions and sampling percentages can be achieved with an (image independent) *multilevel random sampling* strategy. This scheme was originally developed for wavelet sparsifying transforms in [4]. We show that it also achieves similarly good performance for TV minimization.

1.4. Outline. We begin in §2 with preliminaries. We state our main results for Fourier and Walsh sampling in §3–§4 and §5 respectively. In §6 we present several numerical experiments. Finally, in §7–§8 we give the proofs of the main results. The Supplementary Material contains some supporting material and proofs of several of the minor results.

2. Preliminaries. We first introduce some notation and background material.

2.1. Notation. We denote the ℓ^p -norm on \mathbb{C}^N by $\|\cdot\|_{\ell^p}$ and the ℓ^2 -inner product by $\langle \cdot, \cdot \rangle$. For $1 \leq p, q < \infty$, we define the $\ell^{p,q}$ -norm on $\mathbb{C}^{N \times M}$ as

$$\|X\|_{\ell^{p,q}} = \left(\sum_{i=1}^N \left(\sum_{j=1}^M |x_{ij}|^p \right)^{q/p} \right)^{1/q}, \quad X = (x_{ij})_{i,j=1}^{N,M}.$$

Note that $\|X\|_{\ell^{2,2}} = \|X\|_F$ is the Frobenius norm of F . We define the ℓ^0 -norm of a vector $x = (x_i)_{i=1}^N$ as $\|x\|_{\ell^0} = |\text{supp}(x)|$, where $\text{supp}(x) = \{i : x_i \neq 0\}$ is the support of x . For a matrix $X = (x_{ij})_{i,j=1}^{N,M} \in \mathbb{C}^{N \times M}$ we define the $\ell^{2,0}$ -norm as

$$\|X\|_{\ell^{2,0}} = |\text{supp}(X)|, \quad \text{supp}(X) = \left\{ i : \sum_{j=1}^M |x_{ij}|^2 \neq 0 \right\}.$$

Given a subset $\Delta \subseteq \{1, \dots, N\}$, we write $P_\Delta \in \mathbb{C}^{N \times N}$ for the diagonal matrix corresponding to the orthogonal projection with range $\text{span}\{e_i : i \in \Delta\}$, where $\{e_i\}_{i=1}^N$ is the canonical basis of \mathbb{C}^N . Note that for $x \in \mathbb{C}^N$, the vector $P_\Delta x$ is isometrically isomorphic to a vector in $\mathbb{C}^{|\Delta|}$, and similarly for $X \in \mathbb{C}^{N \times M}$, $P_\Delta X$ is isomorphic to an element of $\mathbb{C}^{|\Delta| \times M}$. On occasion we therefore slightly abuse notation and consider $P_\Delta x$ as an element of $\mathbb{C}^{|\Delta|}$ or $P_\Delta X$ as an element of $\mathbb{C}^{|\Delta| \times N}$.

We write $C > 0$ for a numerical constant and $C_x > 0$ for a constant depending on a variable x . We use the notation $a \lesssim b$ to mean there exists $C > 0$ such that $a \leq Cb$, and likewise for the symbol \gtrsim . We also write $a \lesssim_x b$ when $a \leq C_x b$ for some $C_x > 0$ depending on a variable x , and likewise for \gtrsim_x . We write $a \asymp b$ or $a \asymp_x b$ if $a \lesssim b \lesssim a$ or $a \lesssim_x b \lesssim_x a$.

2.2. Discrete images. We consider discrete, d -dimensional complex images

$$X = (X_{i_1, \dots, i_d})_{i_1, \dots, i_d=1}^N \in \mathbb{C}^{N \times \dots \times N},$$

where N is its *resolution*. The motivation to consider complex images stems primarily from MRI, where the images are often complex. We assume throughout this paper that $N = 2^r$ is a power of two, where $r \geq 1$. We often reshape X into a vector using lexicographical ordering. Let $\varsigma : \{1, \dots, N^d\} \rightarrow \{1, \dots, N\}^d$ be the bijection corresponding to this ordering, defined via its inverse as

$$\varsigma^{-1}(i_1, \dots, i_d) = N^{d-1}i_1 + N^{d-2}i_2 + \dots + i_d, \quad (i_1, \dots, i_d) \in \{1, \dots, N\}^d.$$

Given X , we let $x = (x_i)_{i=1}^{N^d} \in \mathbb{C}^{N^d}$ be such that $x_i = X_{\varsigma(i)}$ and write $x = \text{vec}(X)$.

2.3. The Discrete Fourier Transform and recovery problem. We order frequency from lowest to highest in absolute value. Define the bijection

$$(2.1) \quad \varrho : \{1, \dots, N\} \rightarrow \{-N/2 + 1, \dots, N/2\}, \quad i \mapsto (-1)^i \lfloor i/2 \rfloor.$$

With this order, we define the one-dimensional Discrete Fourier Transform (DFT) matrix $F = F^{(1)} \in \mathbb{C}^{N \times N}$ as

$$F_{ij} = \exp(-2\pi i \varrho(i)(j-1)/N), \quad i, j = 1, \dots, N,$$

(this differs from the usual DFT matrix by a row permutation and diagonal scaling, but is beneficial for our purposes as it orders frequencies from lowest to highest).

The d -dimensional DFT $F = F^{(d)} \in \mathbb{C}^{N^d \times N^d}$ is given by $F^{(d)} = F^{(1)} \otimes \dots \otimes F^{(1)}$, where \otimes denotes the Kronecker product. Notice that $N^{-d} F^* F = I$ is the identity matrix. The rows

of $F^{(d)}$ correspond to the d -dimensional frequency space $\{-N/2 + 1, \dots, N/2\}^d$. Specifically, let $\varrho = \varrho^{(d)} : \{1, \dots, N^d\} \rightarrow \{-N/2 + 1, \dots, N/2\}^d$ be the bijection defined by

$$(2.2) \quad \varrho^{(d)}(i) = (\varrho(\varsigma(i)_1), \dots, \varrho(\varsigma(i)_d)), \quad i \in \{1, \dots, N^d\},$$

where ς is the lexicographical ordering and ϱ is the one-dimensional bijection (2.1). Then the i^{th} row of $F^{(d)}$ corresponds to the frequency $\omega = \varrho(i)$.

In the first part of this paper, we consider the problem of recovering a vectorized image x from m of its Fourier frequencies. The choice of frequencies is variously referred to as a *sampling scheme, strategy, map or pattern*. We consider two main types of sampling schemes:

Definition 2.1 (Uniform random sampling). *A d -dimensional uniform random sampling scheme of order m is a subset of frequencies $\Omega = \{\omega_1, \dots, \omega_m\} \subseteq \{-N/2 + 1, \dots, N/2\}^d$ where the ω_i are chosen independently and uniformly from $\{-N/2 + 1, \dots, N/2\}^d$.*

Definition 2.2 (Variable density sampling). *Let $p = (p_\omega)$ be a probability distribution on $\{-N/2 + 1, \dots, N/2\}^d$. A d -dimensional variable density sampling scheme of order m is a subset of frequencies $\Omega = \{\omega_1, \dots, \omega_m\} \subseteq \{-N/2 + 1, \dots, N/2\}^d$ where the ω_i are chosen i.i.d. according to p .*

Let Ω be given by one of these schemes. With slight abuse of notation, write $P_\Omega \in \mathbb{C}^{N^d \times N^d}$ for the orthogonal projection onto the indices in Ω (technically, this should be $P_{\varrho^{-1}(\Omega)}$ with ϱ as in (2.2)). Then we write $A = \frac{1}{\sqrt{m}} P_\Omega F \in \mathbb{C}^{m \times N^d}$ for the corresponding measurement matrix. This is an example of a subsampled DFT matrix: the vector Ax consists of the m frequency values of the vectorized image x from the set Ω . We assume these values are also corrupted by noise, giving the vector of measurements

$$y = Ax + e \in \mathbb{C}^m,$$

where $e \in \mathbb{C}^m$ is a noise vector. With this in hand, the recovery problem we aim to solve is the following: *given $y = Ax + e$, recover x .*

2.4. The Discrete Walsh–Hadamard Transform and recovery problem. We now define the Discrete Walsh–Hadamard Transform. Recall that the one-dimensional (sequency-ordered) Walsh functions on $[0, 1)$ are defined by

$$v_n(x) = (-1)^{\sum_{i=1}^{\infty} (n_i + n_{i+1})x_i}, \quad 0 \leq x < 1, \quad n \in \mathbb{N}_0,$$

where $(n_i)_{i \in \mathbb{N}} \in \{0, 1\}^{\mathbb{N}}$ and $(x_i)_{i \in \mathbb{N}} \in \{0, 1\}^{\mathbb{N}}$ are the dyadic expansions of n and x respectively (see [6, 17, 18] and references therein for further information on Walsh functions). The number $n \in \mathbb{N}_0$ is the *sequency* (number of sign changes) of the Walsh function; it is therefore analogous to the Fourier frequency. The functions v_n take values in $\{+1, -1\}$ and form an orthonormal basis of $L^2([0, 1))$. When $N = 2^r$, the *Discrete Walsh–Hadamard Transform* (DHT) arises by sampling this basis on an equispaced grid in $[0, 1)$:

$$H = H^{(1)} = (v_m(n/N))_{m,n=0}^{N-1} \in \{-1, 1\}^{N \times N}.$$

Note that other orderings of the Walsh functions (or equivalently the rows of H) could be considered here, e.g. the Paley or ordinary orderings. The sequency ordering is convenient due to its connection to frequency; we therefore use it throughout. When $d \geq 2$, we write $H^{(d)} = H^{(1)} \otimes \dots \otimes H^{(1)}$ for the d -dimensional DHT matrix. Note that in any dimension, H is a symmetric matrix and is orthogonal up to a constant: specifically, $N^{-d} H^\top H = I$.

In $d \geq 1$ dimensions, the transform $x \mapsto Hx$ computes the discrete Walsh–Hadamard measurements of a vectorized image x corresponding to the frequencies in $\{0, \dots, N-1\}^d$. Specifically, let $\varrho : \{1, \dots, N^d\} \rightarrow \{0, \dots, N-1\}^d$ be the bijection defined by

$$\varrho(i) = (\varsigma(i)_1 - 1, \dots, \varsigma(i)_d - 1), \quad i \in \{1, \dots, N^d\},$$

where ς is the lexicographical ordering. Then the i^{th} row of $H^{(d)}$ corresponds to the Walsh frequency $n = \varrho(i)$ with i^{th} entry of Hx being the Walsh frequency of x .

Similar to the Fourier case, we consider sampling schemes $\Omega = \{\omega_1, \dots, \omega_m\} \subseteq \{0, \dots, N-1\}^d$. Uniform random and variable density sampling schemes are all defined in the analogous manner, with the notable difference that in Walsh–Hadamard sampling the frequencies are nonnegative numbers only, as opposed to arbitrary integers. As in the Fourier case, we write $A = \frac{1}{\sqrt{m}} P_\Omega H \in \mathbb{R}^{m \times N^d}$ for the subsampled DHT matrix. Hence the noisy measurements are given by $y = Ax + e \in \mathbb{R}^m$ and the recovery problem is to recover x from y .

2.5. Gradient operators and TV semi-norms. We consider periodic gradient operators. The one-dimensional discrete gradient operator $\nabla : \mathbb{C}^N \rightarrow \mathbb{C}^N$ is defined by

$$(\nabla x)_i = x_{i+1} - x_i, \quad i = 1, \dots, N,$$

where $x = (x_i)_{i=1}^N$ and $x_{N+1} = x_1$. The one-dimensional *Total Variation* semi-norm $\|\cdot\|_{\text{TV}}$ is $\|x\|_{\text{TV}} = \|\nabla x\|_{\ell^1}$. Note that ∇ is the circulant matrix generated by the vector $(-1, 0, \dots, 0, 1)$.

In d dimensions, we define the j^{th} partial derivative operator $\nabla_j : \mathbb{C}^{N^d} \rightarrow \mathbb{C}^{N^d}$ as

$$\nabla_j = \underbrace{I \otimes \dots \otimes I}_{d-j} \otimes \nabla \otimes \underbrace{I \otimes \dots \otimes I}_{j-1},$$

where ∇ is the one-dimensional discrete gradient operator and $I \in \mathbb{C}^{N \times N}$ is the identity matrix. When $d \geq 2$, there is more than one way to define the TV semi-norm. We define the d -dimensional *isotropic* discrete gradient operator as

$$(2.3) \quad \nabla : \mathbb{C}^{N^d} \rightarrow \mathbb{C}^{N^d \times d}, \quad x \mapsto \nabla x = \begin{pmatrix} \nabla_1 x & \dots & \nabla_d x \end{pmatrix}.$$

The d -dimensional *isotropic* TV semi-norm is $\|x\|_{\text{TV}_i} = \|\nabla x\|_{\ell^{2,1}}$, where $\nabla x \in \mathbb{C}^{N^d \times d}$ is as in (2.3). Alternatively, the d -dimensional *anisotropic* discrete gradient operator is

$$(2.4) \quad \nabla : \mathbb{C}^{N^d} \rightarrow \mathbb{C}^{dN^d}, \quad x \mapsto \nabla x = \begin{pmatrix} \nabla_1 x \\ \vdots \\ \nabla_d x \end{pmatrix},$$

and d -dimensional *anisotropic* TV semi-norm is $\|x\|_{\text{TV}_a} = \|\nabla x\|_{\ell^1}$, where $\nabla x \in \mathbb{C}^{dN^d}$ is as in (2.4). Notice that these semi-norms are equivalent up to a constant:

$$(2.5) \quad \|x\|_{\text{TV}_i} \leq \|x\|_{\text{TV}_a} \leq \sqrt{d}\|x\|_{\text{TV}_i}.$$

2.6. TV minimization problem. Let $X \in \mathbb{C}^{N \times \dots \times N}$ be an image, $x \in \mathbb{C}^{N^d}$ be its vectorization, $A \in \mathbb{C}^{m \times N^d}$ be a measurement matrix, as defined above, and $y = Ax + e$ be noisy measurements. To recover x from y , we consider the constrained TV minimization problem

$$(2.6) \quad \min_{z \in \mathbb{C}^{N^d}} \|z\|_{\text{TV}} \text{ subject to } \|Az - y\|_{\ell^2} \leq \eta,$$

where $\eta \geq \|e\|_{\ell^2}$ is an upper bound on the noise level, and $\|\cdot\|_{\text{TV}}$ denotes either the isotropic or anisotropic TV norm. We write $\hat{x} \in \mathbb{C}^{N^d}$ for a minimizer of this problem, which is the reconstruction of x , and $\hat{X} \in \mathbb{C}^{N \times \dots \times N}$ for the corresponding reconstruction of X .

2.7. Gradient sparsity and best s -term approximation. In what follows, we derive conditions on Ω and m under which the error $\|x - \hat{x}\|_{\ell^2} = \|X - \hat{X}\|_{\ell^2, 2}$ satisfies a bound depending on the gradient sparsity of the image. To this end, we define the ℓ^1 -norm best s -term approximation error of a vector $x \in \mathbb{C}^N$ as

$$\sigma_s(x)_{\ell^1} = \min\{\|x - z\|_{\ell^1} : \|z\|_{\ell^0} \leq s\}.$$

Similarly, we define the $\ell^{2,1}$ -norm best s -term approximation error of a matrix $X \in \mathbb{C}^{N \times M}$ as

$$\sigma_s(X)_{\ell^{2,1}} = \min\{\|X - Z\|_{\ell^{2,1}} : \|Z\|_{\ell^{2,0}} \leq s\}.$$

3. Main results on Fourier sampling. We now present our main results on Fourier sampling. We consider Walsh sampling in §5.

3.1. Uniform random Fourier sampling. Based on [26], we first consider uniform random Fourier sampling, as in Definition 2.1. For reasons that will become clear, we separate our results into the $d = 1$ and $d \geq 2$ cases:

Theorem 3.1 (Uniform Fourier sampling, one dimension). *Let $d = 1$, $0 < \varepsilon < 1$, $2 \leq s, m \leq N$ and $\Omega = \Omega_1 \cup \Omega_2$, where $\Omega_1 \subseteq \{-N/2 + 1, \dots, N/2\}$ is a uniform random sampling scheme of order $m - 1$ and $\Omega_2 = \{0\}$. Let $A = \frac{1}{\sqrt{m}}P_\Omega F \in \mathbb{C}^{m \times N}$ and*

$$m \gtrsim s \cdot \log(s) \cdot (\log(s) \cdot \log(N) + \log(\varepsilon^{-1})).$$

Then the following holds with probability at least $1 - \varepsilon$. For all $x \in \mathbb{C}^N$ and $y = Ax + e \in \mathbb{C}^m$, where $\|e\|_{\ell^2} \leq \eta$ for some $\eta \geq 0$, every minimizer \hat{x} of (2.6) satisfies

$$(3.1) \quad \|\nabla x - \nabla \hat{x}\|_{\ell^2} \lesssim \frac{\sigma_s(\nabla x)_{\ell^1}}{\sqrt{s}} + \eta, \quad \|x - \hat{x}\|_{\text{TV}} \lesssim \sigma_s(\nabla x)_{\ell^1} + \sqrt{s}\eta,$$

and

$$(3.2) \quad \frac{\|x - \hat{x}\|_{\ell^2}}{\sqrt{N}} \lesssim \sigma_s(\nabla x)_{\ell^1} + \sqrt{s}\eta.$$

Theorem 3.2 (Uniform Fourier sampling, $d \geq 2$ dimensions). *Let, $d \geq 2$, $0 < \varepsilon < 1$, $2 \leq s, m \leq N^d$ and $\Omega = \Omega_1 \cup \Omega_2$, where Ω_1 is a d -dimensional uniform random sampling map of order $m - 1$ and $\Omega_2 = \{(0, 0, \dots, 0)\}$. Let $A = \frac{1}{\sqrt{m}} P_\Omega F \in \mathbb{C}^{m \times N^d}$ and*

$$(3.3) \quad m \gtrsim s \cdot \log(s) \cdot (d \cdot \log(s) \cdot \log(N) + \log(\varepsilon^{-1})),$$

Then the following holds with probability at least $1 - \varepsilon$. For all $x \in \mathbb{C}^{N^d}$ and $y = Ax + e \in \mathbb{C}^m$, where $\|e\|_{\ell^2} \leq \eta$ for some $\eta \geq 0$, every minimizer \hat{x} of (2.6) satisfies

$$(3.4) \quad \|\nabla x - \nabla \hat{x}\|_{\ell^2} \lesssim \frac{\sigma_s(\nabla x)_{\ell^1}}{\sqrt{s}} + d\eta, \quad \|x - \hat{x}\|_{\text{TV}_a} \lesssim \sigma_s(\nabla x)_{\ell^1} + \sqrt{sd}\eta \quad (\text{anisotropic}),$$

$$(3.5) \quad \|\nabla \hat{x} - \nabla x\|_{\ell^{2,2}} \lesssim \frac{\sigma_s(\nabla x)_{\ell^{2,1}}}{\sqrt{s}} + \sqrt{d}\eta, \quad \|x - \hat{x}\|_{\text{TV}_i} \lesssim \sigma_s(\nabla x)_{\ell^{2,1}} + \sqrt{s}\sqrt{d}\eta \quad (\text{isotropic}),$$

and

$$(3.6) \quad \|\hat{x} - x\|_{\ell^2} \lesssim 2^{-d/2} \sigma_s(\nabla x)_{\ell^1} + (1 + 2^{-d/2} \sqrt{sd})\eta \quad (\text{anisotropic}),$$

$$(3.7) \quad \|\hat{x} - x\|_{\ell^2} \lesssim 2^{-d/2} \sqrt{d} \sigma_s(\nabla x)_{\ell^{2,1}} + (1 + 2^{-d/2} \sqrt{sd})\eta \quad (\text{isotropic}).$$

These results assert recovery of x from roughly $s \cdot \log^2(s) \cdot \log(N)$ measurements for fixed d , i.e. linear in s up to the log factors. The gradient error bound in the ℓ^2 -norm (or $\ell^{2,2}$ -norm in the case of the anisotropic TV semi-norm) is the typical stable and robust recovery guarantee found ubiquitously in compressed sensing [16]. Specifically, the error depends on a best s -term approximation error $\sigma_s(\nabla x)_{\ell^1} / \sqrt{s}$ (stability) and the noise level η (robustness).

Conversely, the recovery of the image x is worse by a factor of \sqrt{s} than the recovery of its gradient – compare, for example, (3.4) with (3.6). As observed previously in [26], this is due to the choice of a uniform random sampling scheme. In the next section we improve the stability and robustness of the image recovery by adding samples drawn from a variable density. We remark in passing that the one-dimensional signal recovery bound (3.2) involves a factor of $1/\sqrt{N}$. This factor is natural when considering x as the discretization of a continuous image [26, Rem. 2.1].

As noted, nonuniform recovery guarantees for uniform random Fourier sampling were shown in [26]. In one dimension, [26, Thm. 2.3] asserts that

$$m \gtrsim s \cdot \log(N) \cdot (1 + \log(\varepsilon^{-1})),$$

measurements are sufficient for an error bound of the form

$$\frac{\|x - \hat{x}\|_{\ell^2}}{\sqrt{N}} \lesssim \log^{1/2}(m) \log(s) \sigma_s(\nabla x)_{\ell^1} + \eta \sqrt{s}.$$

Our uniform recovery guarantee (Theorem 3.1) imposes a higher sample complexity (by a factor of $\log^2(s)$), but obtains an improved error bound (3.2), in which no log factors appear. The same comparison can be made in $d = 2$ dimensions. See Theorem 3.2 and [26, Thm. 2.4].

3.2. Variable density Fourier sampling. Using an idea of [26], we now consider a sampling strategy where the uniform random samples (which are sufficient to recover the gradient stably and robustly) are augmented by a set of variable density Fourier samples to enhance the image recovery. Following Definition 2.2, let $p = (p_\omega)$ be a probability distribution on $\{-N/2 + 1, \dots, N/2\}^d$. We also require several additional concepts. First, if $\omega \in \mathbb{R}$, we let $\bar{\omega} = \max\{1, |\omega|\}$. Second, if $\omega = (\omega_1, \dots, \omega_d) \in \mathbb{R}^d$, we let $\pi : \{1, \dots, d\} \rightarrow \{1, \dots, d\}$ be a bijection such that $\bar{\omega}_{\pi(1)} \geq \bar{\omega}_{\pi(2)} \geq \dots \geq \bar{\omega}_{\pi(d)}$. Next, we define $q = (q_\omega)$ by

$$(3.8) \quad q_\omega = \bar{\omega}_{\pi(1)} \cdots \bar{\omega}_{\pi(d/2)}, \quad d \text{ even},$$

and

$$(3.9) \quad q_\omega = \bar{\omega}_{\pi(1)} \cdots \bar{\omega}_{\pi((d-1)/2)} \sqrt{\bar{\omega}_{\pi((d+1)/2)}}, \quad d \text{ odd}.$$

Finally, we let $\Gamma(p)$ be the smallest positive constant such that

$$(3.10) \quad (q_\omega)^{-2} \leq \Gamma(p)p_\omega, \quad \forall \omega \in \{-N/2 + 1, \dots, N/2\}^d.$$

Notice that $\Gamma(p) \geq 1$, since p is a probability distribution and $q_0 = 1$.

Theorem 3.3 (Variable density Fourier sampling, one dimension). *Let $d = 1$, $0 < \varepsilon < 1$, $2 \leq s, m \leq N$ and $\Omega = \Omega_1 \cup \Omega_2 \subseteq \{-N/2 + 1, \dots, N/2\}$, where Ω_1 is a uniform random sampling scheme of order $m/2$ and Ω_2 is a variable density sampling scheme of order $m/2$ corresponding to a probability distribution $p = (p_\omega)$. Let $A = \frac{1}{\sqrt{m}}P_\Omega F \in \mathbb{C}^{m \times N}$ and*

$$(3.11) \quad m \gtrsim \Gamma(p) \cdot s \cdot \log(\Gamma(p)s) \cdot (\log(\Gamma(p)s) \cdot \log(N) + \log(2\varepsilon^{-1})).$$

Then the following holds with probability at least $1 - \varepsilon$. For all $x \in \mathbb{C}^N$ and $y = Ax + e \in \mathbb{C}^m$, where $\|e\|_{\ell^2} \leq \eta$ for some $\eta \geq 0$, every minimizer \hat{x} of (2.6) satisfies

$$(3.12) \quad \|\nabla x - \nabla \hat{x}\|_{\ell^2} \lesssim \frac{\sigma_s(\nabla x)_{\ell^1}}{\sqrt{s}} + \eta, \quad \|x - \hat{x}\|_{\text{TV}} \lesssim \sigma_s(\nabla x)_{\ell^1} + \sqrt{s}\eta,$$

and

$$(3.13) \quad \frac{\|x - \hat{x}\|_{\ell^2}}{\sqrt{N}} \lesssim \frac{\sigma_s(\nabla x)_{\ell^1}}{s} + \left(\sqrt{\frac{\Gamma(p)}{N}} + \frac{1}{\sqrt{s}} \right) \eta.$$

Theorem 3.4 (Variable density Fourier sampling, $d \geq 2$ dimension). *Let, $d \geq 2$, $0 < \varepsilon < 1$, $2 \leq s, m \leq N^d$ and $\Omega = \Omega_1 \cup \Omega_2$, where Ω_1 is a d -dimensional uniform random sampling pattern of order $m/2$ and Ω_2 is a variable density sampling scheme of order $m/2$ corresponding to a probability distribution $p = (p_\omega)$. Let $A = \frac{1}{\sqrt{m}}P_\Omega F \in \mathbb{C}^{m \times N^d}$ and*

$$(3.14) \quad m \gtrsim_d \Gamma(p) \cdot s \cdot \log^2(N) \cdot \log(\Gamma(p) \log(N)s) \cdot (\log(\Gamma(p) \log(N)s) \cdot \log(N) + \log(2\varepsilon^{-1})),$$

where $\Gamma(p)$ is as in (3.10). Then the following holds with probability at least $1 - \varepsilon$. For all $x \in \mathbb{C}^{N^d}$ and $y = Ax + e \in \mathbb{C}^m$, where $\|e\|_{\ell^2} \leq \eta$ for some $\eta \geq 0$, every minimizer \hat{x} of (2.6) satisfies

$$(3.15) \quad \|\nabla x - \nabla \hat{x}\|_{\ell^2} \lesssim \frac{\sigma_s(\nabla x)_{\ell^1}}{\sqrt{s}} + d\eta, \quad \|x - \hat{x}\|_{\text{TV}_a} \lesssim \sigma_s(\nabla x)_{\ell^1} + \sqrt{s}d\eta \quad (\text{anisotropic}),$$

$$(3.16) \quad \|\nabla \hat{x} - \nabla x\|_{\ell^{2,2}} \lesssim \frac{\sigma_s(\nabla x)_{\ell^{2,1}}}{\sqrt{s}} + \sqrt{d}\eta, \quad \|x - \hat{x}\|_{\text{TV}_i} \lesssim \sigma_s(\nabla x)_{\ell^{2,1}} + \sqrt{s}\sqrt{d}\eta \quad (\text{isotropic}),$$

and

$$(3.17) \quad \|\hat{x} - x\|_{\ell^2} \lesssim \frac{\sigma_s(\nabla x)_{\ell^1}}{\sqrt{s}} + \left(\sqrt{\Gamma(p)} + d\right)\eta \quad (\text{anisotropic}),$$

$$(3.18) \quad \|\hat{x} - x\|_{\ell^2} \lesssim \sqrt{d} \frac{\sigma_s(\nabla x)_{\ell^{2,1}}}{\sqrt{s}} + \left(\sqrt{\Gamma(p)} + d\right)\eta \quad (\text{isotropic}).$$

These results are general in the sense that they permit any variable density sampling scheme. Moreover, the effect of the density p is seen clearly through the constant $\Gamma(p)$: the smaller $\Gamma(p)$, the better the measurement conditions (3.11) and (3.14) and the image recovery bounds (3.13), (3.17) and (3.18). In the next section, we discuss the choice of p . Specifically, we identify densities for which $\Gamma(p)$ satisfies the optimal bound $\Gamma(p) \lesssim \log_d(N)$.

With this in mind, these results can be understood as follows. Suppose that p is chosen so that $\Gamma(p) \lesssim_d \log(N)$. Then by incorporating variable density samples we achieve better stability and robustness in the image recovery by a factor of \sqrt{s} over the case when only uniform random samples are used (Theorems 3.1 and 3.2). In particular, the image error bounds, up to the factor of $\Gamma(p)$, depend on $\sigma_s(\nabla x)_{\ell^1}/\sqrt{s}$ and η , exactly as in the gradient error bounds. Moreover, to achieve these estimates we need a number of measurements scaling linearly in s , up to log factors. We note also that the anisotropic and isotropic TV semi-norms give the same recovery guarantees, up to factors in d .

3.3. Discussion. To illustrate this difference, in Fig. 1 we compare the stability and robustness of the recovery of a two-dimensional image and its gradient. In this figure, we perturb either the image x (to study stability) or the measurements y (to study robustness) and compute the error in the reconstructed image and its gradient. We use the standard Shepp–Logan phantom, since its gradient is exactly sparse, and compare the recovery from uniform random and variable density samples.

For both types of perturbations, observe that the image recovery error is better for variable density samples than uniform random samples, whereas the gradient recovery errors are very similar. This confirms the results of the previous section, which assert that uniform random sampling provides adequate recovery of the image gradient, matching the stability and robustness of variable density sampling, but that the image recovery error is worse by a factor of \sqrt{s} .

The intuition for this discrepancy is quite straightforward. Since it has periodic boundary conditions, the gradient operator commutes with the DFT matrix (see Lemma 7.1). Hence recovery of the image gradient is equivalent to recovering a sparse vector from samples of its Fourier transform. It is well known that uniform random sampling is a suitable (in fact, optimal) sampling strategy for recovering a sparse vector from samples of its Fourier transform. Hence, we expect adequate recovery of the gradient from such measurements. On the other hand, since the constant vector lies in the null space of the gradient operator, it is impossible to recover x from ∇x . This is why the zero frequency is added in Theorems 3.1 and 3.2. However,

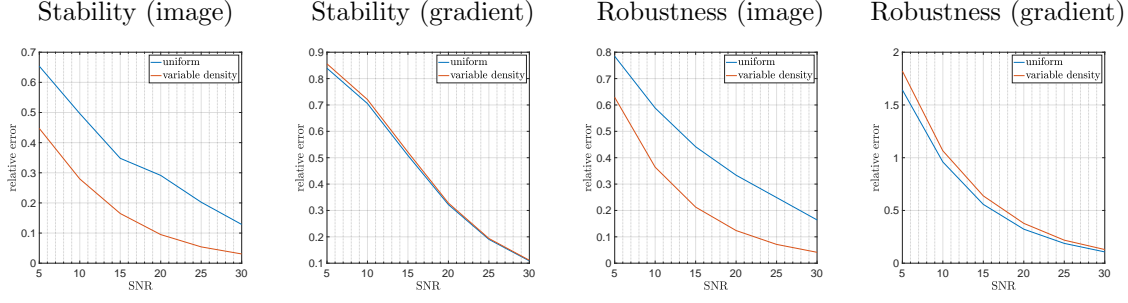


Figure 1. Recovery of the discrete 256^2 Shepp–Logan phantom from 25% Fourier measurements using either uniform random sampling or variable density sampling according to (4.6). The horizontal axis shows the signal-to-noise ratio (SNR) of the perturbation and the vertical axis shows the relative error in the recovered image or recovered image gradient. For the stability experiment (left), the image x is perturbed to $x + h$. The SNR and relative error are defined as $20 \log_{10}(\|x\|_{\ell^2}/\|h\|_{\ell^2})$ and $\|z - (x + h)\|_{\ell^2}/\|x + h\|_{\ell^2}$ or $\|\nabla(z - (x + h))\|_{\ell^2}/\|\nabla(x + h)\|_{\ell^2}$ respectively, where z is the reconstruction of $x + h$. For the robustness experiment (right), the measurements y are perturbed to $y + h$. The SNR and relative error are defined as $20 \log_{10}(\|y\|_{\ell^2}/\|h\|_{\ell^2})$ and $\|z - x\|_{\ell^2}/\|x\|_{\ell^2}$ or $\|\nabla(z - x)\|_{\ell^2}/\|\nabla(x)\|_{\ell^2}$ respectively, where z is the reconstruction obtained from measurements $y + h$.

the stability and robustness of the image recovery is worse, since the gradient operator ∇ is ill-conditioned for large N . In particular, smooth functions (i.e. image textures) lie approximately in its null space. Yet, the Fourier transform of a smooth function decays rapidly with increasing frequency. Hence, variable density sampling overcomes this issue by sampling more densely near the origin, thus stabilizing the recovery of the smooth image components.

4. Choice of Fourier sampling pattern. As noted above, Theorems 3.3 and 3.4 allow for any variable density sampling scheme. We now discuss this choice in more detail.

4.1. Theoretically-optimal sampling patterns. We commence by deriving sampling patterns that are theoretically optimal, in the sense that they give the optimal scaling of $\Gamma(p)$ with respect to N (for fixed d):

Lemma 4.1. *Let $p = (p_\omega)$ be a probability distribution and $\Gamma(p)$ be as in (3.10). Then $\Gamma(p) \gtrsim \log(N)$. Moreover, if*

$$p_\omega = \frac{C_{N,d}}{(q_\omega)^2}, \quad \omega \in \{-N/2 + 1, \dots, N/2\}^d,$$

where q_ω is as in (3.8)–(3.9), then $\Gamma(p) \lesssim_d \log(N)$.

Using this, we immediately deduce the following:

Corollary 4.2 (Theoretically-optimal variable density Fourier sampling, one dimension). *Consider the setup of Theorem 3.3 with $p = (p_\omega)$ given by*

$$(4.1) \quad p_\omega = \frac{C_N}{\max\{1, |\omega|\}}, \quad \omega \in \{-N/2 + 1, \dots, N/2\},$$

and $s \gtrsim \log(N)$. Then the conclusions of Theorem 3.3 hold (with $\Gamma(p) \lesssim \log(N)$ in the case

of (3.13), provided m satisfies

$$(4.2) \quad m \gtrsim s \cdot \log(s) \cdot \log(N) \cdot (\log(s) \cdot \log(N) + \log(2\varepsilon^{-1})).$$

Note that the condition $s \gtrsim \log(N)$ is imposed merely to simplify the measurement condition (it allows one to replace terms such as $\log(\log(N)s)$ by $\log(s)$). It is informative to compare this result with Theorem 3.1. The measurement condition (4.2) prescribes an additional $\log(N)$ samples over Theorem 3.1, taken according to the density (4.1). However, this leads to an improved signal recovery error of the form

$$(4.3) \quad \frac{\|\hat{x} - x\|_{\ell^2}}{\sqrt{N}} \lesssim \frac{\sigma_s(\nabla x)_{\ell^1}}{s} + \left(\sqrt{\frac{\log(N)}{N}} + \frac{1}{\sqrt{s}} \right) \eta.$$

Note that a nonuniform recovery guarantee of similar flavour to Corollary 4.2 was first proved in [26, Thm. 2.1]. Therein $m \gtrsim s \cdot \log(N) \cdot (1 + \log(\varepsilon^{-1}))$ samples taken in the same way (in particular, with the same variable density (4.1)) were shown to give a recovery error

$$\frac{\|\hat{x} - x\|_{\ell^2}}{\sqrt{N}} \lesssim \log^2(s) \log(N) \log(m) \left(\log(s) \log^{1/2}(m) \frac{\sigma_s(\nabla x)_{\ell^1}}{s} + \frac{1}{\sqrt{s}} \eta \right).$$

By contrast, Corollary 4.2 is a uniform recovery guarantee. While it imposes a more stringent measurement condition (4.2), specifically, by a factor of $\log^2(s) \log(N)$, it leads to an improved recovery guarantee (4.3). For instance, the best s -term approximation error term $\sigma_s(\nabla x)_{\ell^1}/s$ is improved by a factor of $\log^3(s) \log(N) \log^{3/2}(m)$.

Corollary 4.3 (Theoretically-optimal variable density Fourier sampling, two dimensions). *Let $d = 2$ and consider the setup of Theorem 3.4 with $p = (p_\omega)$ given by*

$$(4.4) \quad p_\omega = \frac{C_N}{(\max\{1, |\omega_1|, |\omega_2|\})^2}, \quad \omega = (\omega_1, \omega_2) \in \{-N/2 + 1, \dots, N/2\}^2,$$

and $s \gtrsim \log(N)$. Then the conclusions of Theorem 3.4 hold (with $\Gamma(p) \lesssim \log(N)$ in the case of (3.17) and (3.18)), provided m satisfies

$$(4.5) \quad m \gtrsim s \cdot \log(s) \cdot \log^3(N) \cdot (\log(s) \cdot \log(N) + \log(2\varepsilon^{-1}))$$

Furthermore, the same conclusion holds (with possibly different numerical constant) if (4.4) is replaced by

$$(4.6) \quad p_\omega = \frac{C_N}{1 + (\omega_1)^2 + (\omega_2)^2}, \quad \omega = (\omega_1, \omega_2) \in \{-N/2 + 1, \dots, N/2\}^2,$$

or more generally, if $\|\cdot\|$ is any norm on \mathbb{R}^2 , by

$$(4.7) \quad p_\omega = \frac{C_N}{1 + \|\omega\|^2}, \quad \omega \in \{-N/2 + 1, \dots, N/2\}^2.$$

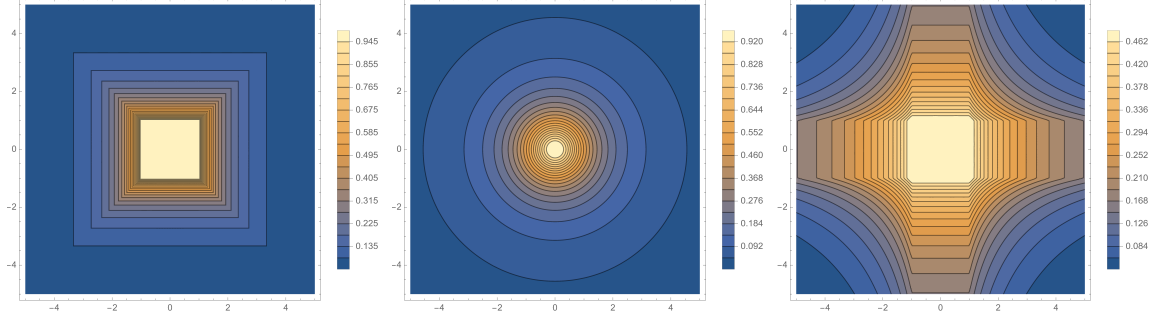


Figure 2. Level curves for the 2D (left) theoretically optimal (4.4), (middle) inverse square (4.6) and (right) hyperbolic cross (4.13) densities.

Note that (4.4) follows immediately from the observation that $q_\omega = \max\{1, |\omega_1|, |\omega_2|\}$ when $d = 2$. The results for (4.6) and (4.7) follow in turn simply because of the equivalence of norms on a finite-dimensional vector space.

The scheme (4.6) is known as *inverse square law* sampling. It is a standard and well-known variable density sampling strategy for compressed sensing recovery from Fourier measurements [20, 26]. Interesting, this result also shows that there are many different sampling strategies that give the same recovery guarantees up to constants. The critical factor is the asymptotic decay rate as $\omega \rightarrow \infty$. Fig. 2 visualizes the level curves of several such sampling strategies. Notice that the schemes (4.7) depend on the distance of ω from the zero frequency (with respect to some norm). We therefore informally refer to them as *radially symmetric*.

Similar results to Corollary 4.3 were shown in [20, 26]. In [20, Thm. 1] a uniform recovery guarantee was proved for inverse square law sampling (4.6), with the measurement condition

$$(4.8) \quad m \gtrsim s \cdot \log^3(s) \cdot \log^5(N),$$

implying a image recovery bound

$$\|x - \hat{x}\|_{\ell^2} \lesssim \frac{\sigma_s(\nabla x)_{\ell^1}}{\sqrt{s}} + \eta,$$

for the anisotropic TV semi-norm with a particular probability, where η is a bound for the noise in a certain weighted ℓ^2 -norm. Corollary 4.3 improves on this result in several ways. First, the log factors in the measurement condition (4.5) are reduced by a factor of $\log(s) \cdot \log(N)$ over (4.8). Second, this result gives a robustness bound where the noise is measured in an unweighted ℓ^2 -norm. Third, this result establishes the same recovery guarantee for the family of sampling schemes (4.7), as opposed to just the inverse square law (4.6).

On the other hand, a nonuniform recovery guarantee was shown in [26, Thm. 2.2]. Therein

$$(4.9) \quad m \gtrsim s \cdot \log(N) \cdot (1 + \log(\varepsilon^{-1})),$$

taken in the same way (in particular, using the inverse square law) were shown to yield a image recovery bound of the form

$$\|\hat{x} - x\|_{\ell^2} \lesssim \log(s) \log(N^2/s) \log^{1/2}(N) \log^{1/2}(m) \left(\log^{1/2}(m) \log(s) \frac{\sigma_s(\nabla x)_{\ell^{2,1}}}{\sqrt{s}} + \eta \right),$$

for the isotropic TV semi-norm. In comparison, Corollary 4.3 is a uniform recovery guarantee with an image recovery error bound of the form

$$\|\hat{x} - x\|_{\ell^2} \lesssim \frac{\sigma_s(\nabla x)_{\ell^{2,1}}}{\sqrt{s}} + \sqrt{\log(N)\eta},$$

for the isotropic TV semi-norm. As in the one-dimensional case, the tradeoff for the worse log term in the measurement condition (4.5) (by a factor of $\log^2(s)\log^3(N)$ over (4.9)) is a better image recovery bound by several log factors.

Finally, we consider the case of $d \geq 2$ dimensions. Note that this problem was not considered in either [20] or [26]:

Corollary 4.4 (Theoretically-optimal variable density Fourier samples, $d \geq 2$ dimensions). *Let $d \geq 2$ and consider the setup of Theorem 3.4 with $p = (p_\omega)$ given by*

$$(4.10) \quad p_\omega = \frac{C_{N,d}}{(q_\omega)^2}, \quad \omega \in \{-N/2 + 1, \dots, N/2\}^d,$$

and $s \gtrsim \log(N)$. Then the conclusions of Theorem 3.4 hold (with $\Gamma(p) \lesssim_d \log(N)$ in the case of (3.17) and (3.18)), provided m satisfies

$$(4.11) \quad m \gtrsim_d s \cdot \log(s) \cdot \log^3(N) \cdot (\log(s) \cdot \log(N) + \log(2\varepsilon^{-1})).$$

In particular, when $d = 3$, (4.10) can be expressed as

$$(4.12) \quad p_\omega = C_N \left(\left(\max_{i=1,2,3} \{\bar{\omega}_i\} \right)^2 \left(\sum_{i=1}^3 \bar{\omega}_i - \max_{i=1,2,3} \{\bar{\omega}_i\} - \min_{i=1,2,3} \{\bar{\omega}_i\} \right) \right)^{-1}.$$

Several remarks are in order. First, the measurement condition (4.11) and recovery error bounds are exactly the same as the two-dimensional measurement condition (4.5) and error bounds, except possibly for d -dependent constants. Second, as shown by (4.12), theoretically-optimal sampling strategies cease to be radially-symmetric in $d \geq 3$ dimensions. We shall discuss this further in the next section. But first, it is interesting to visualize the shape of the density (4.12). Fig. 3 plots a typical level set of this function. We observe in particular the axis-aligned spikes, and the nonsmooth transitions along the edges of the cube.

4.2. Sub-optimality of radially-symmetric sampling. As shown in Corollary 4.3, radially-symmetric sampling schemes are theoretically optimal in $d = 2$ dimensions. We now show that this ceases to be the case when $d \geq 3$.

Lemma 4.5. *Let $d \geq 2$ and $p = (p_\omega)$ be defined by*

$$p_\omega = \frac{C_{N,d,\alpha}}{(1 + \|\omega\|)^\alpha}, \quad \omega \in \{-N/2 + 1, \dots, N/2\}^d,$$

where $\|\cdot\|$ is any norm on \mathbb{R}^d and $\alpha > 0$. Then

$$\Gamma(p) \asymp_{d,\alpha} \begin{cases} N^{d-\alpha} & \alpha < 2 \\ N^{d-2} & 2 \leq \alpha < d \\ N^{d-2} \log(N) & \alpha = d \\ N^{\alpha-2} & \alpha > d \end{cases},$$

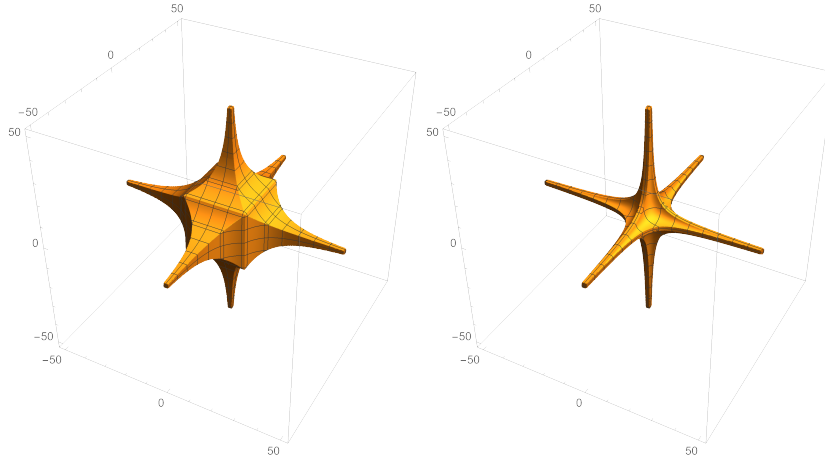


Figure 3. Level sets of the (left) theoretically-optimal (4.12) and (right) hyperbolic cross (4.13) densities in three dimensions.

(note that the second case is only possible when $d \geq 3$). In particular, the best scaling for $\Gamma(p)$ is $\Gamma(p) \asymp \log(N)$ when $d = 2$ and $\Gamma(p) \asymp N^{d-2}$ when $d \geq 3$, and these correspond to the choice $\alpha = 2$.

In particular, this result means that in $d = 3$ dimensions any radially-symmetric sampling pattern will yield a measurement condition that scales linearly with N . This, in view of Corollary 4.4 is theoretically suboptimal.

Remark 4.6 (Why radially-symmetric Fourier sampling is suboptimal). This arises from the proof of Theorem 3.4, which, following [24, 25], relies on Haar wavelets. This proof relates the recovery properties of a variable-density scheme for gradient sparse images to its recovery properties for images which are sparse in the discrete Haar wavelet basis. The study of Fourier sampling with wavelets has been considered extensively in [4, 20, 21] and elsewhere. In essence, the optimal variable density scheme is determined by the behaviour of Haar wavelets in frequency space. In one or two dimensions, the Fourier transform of a Haar wavelet decays sufficiently rapidly in all directions to allow for radially-symmetric sampling strategies to be optimal. However, as shown in [1], in three or more dimensions, the slow decay of the Fourier transform of a multi-dimensional Haar wavelet means that the optimal sampling scheme is no longer, as termed therein, *isotropic* (i.e. radially symmetric), but rather *anisotropic*, similar to what is described in Corollary 4.4.

4.3. Near-optimal sampling using hyperbolic cross densities. In $d \geq 3$ dimensions, it is interesting to determine other densities which offer theoretically optimal or near-optimal performance. As seen in Fig. 3, the three-dimensional theoretically-optimal density (4.12) has level curves that fail to be smooth at certain points. To conclude this section, we now identify a different density possessing smooth level curves which is optimal up to the log factor. This is based on hyperbolic cross sampling:

Corollary 4.7 (Near-optimal hyperbolic cross Fourier sampling, $d \geq 2$ dimensions). *Let $d \geq 2$ and consider the setup of Theorem 3.4 with $p = (p_\omega)$ given by*

$$(4.13) \quad p_\omega = \frac{C_{N,d}}{\bar{\omega}_1 \cdots \bar{\omega}_d}, \quad \omega \in \{-N/2 + 1, \dots, N/2\}^d,$$

and $s \gtrsim \log(N)$. Then the conclusions of Theorem 3.4 hold (with $\Gamma(p) \lesssim_d \log^d(N)$ in the case of (3.17) and (3.18)), provided m satisfies

$$(4.14) \quad m \gtrsim_d s \cdot \log(s) \cdot \log^{d+2}(N) \cdot (\log(s) \cdot \log(N) + \log(2\varepsilon^{-1})).$$

This result shows that hyperbolic cross sampling is near optimal. In particular, the measurement condition (4.14) is worse than the optimal condition (4.11) only by a factor of $\log^{d-1}(N)$. Fig. 2 plots the level curves of two-dimensional hyperbolic cross sampling and Fig. 3 shows a three-dimensional level set. Notice that this strategy mimics the spikes of the theoretically-optimal pattern, but is less dense near the centre. However, it is a smooth function of $\bar{\omega}_1, \dots, \bar{\omega}_d$, unlike in the case of the latter. We note in passing that the hyperbolic cross is a well-known object in multivariate approximation theory [28], where it is used to overcome the curse of dimensionality.

5. Main results on Walsh sampling. We now consider Walsh sampling. The major difference between this and the previous case is that the Walsh–Hadamard transform does not commute with the discrete gradient operator. For this reason, we do not provide gradient recovery estimates, we only consider variable density sampling and we assume throughout that $d \geq 2$ (see §8 for some further discussion on this point). For simplicity, we state our results for anisotropic TV only in this section. However, results for isotropic TV can be readily proved as well.

Recall from §2.4 that Walsh frequencies are indexed over $\{0, \dots, N-1\}^d$. Thus, we now consider variable density sampling according to probability distributions $p = (p_i)_{i \in \{0, \dots, N-1\}^d}$ over this set. We let $\Gamma(p) \geq 0$ be the smallest constant such that

$$(5.1) \quad \left(1 + \|i\|_{\ell^\infty}^d\right)^{-1} \leq \Gamma(p)p_i, \quad \forall i \in \{0, \dots, N-1\}^d.$$

Once more we notice that $\Gamma(p) \geq 1$, since the p is a probability distribution and the left-hand side is equal to one when $i = (0, \dots, 0)$. Our main result is the following:

Theorem 5.1 (Variable density Walsh sampling, $d \geq 2$ dimensions). *Let $d \geq 2$, $0 < \varepsilon < 1$, $2 \leq s, m \leq N^d$ and $\Omega \subseteq \{0, \dots, N-1\}^d$ be a variable density sampling scheme of order m corresponding to a probability distribution $p = (p_i)$. Let $A = \frac{1}{\sqrt{m}}P_\Omega H$ and suppose that*

$$m \gtrsim_d \Gamma(p) \cdot s \cdot \log^2(N/s) \cdot \log(N) \cdot \log(\Gamma(p)s) \cdot (\log(\Gamma(p)s) \cdot \log(N) + \log(\varepsilon^{-1})),$$

where $\Gamma(p)$ is as in (5.1). Then the following holds with probability at least $1 - \varepsilon$. For all $x \in \mathbb{C}^{N^d}$ and $y = Ax + e \in \mathbb{C}^m$, where $\|e\|_{\ell^2} \leq \eta$ for some $\eta \geq 0$, every minimizer \hat{x} of (2.6) satisfies

$$(5.2) \quad \|x - \hat{x}\|_{\ell^2} \lesssim \frac{\sigma_s(\nabla x)_{\ell^1}}{\sqrt{s \log(N)}} + \sqrt{\Gamma(p)}\eta.$$

Similar to Fourier sampling, this result asserts stable and robust recovery of the image x from Walsh measurements, up to log factors, taken according to the appropriate variable density strategy. We now consider the choice of sampling strategy:

Lemma 5.2. *Let $p = (p_i)$ be a probability distribution and $\Gamma(p)$ be as in (5.1). Then $\Gamma(p) \gtrsim_d \log(N)$. Moreover, if*

$$p_i = \frac{C_{N,d}}{1 + \|i\|^d}, \quad i \in \{0, \dots, N-1\}^d,$$

where $\|\cdot\|$ is any norm, then $\Gamma(p) \lesssim_d \log(N)$.

Corollary 5.3 (Theoretically-optimal variable density Walsh sampling, $d \geq 2$ dimensions). *Consider the setup of Theorem 5.1 with $p = (p_i)$ given by*

$$p_i = \frac{C_{N,d}}{1 + \|i\|^d}, \quad i \in \{0, \dots, N-1\}^d,$$

where $\|\cdot\|$ is any norm on \mathbb{R}^d , and $s \gtrsim \log(N)$. Then the conclusions of Theorem 5.1 hold (with $\Gamma(p) \lesssim_d \log(N)$ in (5.2)), provided m satisfies

$$m \gtrsim_d s \cdot \log(s) \cdot \log^2(N/s) \cdot \log^2(N) (\log(s) \cdot \log(N) + \log(\varepsilon^{-1})).$$

Much like with Fourier sampling (Corollary 4.4), this result asserts a class of theoretically-optimal sampling strategies which ensure stable and robust recovery in $d \geq 2$ dimensions from Walsh measurements. We are unaware of any similar result in the literature. It is notable, however, that the optimal sampling strategy is radially symmetric in all dimensions, unlike in the Fourier case. See Remark 5.4 below. We also note that the log term in Corollary 5.3 is worse by a factor of $\log^2(N/s)/\log(N)$ than that of Corollary 4.4. This stems from the proof strategy, and specifically the different technique that is used in the Walsh case in the absence of the commuting property.

Remark 5.4. Similar to the Fourier case (Remark 4.6), the explanation for why radially-symmetric sampling works in any dimensions for Walsh sampling can be traced to the use of Haar wavelets in the proof. Haar wavelets and Walsh functions are intimately related, see (8.5). This means that the Walsh transform of a Haar wavelet behaves far more nicely than its Fourier transform, which in turn allows one to use a radially-symmetric sampling pattern in any dimension. See also [1]. By contrast, as shown in §4.2 the use of a radially-symmetric sampling pattern in the Fourier case leads to a measurement condition with a factor of N^{d-2} .

6. Experiments and discussion. We now show a series of further numerical experiments.

6.1. Experimental setup. We first describe the details of these experiments. We focus on reconstructing either three-dimensional MRI or test data, Fig. 4, with Fourier sampling or two-dimensional natural images with Walsh sampling, Fig. 5. For each of our experiments, we run 20 trials of reconstructing the given image using a modified version of the NESTA solver [8] which allows for reconstruction of two- or three-dimensional images via TV-minimization. The NESTA parameters used are designed for images whose values lie in the range $[0, 100]$,

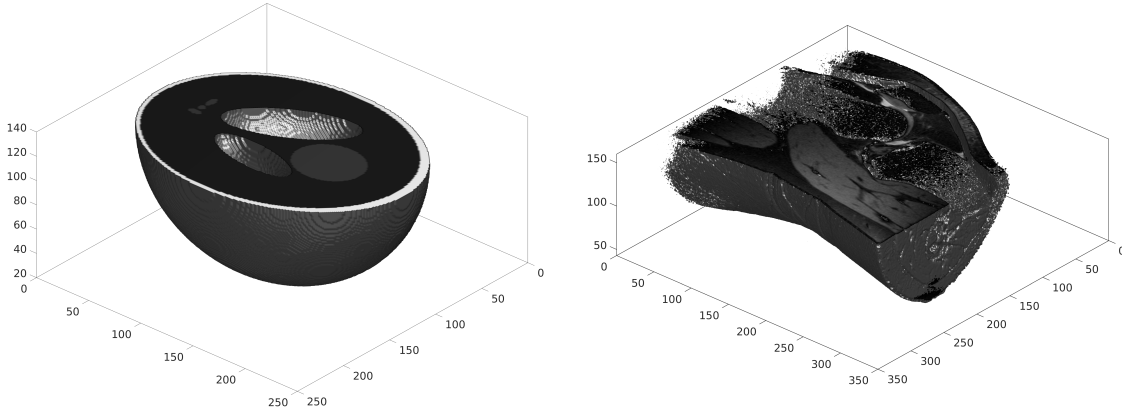


Figure 4. The (left) Shepp-Logan phantom (size 256^3) generated with <https://www.mathworks.com/matlabcentral/fileexchange/9416-3d-shepp-logan-phantom> and (right) “knee MRI” (size $320^2 \times 256$) three-dimensional test images for Fourier sampling. The “knee MRI” test image is generated from the MRI data from case 11 of the “Stanford Fullysampled 3D FSE Knees” dataset available at <https://mridata.org>, and was zero-padded to obtain a test image of size 320^3 .



Figure 5. The (left) “cameraman” (size 256^2), (middle) “donkey” (size 512^2) and (right) “man” (available in sizes 256^2 , 512^2 and 1024^2) test images for Walsh-Hadamard sampling.

and therefore we rescale all images to this range. These parameters are $\mu = 0.2$, 5 outer iterations, 5000 inner iterations, a tolerance of 10^{-5} and $\delta = 10^{-5}$. We run 20 random trials, each with a different seed, and plot the average PSNR values.

We consider six sampling patterns, four of which have already been introduced in this paper. These are: *uniform random*, *hyperbolic cross* (4.13), the *theoretically-optimal* pattern (see Corollaries 4.4 and 5.3 for Fourier and Walsh-Hadamard respectively) and the *inverse square law*. We also consider two further sampling patterns, *half-half* sampling and *multilevel random* subsampling. The former fully samples the lowest $m/2$ frequencies and then randomly subsamples the remainder. The latter was introduced in [4]. In this scheme, one first divides frequency space into r annular regions B_1, \dots, B_r of equal width. Next, one defines a

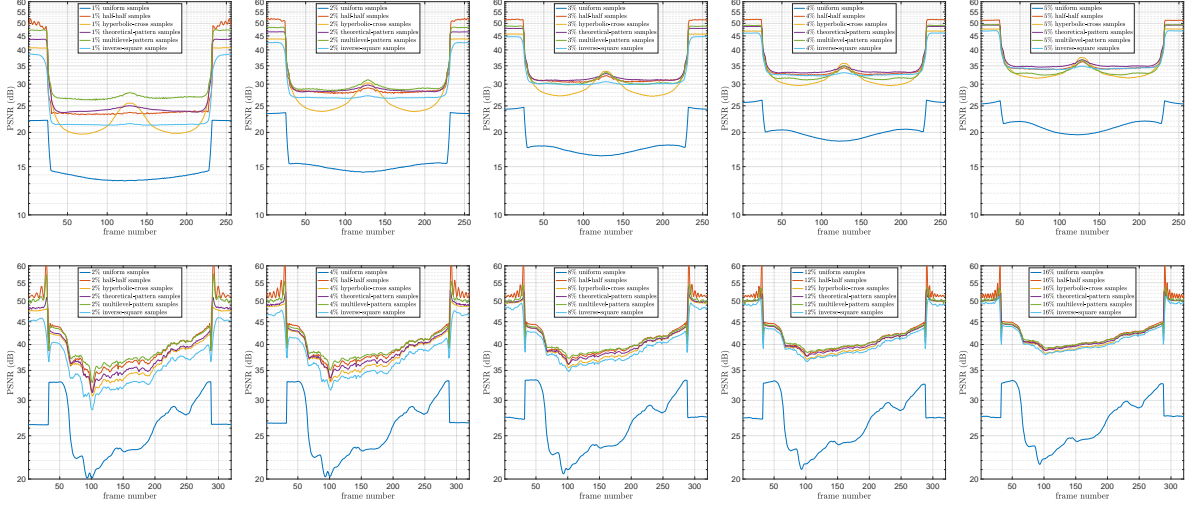


Figure 6. Average PSNR values over 20 trials in reconstructing each frame of the Shepp-Logan phantom (top) and knee MRI (bottom) test image from Fig. 4 with different Fourier sampling strategies as the sampling percentage is increased from left to right. Here the multilevel sampling is performed with $a = 1$, $r = 20$ and $r_0 = 1$ (top) or $r_0 = 2$ (bottom).

decreasing sampling fraction $p_k = m_k/|B_k|$ as

$$p_k = 1, \quad k = 1, \dots, r_0, \quad p_k = \exp\left(-\left(\frac{b(k-r_0)}{r-r_0}\right)^a\right), \quad k = r_0 + 1, \dots, r,$$

where r_0 and a are parameters, and b is chosen so that $\sum_{k=1}^r m_k = m$. Finally, within each region B_k one selects m_k samples uniformly and randomly. We refer to [4] for further details.

6.2. Fourier sampling. Fig. 6 displays the PSNR values for reconstructing the two Fourier test images shown in Fig. 4. Note that the reconstruction is performed in three dimensions, while the Fig. shows the PSNR versus frame number in the z -direction.

As expected, uniform random sampling performs very poorly in comparison to all other schemes. Similar, as predicted in §4.2, the inverse-square law generally performs relatively poorly in comparison to the others, especially for the more complicated knee MRI image.

Interestingly, the multilevel scheme performs amongst the best, especially at low sampling percentages. Often, it outperforms the theoretically-optimal pattern. This is in spite of the fact that the multilevel scheme is radially symmetric, whereas it was argued in §4.2 radially-symmetric patterns, at least those that draw samples from a single density, are theoretically suboptimal in three dimensions.

Typically, in the experiments, the second and third best performers are the theoretically-optimal and half-half schemes. It should come as little surprise that the latter performs worse than multilevel random sampling: full sampling followed by uniform random sampling is a relatively crude strategy. Interestingly, the behaviour of the hyperbolic cross scheme is much more heavily dependent on the frame for the Shepp-Logan phantom – it is clearly too anisotropic to recover the details in some of the frames – than the other patterns. But its

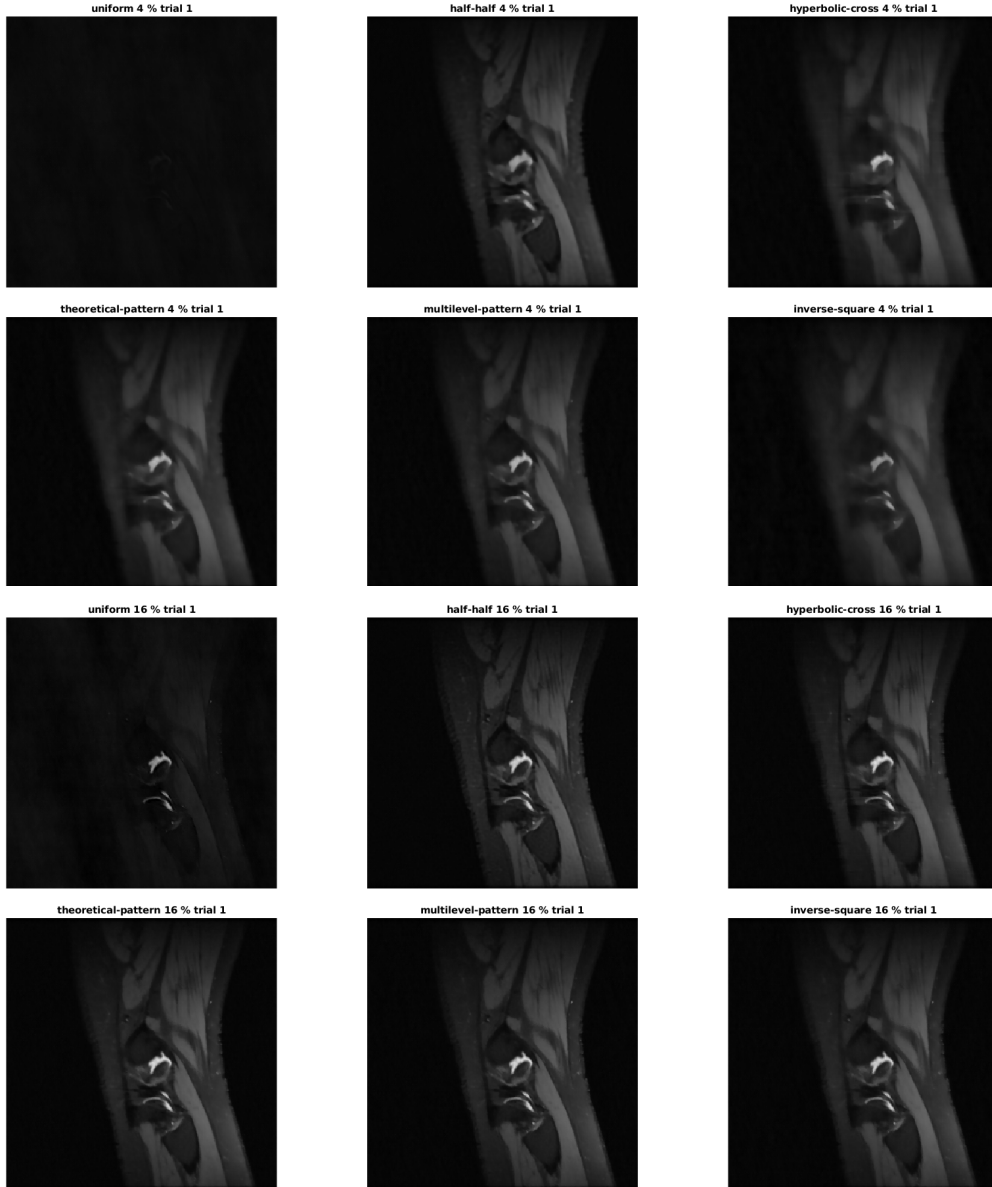


Figure 7. Comparison of reconstructions from trial 1 of 20 of frame 102 from the zero-padded “knee MRI” data with each method at (rows 1 & 2) 4% and (rows 3 & 4) 16% subsampling.

relative frame-by-frame performance on the knee MRI image is similar to the other patterns.

In Fig. 7 we show the recovery of an individual frame for two different sampling percentages. In both cases the half-half and multilevel patterns give a slightly sharper image in comparison to the theoretical pattern, which is slightly more blurred. As one would expect, the hyperbolic cross and inverse-square law both present substantial additional artefacts.

6.3. Walsh sampling. In Fig. 8 we consider two-dimensional Walsh sampling for the images in Fig. 5. Across all images and all sampling percentages, the multilevel scheme

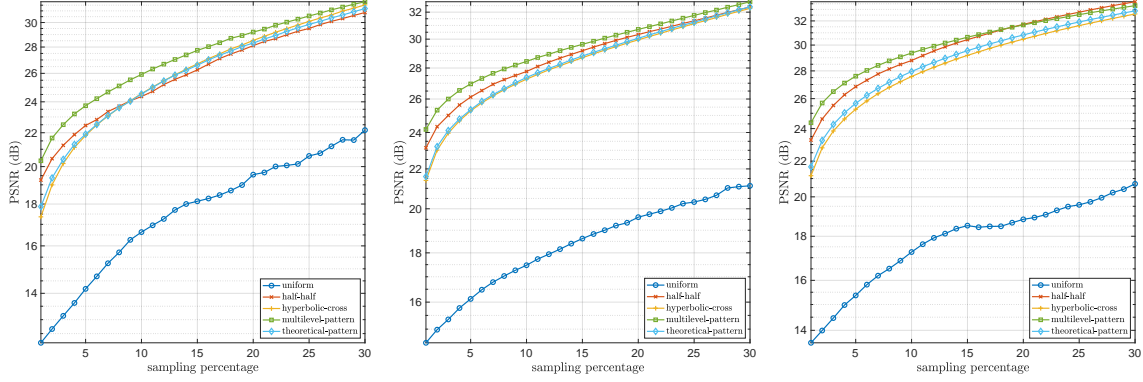


Figure 8. Comparison of the average PSNR values over 20 trials for various sampling patterns in reconstructing the (left) “cameraman,” (middle) “donkey,” and (right) “man” test images with Walsh sampling. Here the multilevel sampling is performed with $a = 2$, $r = 30$ and $r_0 = 2$.

consistently performs amongst the best, with generally the theoretically-optimal or half-half pattern performing second best. The relative performance of the half-half scheme is quite heavily dependent on the image, with it performing worse on the “cameraman” image but better on the “donkey” and “man” images. This is not surprising. The “cameraman” image is relatively simple, meaning the half-half scheme likely oversamples the high frequency regime. Conversely, the “donkey” and “man” images are more complex, meaning more sampling is needed at higher frequencies to resolve the fine details. This effect can be further examined by considering the “man” image at different resolutions, as we do in Fig. 9. At low resolution the half-half scheme actually outperforms the multilevel scheme whenever the sampling percentage is greater than 12%, whereas at higher resolution this only occurs after 21%. This can once more be traced to the properties of the image. At low resolution, the edges of the image are relatively closer together, thus requiring more higher-frequency samples to resolve, whereas at higher resolutions they are relatively better separated.

This observation is related to the previous discussion in §1.3. As originally considered in [26], the optimal sampling strategy in practice depends on the image, resolution and sampling percentage – in particular, the geometry of its edges. This is not reflected in our theoretically-optimal sampling strategies (which are image independent). Yet it is notable that good all-round performance can be achieved with the multilevel random sampling strategy.

7. Proofs Part I: Theorems 3.1–3.4. The remainder of this paper is devoted to the proofs of the main results. We divide this into two sections: Fourier sampling in this section and Walsh sampling in the next. Note that in both these sections we rely on some background results which are found in the Supplementary Materials. In §C we also prove several of the ancillary lemmas stated in previously.

7.1. Overview. Our proof is divided into three parts. First, in §7.2, we assert stable and robust recovery of the gradient ∇x . Following [26], this made possible by the uniform random samples Ω_1 and relies crucially on the commuting property of the Fourier transform and the gradient operator (Lemma 7.1).

Next, in the §7.3 and §7.4, we address the recovery of the image itself. In the case of

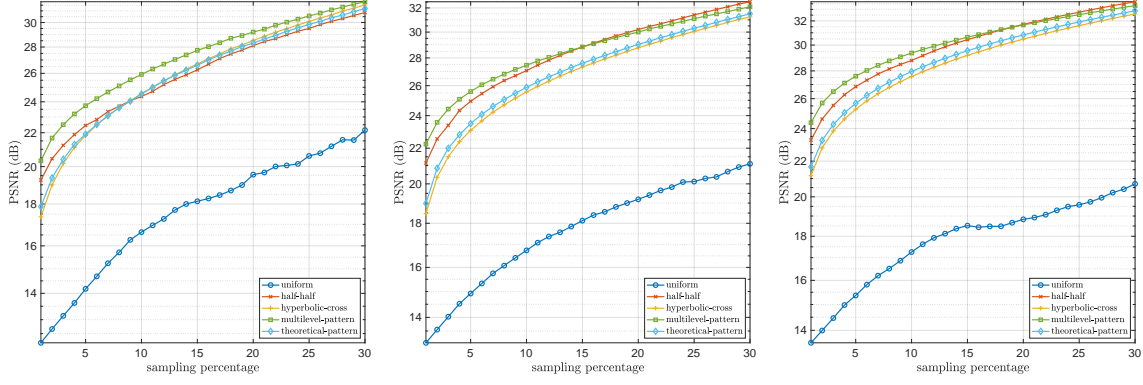


Figure 9. Comparison of the average PSNR values over 20 trials for various sampling patterns in reconstructing the “man” test image with (left) $N = 256$ (middle) $N = 512$, and (right) $N = 1024$ with Walsh sampling. Here the multilevel sampling is performed with $a = 2$, $r = 30$ and $r_0 = 2$.

uniform random sampling, we follow [26] and use the following discrete *Poincaré inequality*

$$(7.1) \quad \begin{aligned} \|z\|_{\ell^2} &\leq \|Az\|_{\ell^2} + \sqrt{N}\|z\|_{TV}, \quad \forall z \in \mathbb{C}^N, \\ \|z\|_{\ell^2} &\leq \|Az\|_{\ell^2} + 2^{1-d/2}\|z\|_{TV_a}, \quad \forall z \in \mathbb{C}^{N^d}. \end{aligned}$$

See Lemma 7.3. The estimates for $\|\hat{x} - x\|_{\ell^2}$ then follow by setting $z = \hat{x} - x$ and using the existing gradient error bounds. For variable density sampling in §7.4, based on ideas of [24, 25], we derive a strengthened Poincaré inequality for any measurement matrix that is incoherent with Haar wavelets (Lemma 7.4). The rest of the proof is then devoted to showing that variable density Fourier samples are sufficiently incoherent with Haar wavelets. For this we use tools from §A.2.

7.2. Gradient recovery. In this section, we prove the error bounds (3.1), (3.4), (3.5), (3.12), (3.15) and (3.16) for gradient recovery using uniform random and variable density Fourier sampling. This relies on the commuting property:

Lemma 7.1 (Commuting property). *Let $F^{(d)}$ be the d -dimensional DFT matrix and ∇_j be the j^{th} partial derivative operator. Then*

$$F^{(d)}\nabla_j = (I^{(d-j)} \otimes \text{diag}(\lambda) \otimes I^{(j-1)})F^{(d)},$$

where $\lambda = (\lambda_j)_{j=1}^N \in \mathbb{C}^N$ has entries $\lambda_j = \exp(2\pi i \varrho(j)/N) - 1$ and ϱ is defined in (2.1). $I^{(d-j)} = \underbrace{I \otimes I \otimes \dots \otimes I}_{d-j}$, $I^{(j-1)} = \underbrace{I \otimes \dots \otimes I}_{j-1}$, and $I \in \mathbb{C}^{N \times N}$ is the identity matrix.

Proof. The $d = 1$ case is a simple exercise. Now consider the $d \geq 2$ case. We have

$$\begin{aligned} F^{(d)}\nabla_j &= \underbrace{(F \otimes F \otimes \dots \otimes F)}_d \underbrace{(I \otimes I \otimes \dots \otimes I)}_{d-j} \nabla_j \underbrace{(I \otimes \dots \otimes I)}_{j-1} \\ &= F^{(d-j)} \otimes ((\text{diag}(\lambda)F) \otimes (F^{(j-1)}I^{(j-1)})) \\ &= (I^{(d-j)}F^{(d-j)}) \otimes (\text{diag}(\lambda) \otimes I^{(j-1)}F^{(j)}) = I^{(d-j)} \otimes \text{diag}(\lambda) \otimes I^{(j-1)}F^{(d)}, \end{aligned}$$

as required. ■

Next, since the sampling map has the form $\Omega = \Omega_1 \cup \Omega_2$ in all cases, we can write A as

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \quad A_i = \frac{1}{\sqrt{m}} P_{\Omega_i} F, \quad i = 1, 2.$$

The matrix $N^{-d/2}F$ is unitary, and therefore A_1 is a randomly-sampled unitary matrix (see §A) with the uniform probability distribution $q = (1/N^d)_{i=1}^{N^d}$. Since $|F_{jk}| = 1$, the bounded orthonormal system constant $\Theta = 1$. Hence, by (A.3), A_1 satisfies the RIP of order $2s$ with $\delta_{2s} \leq 1/2$ (this factor is arbitrary, any number less than $4/\sqrt{41}$ will do) and probability at least $1 - \varepsilon$, provided (after simplifying the log factor using the fact that $N^d \geq s \geq 2$),

$$m \gtrsim s \cdot \log(s) \cdot (\log(s) \cdot \log(N^d) + \log(\varepsilon^{-1})).$$

For the next steps of the proof, we split into the anisotropic and isotropic cases.

7.2.1. Anisotropic TV: (3.1), (3.12), (3.4) and (3.15).

Proof of (3.1), (3.12), (3.4) and (3.15). We use Lemma A.4 Let $B = B_1$ be as in this lemma with $A = A_1$. Since A_1 has the RIP of order $2s$ with constant $\delta_{2s} \leq 1/2$, it also has the rNSP of order s . We now apply Lemma A.2 to B_1 with the vectors $\nabla \hat{x}$ and ∇x and use the fact that $\|\nabla \hat{x}\|_1 = \|\hat{x}\|_{\text{TV}_a} \leq \|x\|_{\text{TV}_a} = \|\nabla x\|_{\ell^1}$ to get

$$\begin{aligned} \|\nabla \hat{x} - \nabla x\|_{\ell^1} &\lesssim \sigma_s(\nabla x)_{\ell^1} + \sqrt{sd} \|B_1 \nabla(\hat{x} - x)\|_{\ell^2}, \\ \|\nabla \hat{x} - \nabla x\|_{\ell^2} &\lesssim \frac{\sigma_s(\nabla x)_{\ell^1}}{\sqrt{s}} + \sqrt{d} \|B_1 \nabla(\hat{x} - x)\|_{\ell^2}. \end{aligned}$$

For the second term, we use the fact that $F^{(d)} \nabla_i = (I^{(d-i)} \otimes \text{diag}(\lambda) \otimes I^{(i-1)}) F^{(d)}$ (Lemma 7.1) and the bound $\|\lambda\|_{\ell^\infty} \leq 2$ to get

$$\begin{aligned} \|B_1 \nabla(\hat{x} - x)\|_{\ell^2} &= \sqrt{\sum_{i=1}^d \|A_1 \nabla_i(\hat{x} - x)\|_{\ell^2}^2} = \sqrt{\sum_{i=1}^d \left\| \frac{1}{\sqrt{m}} P_{\Omega_1} F^{(d)} \nabla_i(\hat{x} - x) \right\|_{\ell^2}^2} \\ &\leq 2\sqrt{d} \left\| \frac{1}{\sqrt{m}} P_{\Omega_1} F^{(d)}(\hat{x} - x) \right\|_{\ell^2} \\ &= 2\sqrt{d} \|A_1(\hat{x} - x)\|_{\ell^2} \leq 2\sqrt{d} \|A(\hat{x} - x)\|_{\ell^2} \leq 4\sqrt{d}\eta. \end{aligned}$$

Note that in the last step we have used the fact that \hat{x} and x are feasible for (2.6). Substituting this into the previous estimates now gives the result. ■

7.2.2. Isotropic TV: (3.5) and (3.16). For isotropic TV, we use the matrix recovery techniques from §A.3.

Proof of (3.5) and (3.16). The matrix A_1 satisfies the RIP of order $2s$ with constant $\delta_{2s} \leq 1/2$. Hence, by [15, Prop. 4.3] it also has the $\ell^{2,2}$ -rNSP of order s with constants ρ and γ depending on δ_{2s} . Using $\nabla \hat{x}$ and ∇x in Lemma A.6, we get

$$(7.2) \quad \|\nabla \hat{x} - \nabla x\|_{\ell^{2,1}} \lesssim \sigma_s(\nabla x)_{\ell^{2,1}} + \sqrt{s} \|A_1 \nabla(\hat{x} - x)\|_{\ell^{2,2}},$$

$$(7.3) \quad \|\nabla \hat{x} - \nabla x\|_{\ell^{2,2}} \lesssim \frac{\sigma_s(\nabla x)_{\ell^{2,1}}}{\sqrt{s}} + \|A_1 \nabla(\hat{x} - x)\|_{\ell^{2,2}}.$$

For the second term of (7.2) and (7.3), by the commuting property, we have,

$$\begin{aligned} \|A_1 \nabla(\hat{x} - x)\|_{\ell^{2,2}} &= \|(A_1 \nabla_1(\hat{x} - x) \cdots A_1 \nabla_d(\hat{x} - x))\|_{\ell^{2,2}} \\ &= \left\| \left(\frac{1}{\sqrt{m}} P_{\Omega_1} F^{(d)} \nabla_1(\hat{x} - x) \cdots \frac{1}{\sqrt{m}} P_{\Omega_1} F^{(d)} \nabla_d(\hat{x} - x) \right) \right\|_{\ell^{2,2}} \\ &\leq 2 \|(A_1(\hat{x} - x) \cdots A_1(\hat{x} - x))\|_{\ell^{2,2}} \\ &\leq 2 \|(A(\hat{x} - x) \cdots A(\hat{x} - x))\|_{\ell^{2,2}} \leq 4\sqrt{d}\eta. \end{aligned}$$

In the last step we used the fact that \hat{x} and x are feasible for (2.6). Substituting this in (7.2) and (7.3) and recalling that $\|\cdot\|_{\text{TV}_i} = \|\nabla \cdot\|_{\ell^{2,1}}$ yields (3.5) and (3.16). \blacksquare

7.3. Image recovery for uniform random sampling. We now prove (3.6) and (3.7). These proofs rely on a discrete Poincaré inequality (Lemma 7.2). To prove this, as well as several later results, we will follow ideas from [24, 25] and relate the TV semi-norm to the decay rate of Haar wavelet coefficients. For notation and background on Haar wavelets, see §B.

Lemma 7.2 (Discrete Poincaré inequality). *Let $x \in \mathbb{C}^{N^d}$ with $\sum_{i=1}^{N^d} x_i = 0$. Then*

$$\|x\|_{\ell^2} \leq \sqrt{N} \|x\|_{\text{TV}}, \quad d = 1, \quad \|x\|_{\ell^2} \lesssim \frac{\|x\|_{\text{TV}_a}}{2^{d/2-1}} \leq \frac{\sqrt{d} \|x\|_{\text{TV}_i}}{2^{d/2-1}}, \quad d \geq 2.$$

Proof. See [26, Lem. 4.1] for the $d = 1$ result. Now consider $d \geq 2$. Let $x \in \mathbb{C}^{N^d}$ with mean zero and f be its isometric embedding, i.e. $f(i/N) = N^{d/2} X_i$ where $i = (i_1, \dots, i_d) \in \{0, \dots, N-1\}^d$ and $x = \text{vec}(X)$. Note that $\|f\|_{L^2} = \|x\|_{\ell^2}$ and f also has mean zero. Let $c_{j,n}^{(e)}$ denote the Haar wavelet coefficient of f . Since f has mean zero, we have $c_{0,0}^{(0)} = 0$. Write $c_{j,n} \in \mathbb{C}^{2^{d-1}}$ for the vector of coefficients $c_{j,n}^{(e)}$ with $e \in \{0, 1\}^d \setminus \{0\}$. Then Lemma B.1 and Lemma B.2 give that when $d \geq 2$, there exists a constant $C > 0$ such that

$$\|c_{(k)}\|_{\ell^2} \leq C \frac{2^{j(d-2)/2} |f|_{BV}}{k}, \quad |f|_{BV} \leq N^{-d/2+1} \|x\|_{\text{TV}_a}.$$

Since $2^j \leq N/2$ we have

$$(7.4) \quad \|c_{(k)}\|_{\ell^2} \leq C \frac{\|x\|_{\text{TV}_a}}{k \cdot 2^{d/2-1}} \leq C \frac{\sqrt{d} \|x\|_{\text{TV}_i}}{k \cdot 2^{d/2-1}},$$

where in the second inequality we use (2.5). Therefore

$$\|x\|_{\ell^2} = \|f\|_{L^2} = \sqrt{\sum_{k=1}^{\infty} \|c_{(k)}\|_{\ell^2}^2} \leq C \frac{\|x\|_{\text{TV}_a}}{2^{d/2-1}} \leq C \frac{\sqrt{d} \|x\|_{\text{TV}_i}}{2^{d/2-1}},$$

as required. \blacksquare

This now gives the following:

Lemma 7.3. *Let A be the measurement matrix of Theorem 3.1. Then*

$$\|z\|_{\ell^2} \leq \|Az\|_{\ell^2} + \sqrt{N}\|z\|_{\text{TV}}, \quad \forall z \in \mathbb{C}^N.$$

If A is the measurement matrix of Theorem 3.2 then

$$\|z\|_{\ell^2} \leq \|Az\|_{\ell^2} + 2^{1-d/2}\|z\|_{\text{TV}_a} \leq \|Az\|_{\ell^2} + 2^{1-d/2}\sqrt{d}\|z\|_{\text{TV}_i}, \quad \forall z \in \mathbb{C}^{N^d}.$$

Proof. Consider the case $d \geq 2$ first. Let $z \in \mathbb{C}^{N^d}$ and define $\bar{z} = (\bar{z}_i)_{i=1}^{N^d}$ with $\bar{z}_i = z_i - \frac{1}{N^d} \sum_{j=1}^{N^d} z_j$. Then we have $\sum_{i=1}^{N^d} \bar{z}_i = 0$ and applying the Poincaré inequality gives

$$\|\bar{z}\|_{\ell^2} \lesssim 2^{1-d/2}\|\bar{z}\|_{\text{TV}_a} = 2^{1-d/2}\|z\|_{\text{TV}_a}.$$

Since $\sum_{j=1}^{N^d} z_j = (Fz)_0 = \sqrt{m}A_2z$ and $m \leq N^d$ by assumption, we have

$$\begin{aligned} \|z\|_{\ell^2} &\leq \frac{1}{\sqrt{N^d}}\|(Fz)_0\|_{\ell^2} + 2^{1-d/2}\|z\|_{\text{TV}_a} = \sqrt{\frac{m}{N^d}}\|A_2z\|_{\ell^2} + 2^{1-d/2}\|z\|_{\text{TV}_a} \\ &\leq \|Az\|_{\ell^2} + 2^{1-d/2}\|z\|_{\text{TV}_a}. \end{aligned}$$

This gives the first inequality. The second follows from (2.5). When $d = 1$ we use the same arguments, replacing the Poincaré inequality by its one-dimensional version (Lemma 7.2). ■

Proof of (3.2), (3.6) and (3.7). We use Lemma 7.3 with $z = \hat{x} - x$. This gives

$$\|\hat{x} - x\|_{\ell^2} \leq \|A(\hat{x} - x)\|_{\ell^2} + 2^{1-d/2}\|\hat{x} - x\|_{\text{TV}_a} \lesssim 2^{-d/2}\sigma_s(\nabla x)_{\ell^1} + (1 + 2^{-d/2}\sqrt{sd})\eta,$$

when $d \geq 2$, which yields (3.6). Here, for the second inequality, we use (3.4) and the fact that \hat{x} and x are feasible, so that $\|A(\hat{x} - x)\|_{\ell^2} \leq 2\eta$. This isotropic case (3.7) is identical. To prove (3.2), we use Lemma 7.3 with $d = 1$ and (3.1). ■

7.4. Image recovery for variable density Fourier sampling. We now consider variable density samples. We first show a strengthened Poincaré inequality for Haar-incoherent measurements, and then derive conditions under which this holds for variable density samples.

Lemma 7.4 (Poincaré inequality for Haar-incoherent measurements). *Let $W \in \mathbb{R}^{N^d \times N^d}$ be the matrix of the d -dimensional discrete Haar wavelet transform and $B \in \mathbb{C}^{m \times N^d}$. Suppose that BW satisfies the RIP of order $5k$ with constant $\delta_{5k} < 1/3$. Then*

$$\|x\|_{\ell^2} \lesssim \|Bx\|_{\ell^2} + \frac{\sqrt{N}\|x\|_{\text{TV}}}{k}, \quad d = 1,$$

and

$$\|x\|_{\ell^2} \lesssim_d \|Bx\|_{\ell^2} + \frac{\|x\|_{\text{TV}_a}}{\sqrt{k}} \log\left(\frac{N}{k}\right), \quad d \geq 2.$$

Proof. We may assume without loss of generality that x has mean zero. Let $A = BW$, $c = W^*x$ and Δ be the index set of the largest k entries of c in absolute value. Then Lemma A.7 with the trivial choices $\gamma = 1$ and $\sigma = \|P_{\Delta}^{\perp}c\|_{\ell^1}$ gives

$$\|x\|_{\ell^2} = \|c\|_{\ell^2} \lesssim \frac{\|P_{\Delta}^{\perp}c\|_{\ell^1}}{\sqrt{k}} + \|Ac\|_{\ell^2} = \frac{\|P_{\Delta}^{\perp}c\|_{\ell^1}}{\sqrt{k}} + \|Bx\|_{\ell^2}$$

Now, as in the proof of Lemma 7.2, let $c_{(k)} \in \mathbb{C}^{2^d-1}$ denote k^{th} largest wavelet coefficient block in c . Then

$$\|c_{(k)}\|_{\ell^2} \lesssim_d \begin{cases} \sqrt{N}\|x\|_{\text{TV}_a}/k^{3/2} & d = 1 \\ \|x\|_{\text{TV}_i}/k & d \geq 2 \end{cases}.$$

Hence, when $d = 1$, we have $\|P_{\Delta}^{\perp}c\|_{\ell^1} \lesssim \sqrt{N}\|x\|_{\text{TV}} \sum_{i>k} i^{-3/2} \lesssim \sqrt{N/k}\|x\|_{\text{TV}}$, as required. When $d \geq 2$, since Δ contains the index set of the largest k entries of c , we can bound $\|P_{\Delta}^{\perp}c\|_{\ell^1}$ by $\|P_{\Delta'}^{\perp}c\|_{\ell^1}$, where Δ' contains the indices of the largest $\lfloor k/(2^d-1) \rfloor$ of c . Hence

$$\|P_{\Delta}^{\perp}c\|_{\ell^1} \lesssim_d \|x\|_{\text{TV}_a} \sum_{i=\lfloor k/(2^d-1) \rfloor+1}^N i^{-1} \lesssim_d \|x\|_{\text{TV}_a} \log(N/k),$$

as required. ■

Lemma 7.5 (The RIP for the Fourier–Haar matrix). *Let $0 < \delta, \varepsilon < 1$, $2 \leq s \leq N^d$, $\Omega \subseteq \{-\frac{N}{2} + 1, \dots, \frac{N}{2}\}^d$ be a d -dimensional variable sampling pattern corresponding to a probability distribution $p = (p_{\omega})$, with $\Gamma(p)$ as in (3.10) and $D \in \mathbb{C}^{N^d \times N^d}$ be the diagonal matrix with entries $D_{ii} = \frac{1}{\sqrt{p_{\varrho^{-1}(i)}}}$, where $\varrho = \varrho^{(d)}$ is the bijection (2.2). Suppose that*

$$m \gtrsim_d \Gamma(p) \cdot s \cdot \log(\Gamma(p)s) \cdot (\log(\Gamma(p)s) \cdot \log(N) + \log(\varepsilon^{-1})).$$

Then, with probability at least $1 - \varepsilon$, the matrix

$$(7.5) \quad \frac{1}{\sqrt{mN^d}} P_{\Omega} D F W,$$

has the RIP of order s with constant $\delta_s \leq 1/2$, where F and W are the discrete Fourier and Haar wavelet transforms respectively.

Note that the factor $1/2$ here is arbitrary. To prove this, we use the tools introduced in §A.2. To this end, we first require an upper bound on the Fourier transform of the discrete Haar wavelet $\phi_{j,n}^{(e)}$. For this, we use Lemma B.3.

Proof of Lemma 7.5. Since $N^{-d/2}FW = U$ is unitary, the matrix (7.5) is a randomly-subsampled unitary matrix in the sense of §A. Hence it has the RIP of order s provided (A.3) holds, where Θ is as in (A.4). In particular, it suffices to show that $\Theta \leq \sqrt{\Gamma(p)}$. Indeed, if this holds, the log factor in (A.3) simplifies, since $s \geq 2$ and $\Gamma(p) \geq 1$ (this follows from (3.10) and the fact that p is a probability distribution). Using Lemma B.3 and tensor-product nature of

the Fourier transform and Haar wavelets, we see that

$$(7.6) \quad \Theta \lesssim_d \max_{\substack{\omega=(\omega_1, \dots, \omega_d) \\ -N/2 < \omega_1, \dots, \omega_d \leq N/2}} \max_{j=0, \dots, r-1} \left\{ \frac{1}{\sqrt{p_\omega}} \prod_{i=1}^d \frac{2^{j/2}}{\max\{\bar{\omega}_i, 2^j\}} \right\}.$$

Consider the product term on the right-hand side. Let $\pi : \{1, \dots, d\} \rightarrow \{1, \dots, d\}$ be a nonincreasing rearrangement of the $\bar{\omega}_i$, and let $0 \leq l \leq d+1$ be such that $\bar{\omega}_{\pi(l)} \geq 2^j \geq \bar{\omega}_{\pi(l+1)}$. Note that if $l = 0$ this means $2^j \geq \bar{\omega}_{\pi(1)}$ and if $l = d+1$ this means $\bar{\omega}_{\pi(d)} \geq 2^j$. Then

$$\prod_{i=1}^d \frac{2^{j/2}}{\max\{\bar{\omega}_i, 2^j\}} = \prod_{i=1}^l \frac{2^{j/2}}{\bar{\omega}_{\pi(i)}} \prod_{i=l+1}^d \frac{2^{j/2}}{2^j} = \frac{2^{j(l-d/2)}}{\bar{\omega}_{\pi(1)} \cdots \bar{\omega}_{\pi(l)}}.$$

Suppose first that d is even. Then, since $\bar{\omega}_{\pi(i)} \geq 2^j$ for $i = 1, \dots, l$, we can use the smallest $l - d/2$ such terms to bound the denominator, giving

$$\prod_{i=1}^d \frac{2^{j/2}}{\max\{\bar{\omega}_i, 2^j\}} \leq \frac{1}{\bar{\omega}_{\pi(1)} \cdots \bar{\omega}_{\pi(d/2)}}$$

If d is odd, then by a similar argument we obtain

$$\prod_{i=1}^d \frac{2^{j/2}}{\max\{\bar{\omega}_i, 2^j\}} \leq \frac{1}{\bar{\omega}_{\pi(1)} \cdots \bar{\omega}_{\pi((d-1)/2)} \sqrt{\bar{\omega}_{\pi((d+1)/2)}}.$$

Hence, recalling (3.8)–(3.9), and returning to (7.6), we deduce that

$$\Theta \lesssim_d \max_{\substack{\omega=(\omega_1, \dots, \omega_d) \\ -N/2 < \omega_1, \dots, \omega_d \leq N/2}} \left\{ \frac{1}{q_\omega \sqrt{p_\omega}} \right\} \leq \sqrt{\Gamma(p)},$$

as required. ■

We now return to the final arguments. We first require the following:

Lemma 7.6. *Under the conditions of Theorem 3.3, the following holds with probability at least $1 - \varepsilon/2$:*

$$(7.7) \quad \|x\|_{\ell^2} \lesssim \sqrt{\Gamma(p)} \|Ax\|_{\ell^2} + \frac{\sqrt{N} \|x\|_{\text{TV}}}{s}, \quad \forall x \in \mathbb{C}^N.$$

Under the conditions of Theorem 3.4, the following holds with probability at least $1 - \varepsilon/2$:

$$(7.8) \quad \|x\|_{\ell^2} \lesssim \sqrt{\Gamma(p)} \|Ax\|_{\ell^2} + \frac{\|x\|_{\text{TV}_a}}{\sqrt{s}}, \quad \forall x \in \mathbb{C}^{N^d}.$$

Proof. Consider the first case. The condition (3.11) and Lemma 7.5 give that the matrix $BW = \frac{1}{\sqrt{mN^d}} P_{\Omega_2} DFW$ has the RIP of order $2s+1$ with constant $\delta_{2s+1} \leq 1/2$. Hence Lemma 7.4 gives that

$$\|x\|_{\ell^2} \lesssim \|Bx\|_{\ell^2} + \frac{\sqrt{N} \|x\|_{\text{TV}}}{s}, \quad \forall x \in \mathbb{C}^N.$$

For the second case, the condition (3.14) and Lemma 7.5 give that the matrix $BW = \frac{1}{\sqrt{mN^d}} P_{\Omega_2} DFW$ has the RIP of order k with constant $\delta_{2k+1} \leq 1/2$, where $k = \lceil sd^2(\log N)^2 \rceil$. Hence Lemma 7.4 gives that

$$\|x\|_{\ell^2} \lesssim \|Bx\|_{\ell^2} + \frac{\|x\|_{\text{TV}_a}}{\sqrt{s}}, \quad \forall x \in \mathbb{C}^{N^d}.$$

Thus, it remains to show that $\|Bx\|_{\ell^2} \leq \sqrt{\Gamma(p)} \|Ax\|_{\ell^2}$. Observe that $B = \frac{1}{\sqrt{mN^d}} P_{\Omega_2} DF = \frac{1}{\sqrt{N^d}} DA_2$. Therefore

$$\|Bx\|_{\ell^2} \leq \frac{1}{\sqrt{N^d}} \|D\|_{\ell^2} \|Ax\|_{\ell^2} \leq \frac{1}{\sqrt{N^d \min_{\omega} \{\sqrt{p_{\omega}}\}}} \|Ax\|_{\ell^2} \leq \sqrt{\Gamma(p)} \|Ax\|_{\ell^2}.$$

Here, in the penultimate step we use (3.10) and the definition of q_{ω} to write

$$\frac{1}{\sqrt{N^d \sqrt{p_{\omega}}}} \leq \sqrt{\Gamma(p)} \frac{q_{\omega}}{N^{d/2}} \leq \sqrt{\Gamma(p)}.$$

The result now follows. ■

Proof of (3.13), (3.17) and (3.18). We consider the case $d \geq 2$. The case $d = 1$ is identical. As shown in §7.2, the gradient error bounds (3.15) and (3.16) hold with probability at least $1 - \varepsilon/2$. Hence, the bounds (3.15), (3.16) and (7.8) hold simultaneously with probability at least $1 - \varepsilon$. We now apply (7.8) to $\hat{x} - x$ to get

$$\|\hat{x} - x\|_{\ell^2} \lesssim \sqrt{\Gamma(p)} \|A(\hat{x} - x)\|_{\ell^2} + \frac{\|\hat{x} - x\|_{\text{TV}_a}}{\sqrt{s}} \lesssim \sqrt{\Gamma(p)} \eta + \frac{\|\hat{x} - x\|_{\text{TV}_a}}{\sqrt{s}}.$$

Hence (3.17) follows from (3.15). For (3.18), we use (7.8) and the inequality $\|\hat{x} - x\|_{\text{TV}_a} \leq \sqrt{d} \|\hat{x} - x\|_{\text{TV}_i}$ to get

$$\|\hat{x} - x\|_{\ell^2} \lesssim \sqrt{\Gamma(p)} \|A(\hat{x} - x)\|_{\ell^2} + \frac{\|\hat{x} - x\|_{\text{TV}_a}}{\sqrt{s}} \lesssim \sqrt{\Gamma(p)} \eta + \sqrt{d} \frac{\|\hat{x} - x\|_{\text{TV}_i}}{\sqrt{s}}.$$

The result then follows from (3.16). ■

8. Proofs Part II: Theorem 5.1. Since we no longer have the commuting property, our proof strategy is based on ideas from [25], see also [20]. In particular, we first show the following result, which extends [25, Thm. 6] for $d = 2$ to $d \geq 2$ dimensions:

Theorem 8.1. *Let $d \geq 2$, $N = 2^r \geq s \geq 2$, $W \in \mathbb{R}^{N^d \times N^d}$ be the matrix of the d -dimensional discrete Haar wavelet transform and $A \in \mathbb{C}^{m \times N^d}$. Suppose that AW has the RIP of order $t \gtrsim_d s \cdot \log(N) \cdot \log^2(N/s)$ with constant $\delta \leq 1/2$. Then for every $x \in \mathbb{C}^{N^d}$ and $y = Ax + e$, where $\|e\|_{\ell^2} \leq \eta$ for some $\eta \geq 0$, any minimizer \hat{x} of (2.6) satisfies*

$$\|\hat{x} - x\|_{\ell^2} \lesssim_d \frac{\sigma_s(\nabla x)_{\ell^1}}{\sqrt{s \log(N)}} + \eta.$$

This result asserts that any measurement matrix which is incoherent with the Haar wavelet basis yields stable and robust recovery via TV minimization. Hence, much as in the Fourier case, to derive guarantees for Walsh sampling we need to examine its incoherence with the Haar basis. Note that Theorem 8.1 does not apply when $d = 1$ (which is the reason our results for Walsh sampling apply only when $d \geq 2$), since it relies crucially on the multi-dimensional Haar coefficient bound that follows from Lemmas B.1 and B.2.

Proof of Theorem 8.1. Since the proof is similar to that of [25, Thm. 6], we omit some details. Let $z = \hat{x} - x$ and $c = W^*z$ be its discrete Haar coefficients. We may assume z is mean zero. Let $\pi : \{1, \dots, N^d\} \rightarrow \{1, \dots, N^d\}$ be a nonincreasing rearrangement of the entries of c in absolute value. We first show that

$$(8.1) \quad \sum_{j=k+1}^{N^d} |c_{\pi(j)}| \leq C_d \log(N^d/t) \left(\sum_{j=1}^k |c_{\pi(j)}| + \sigma_s(\nabla x)_{\ell^1} \right),$$

where $C_d > 0$ and $k = (2^d - 1)l + 1$ is minimal such that $k \geq \tau_d s \log(N)$ for some constant $\tau_d \geq 1$ to be defined later. Observe that

$$\sum_{j>k} |c_{\pi(j)}| \leq \sum_{i>t} \|c_{(i)}\|_{\ell^2},$$

where $c_{(i)} \in \mathbb{C}^{2^d-1}$ are the wavelet coefficient blocks, sorted in nonincreasing order. Hence Lemmas B.1 and B.2 give

$$(8.2) \quad \sum_{j>k} |c_{\pi(j)}| \lesssim \frac{\|\nabla z\|_{\ell^1}}{2^{d/2}} \log(N^d/t).$$

Let Δ be the index set of the largest s entries of ∇z in absolute value. It is straightforward to show that

$$(8.3) \quad \|P_{\Delta}^{\perp} \nabla z\|_{\ell^1} \leq 2\sigma_s(\nabla x)_{\ell^1} + \|P_{\Delta} \nabla z\|_{\ell^1}.$$

Now consider $\|\nabla z\|_{\ell^1}$. Write $\xi_1, \dots, \xi_{N^d} \in \mathbb{C}^{N^d}$ for the discrete Haar basis and let $\Lambda = \{j : (\nabla \xi_j)_i \neq 0 \text{ for some } i \in \Delta\}$ be the index set of those Haar wavelets that are nonconstant on Δ . It is straightforward to show that $|\Lambda| \lesssim_d s \log(N)$, thus we now let τ_d be such that $|\Lambda| \leq \tau_d s \log(N)$. Write $z = \sum_{j \in \Lambda} c_j \xi_j + \sum_{j \notin \Lambda} c_j \xi_j$. Then $P_{\Delta} \nabla z = \sum_{j \in \Lambda} c_j P_{\Delta} \nabla \xi_j$ by construction, and therefore

$$\|P_{\Delta} \nabla z\|_{\ell^1} \leq \sum_{j \in \Lambda} |c_j| \|\nabla \xi_j\|_{\ell^1} \lesssim_d \sum_{j \in \Lambda} |c_j|.$$

Here, in the second step we use the fact that $\|\nabla \xi_j\|_{\ell^1} \lesssim_d 1$, which follows easily from the definition of the ξ_j . Combining this with (8.3), we have

$$\|\nabla z\|_{\ell^1} \lesssim_d \sigma_s(\nabla x)_{\ell^1} + \sum_{j \in \Lambda} |c_j| \leq \sigma_s(\nabla x)_{\ell^1} + \sum_{j=1}^k |c_{\pi(j)}|,$$

where in the second step we use the definition of π and the fact that $|\Lambda| \leq \tau_d s \log(N) \leq k$. Substituting this into (8.2) now yields (8.1).

To complete the proof we apply Lemma A.7 to the matrix AW , with values $\gamma = \lceil C_d \log(N^d/t) \rceil$, $\sigma = \gamma \sigma_s(\nabla x)_{\ell^1}$ and $\Delta = \{\pi(1), \dots, \pi(k)\}$. The matrix AW satisfies the RIP of order $5k\gamma^2 \lesssim_d s \cdot \log(N) \cdot \log^2(N/s)$. Hence

$$\|\hat{x} - x\|_{\ell^2} = \|c\|_{\ell^2} \lesssim \frac{\sigma}{\gamma\sqrt{k}} + \|AWc\|_{\ell^2} \lesssim_d \frac{\sigma_s(\nabla x)_{\ell^1}}{\sqrt{s \log(N)}} + \|A(\hat{x} - x)\|_{\ell^2}.$$

The result now follows after noting that $\|A(\hat{x} - x)\|_{\ell^2} \leq 2\eta$. ■

Lemma 8.2. *Let $0 < \delta, \varepsilon < 1$, $2 \leq s \leq N^d$, $\Omega \subseteq \{0, \dots, N-1\}^d$ be a d -dimensional variable sampling pattern corresponding to a probability distribution $p = (p_i)$, with $\Gamma(p)$ as in (5.1) and $D \in \mathbb{C}^{N^d \times N^d}$ be the diagonal matrix with entries $D_{ii} = \frac{1}{\sqrt{p_{e^{-1}(i)}}}$. Suppose that*

$$m \gtrsim_d \Gamma(p) \cdot s \cdot \log(\Gamma(p)s) \cdot (\log(\Gamma(p)s) \cdot \log(N) + \log(\varepsilon^{-1})).$$

Then, with probability at least $1 - \varepsilon$, the matrix

$$(8.4) \quad \frac{1}{\sqrt{mN^d}} P_\Omega D H W,$$

has the RIP of order s with constant $\delta_s \leq 1/2$, where H and W are the discrete Walsh–Hadamard and Haar wavelet transforms respectively.

Proof. As in the Fourier case (see the proof of Lemma 7.5), the matrix $N^{-d/2}HW$ is unitary and therefore $A = \frac{1}{\sqrt{mN^d}} P_\Omega D H W$ is a randomly-sampled unitary matrix. Hence it has the RIP of order s whenever (A.3) holds with Θ as in (A.4) for $U = N^{-d/2}HW$. Hence it suffices to show that $\Theta^2 \lesssim_d \Gamma(p)$.

Let v_i denote the one-dimensional Walsh function on $[0, 1)$ and $\psi_{j,n}^{(e)}$ be the one-dimensional Haar wavelet. Then

$$(8.5) \quad \left| \langle v_i, \psi_{j,n}^{(0)} \rangle_{L^2} \right| = \begin{cases} 2^{-j/2} & i < 2^j \\ 0 & \text{otherwise} \end{cases}, \quad \left| \langle v_i, \psi_{j,n}^{(1)} \rangle_{L^2} \right| = \begin{cases} 2^{-j/2} & 2^j \leq i < 2^{j+1} \\ 0 & \text{otherwise} \end{cases},$$

See [2, Thm. 6.8]. In particular, this implies that

$$(8.6) \quad |\langle v_i, \psi_{j,n}^{(e)} \rangle_{L^2}| \leq \begin{cases} 2^{-j/2} & i < 2^{j+1} \\ 0 & \text{otherwise} \end{cases}.$$

Let $\psi_{j,n}^{(e)}$ be the d -dimensional Haar wavelets on $[0, 1]^d$ and $v_i = v_{i_1} \otimes \dots \otimes v_{i_d}$ be the d -dimensional Walsh functions, where $i = (i_1, \dots, i_d)$. Then

$$\Theta = \max \left\{ \frac{1}{\sqrt{p_i}} \left| \langle v_i, \psi_{j,n}^{(e)} \rangle_{L^2} \right| \right\},$$

where the maximum is taken over all $i = (i_1, \dots, i_d) \in \{0, \dots, N-1\}^d$, $n = (n_1, \dots, n_d)$ with $0 \leq n_k < 2^j$, $j = 0, \dots, r-1$ and $e \in \{0, 1\}^d$ with $e \neq 0$ unless $j = 0$. Using (8.6), we have

$$\left| \langle v_i, \psi_{j,n}^{(e)} \rangle_{L^2} \right| = \prod_{k=1}^d \left| \langle v_{i_k}, \psi_{j,n_k}^{(e_k)} \rangle_{L^2} \right| \leq \begin{cases} 2^{-jd/2} & \|i\|_{\ell^\infty} < 2^{j+1} \\ 0 & \text{otherwise} \end{cases}.$$

It follows that $|\langle v_i, \psi_{j,n}^{(e)} \rangle_{L^2}| \lesssim_d (1 + \|i\|_{\ell^\infty}^{d/2})^{-1}$ and therefore

$$\Theta \lesssim_d \max_i \left\{ \frac{1}{\sqrt{p_i}} \left(1 + \|i\|_{\ell^\infty}^{d/2} \right)^{-1} \right\} \leq \sqrt{\Gamma(p)},$$

as required. ■

Proof of Theorem 5.1. Let A' be the matrix defined in (8.4) of Lemma 8.2. This lemma, the condition on m and the fact that $\Gamma(p) \gtrsim \log(N)$ (Lemma 5.2) imply that A' has the RIP of order $t \gtrsim_d s \log(N) \log^2(N/s)$. To complete the proof, we cannot simply invoke Theorem 8.1, since the measurement matrix $A = \frac{1}{\sqrt{m}} P_\Omega H$ is not scaled in such a way for AW to have the RIP. Instead, we follow the same steps as its proof, making necessary adjustments. Let $z = \hat{x} - x$, $c = W^* z$ be its Haar coefficients and k be as in the proof. Then (8.1) holds (this property does not depend on the measurement matrix). We now apply [25, Prop. 3] using the matrix A' and the values $\gamma = \lceil C_d \log(N^d/t) \rceil$ and $\sigma = \gamma \sigma_s(\nabla x)_{\ell^1}$. This gives

$$\|\hat{x} - x\|_{\ell^2} = \|d\|_{\ell^2} \lesssim \frac{\sigma_s(\nabla x)_{\ell^1}}{\sqrt{s \log(N)}} + \|A'c\|_{\ell^2}.$$

Now observe that

$$\|A'c\|_{\ell^2} = \|A'W^*(\hat{x} - x)\|_{\ell^2} = \frac{1}{\sqrt{N^d}} \|D\|_{\ell^2} \|A(\hat{x} - x)\|_{\ell^2} \leq \frac{2}{\sqrt{N^d} \min_i \{\sqrt{p_i}\}} \eta.$$

Observe that

$$\frac{1}{\sqrt{N^d} \sqrt{p_i}} \leq \sqrt{\Gamma(p)} \frac{\sqrt{1 + \|i\|_{\ell^\infty}^d}}{\sqrt{N^d}} \leq \sqrt{2\Gamma(p)}.$$

Hence $\|A'd\|_{\ell^2} \lesssim \sqrt{\Gamma(p)} \eta$, as required. ■

Appendix A. Preliminary results from compressed sensing.

Below we collect some standard compressed sensing results. For further information, see for instance [16].

A.1. Sparsity, rNSP and RIP. Let $N \geq s \geq 2$. Recall that a vector $x \in \mathbb{C}^N$ is s -sparse if it has at most s nonzero entries. We write Σ_s for the set of s -sparse vectors. Let D_s denote the set of all subsets $\Delta \subseteq \{1, \dots, N\}$ for which $|\Delta| \leq s$. Thus, $x \in \Sigma_s$ if and only if its support $\text{supp}(x) = \{i : x_i \neq 0\}$ belongs to D_s .

Definition A.1 (Robust Null Space Property). A matrix $A \in \mathbb{C}^{m \times N}$ satisfies the robust Null Space Property (rNSP) of order s with constants $0 < \rho < 1$ and $\gamma > 0$ if

$$(A.1) \quad \|P_{\Delta}x\|_{\ell^2} \leq \frac{\rho}{\sqrt{s}} \|P_{\Delta}^{\perp}x\|_{\ell^1} + \gamma \|Ax\|_{\ell^2}, \quad \forall x \in \mathbb{C}^N, \Delta \in D_s.$$

Lemma A.2 (rNSP implies ℓ^1 and ℓ^2 distance bounds). Suppose that A has the rNSP of order s with constants $0 < \rho < 1$ and $\gamma > 0$. Let $x, z \in \mathbb{C}^N$. Then

$$\|z - x\|_{\ell^1} \leq \frac{1 + \rho}{1 - \rho} (2\sigma_s(x)_{\ell^1} + \|z\|_{\ell^1} - \|x\|_{\ell^1}) + \frac{2\gamma}{1 - \rho} \sqrt{s} \|A(z - x)\|_{\ell^2},$$

and

$$\|x - z\|_{\ell^2} \leq \frac{(3\rho + 1)(\rho + 1)}{2(1 - \rho)} \left(\frac{2\sigma_s(x)_{\ell^1} + \|z\|_{\ell^1} - \|x\|_{\ell^1}}{\sqrt{s}} \right) + \frac{(3\rho + 5)\gamma}{2(1 - \rho)} \|A(z - x)\|_{\ell^2}.$$

Note that this result is a special case (corresponding to $M = 1$) of a result proved later, Lemma A.6.

Definition A.3. The s^{th} Restricted Isometry Constant (RIC) δ_s of a matrix $A \in \mathbb{C}^{m \times N}$ is the smallest $\delta \geq 0$ such that

$$(A.2) \quad (1 - \delta)\|x\|_{\ell^2}^2 \leq \|Ax\|_{\ell^2}^2 \leq (1 + \delta)\|x\|_{\ell^2}^2, \quad \forall x \in \Sigma_s.$$

If $0 < \delta_s < 1$ then the matrix A is said to have the Restricted Isometry Property (RIP) of order s .

Note that the RIP implies the rNSP. For instance, if A has the RIP of order $2s$ with constant $\delta_{2s} < 4/\sqrt{41}$ then it has the rNSP of order s with constants ρ and γ depending on δ_{2s} [16, Thm. 6.13].

For convenience, we now state one further result:

Lemma A.4. If $A \in \mathbb{C}^{m \times N}$ satisfies the rNSP of order s with constants ρ and γ , then

$$B = \begin{pmatrix} A & & \\ & \ddots & \\ & & A \end{pmatrix} \in \mathbb{C}^{dm \times dN},$$

has the rNSP of order s with constants $\rho' = \rho$ and $\gamma' = \sqrt{d}\gamma$.

Proof. Consider any $x = (x_1^{\top}, \dots, x_d^{\top}) \in \mathbb{C}^{dN}$ with $x_i \in \mathbb{C}^N$. Let $\Lambda \subseteq \{1, \dots, dN\}$ with $|\Lambda| = s$, and write $\Lambda = \Lambda_1 \cup \dots \cup \Lambda_d$ where $\Lambda_i \subseteq \{(i-1)N + 1, \dots, iN\}$. Since $|\Lambda_i| \leq s$ the rNSP for A gives

$$\begin{aligned} \|P_{\Lambda}x\|_{\ell^2} &\leq \|P_{\Lambda_1}x_1\|_{\ell^2} + \dots + \|P_{\Lambda_d}x_d\|_{\ell^2} \\ &\leq \frac{\rho}{\sqrt{s}} (\|P_{\Lambda_1}^{\perp}x_1\|_{\ell^1} + \dots + \|P_{\Lambda_d}^{\perp}x_d\|_{\ell^1}) + \gamma (\|Ax_1\|_{\ell^2} + \dots + \|Ax_d\|_{\ell^2}) \\ &\leq \frac{\rho}{\sqrt{s}} \|P_{\Lambda}^{\perp}x\|_{\ell^1} + \sqrt{d}\gamma \|Bx\|_{\ell^2}, \end{aligned}$$

as required. ■

A.2. Bounded orthonormal systems. Let $\mathcal{D} \subset \mathbb{R}^d$ be a domain with a probability measure ν and ψ_1, \dots, ψ_N be an orthonormal system of complex-valued functions on \mathcal{D} . The system is a *bounded orthonormal system* with constant Θ if

$$\sup_{t \in \mathcal{D}} |\psi_j(t)| \leq \Theta, \quad j = 1, \dots, N.$$

Given such a system, draw t_1, \dots, t_m random and independently from ν and define the measurement matrix

$$A = \frac{1}{\sqrt{m}} (\psi_j(t_i))_{i,j=1}^{m,N} \in \mathbb{C}^{m \times N}.$$

Let $0 < \delta, \epsilon < 1$ and $N \geq s \geq 2$. The following result was shown in [14, Thm. 2.2] (we have slightly simplified the log factor below using the fact that $N \geq s \geq 2$). Suppose that

$$(A.3) \quad m \gtrsim \delta^{-2} \cdot \Theta^2 \cdot s \cdot L, \quad L = \log \left(\frac{\Theta^2 s}{\delta^2} \right) \cdot \left[\frac{1}{\delta^4} \log \left(\frac{\Theta^2 s}{\delta^2} \right) \cdot \log(N) + \frac{1}{\delta} \log \left(\frac{1}{\delta \epsilon} \log \left(\frac{\Theta^2 s}{\delta^2} \right) \right) \right].$$

Then, with probability at least $1 - \epsilon$ the matrix A has the RIP of order s with $\delta_s \leq \delta$.

Randomly-subsampled unitary matrices are important examples of the bounded orthonormal system framework. Let $U \in \mathbb{C}^{N \times N}$ be unitary and $p = (p_i)_{i=1}^N$ be a probability distribution on $\{1, \dots, N\}$. Draw t_1, \dots, t_m independently and randomly from p and consider the measurement matrix

$$A = \frac{1}{\sqrt{m}} P_T D U \in \mathbb{C}^{m \times N}, \quad D = \text{diag}(1/\sqrt{p_1}, \dots, 1/\sqrt{p_N}) \in \mathbb{C}^{N \times N},$$

where $T = \{t_1, \dots, t_m\}$ and P_T is the row selector matrix. Now let $\mathcal{D} = \{1, \dots, N\}$, ν be the probability measure corresponding to p and define $\phi_j(i) = \frac{1}{\sqrt{p_i}} u_{ij}$, where $U = (u_{ij})$. It is straightforward to verify that this is a bounded orthonormal system. The constant Θ is

$$(A.4) \quad \Theta = \max_{i,j=1,\dots,N} \frac{|u_{ij}|}{\sqrt{p_i}}.$$

A.3. Matrix recovery. We now need a more general version of the rNSP, see for instance, [15, Defn. 4.1]:

Definition A.5. A matrix $A \in \mathbb{C}^{m \times N}$ satisfies the $\ell^{2,2}$ -robust Null Space Property (rNSP) of order s with constants $0 < \rho < 1$ and $\gamma > 0$ if

$$\|P_\Delta X\|_{\ell^{2,2}} \leq \frac{\rho}{\sqrt{s}} \|P_\Delta^\perp X\|_{\ell^{2,1}} + \gamma \|AX\|_{\ell^{2,2}}, \quad \forall X \in \mathbb{C}^{N \times M}, \Delta \in D_s.$$

As shown in [15, Prop. 4.3], if $A \in \mathbb{C}^{m \times N}$ satisfies the RIP of order s with constant $\delta_{2s} < \frac{4}{\sqrt{41}}$, then A satisfies the $\ell^{2,2}$ -rNSP of order s with constants ρ and γ depending on δ_{2s} . We also have the following generalization of Lemma A.2:

Lemma A.6 (rNSP implies $\ell^{2,1}$ and $\ell^{2,2}$ distance bounds). *Suppose that A has the $\ell^{2,2}$ -rNSP of order s with constants $0 < \rho < 1$ and $\gamma > 0$. Let $X, Z \in \mathbb{C}^{N \times M}$. Then*

$$(A.5) \quad \|Z - X\|_{\ell^{2,1}} \leq \frac{1+\rho}{1-\rho} (2\sigma_s(X)_{\ell^{2,1}} + \|Z\|_{\ell^{2,1}} - \|X\|_{\ell^{2,1}}) + \frac{2\gamma}{1-\rho} \sqrt{s} \|A(Z - X)\|_{\ell^{2,2}},$$

and

$$(A.6) \quad \|Z - X\|_{\ell^{2,2}} \leq \frac{(3\rho+1)(\rho+1)}{2(1-\rho)} \left(\frac{2\sigma_s(X)_{\ell^{2,1}} + \|Z\|_{\ell^{2,1}} - \|X\|_{\ell^{2,1}}}{\sqrt{s}} \right) + \frac{(3\rho+5)\gamma}{2(1-\rho)} \|A(Z - X)\|_{\ell^{2,2}}.$$

Proof. Consider (A.5). Let $V = Z - X$ and $\Delta \in D_s$ be such that $\|X_\Delta^\perp\|_{\ell^{2,1}} = \sigma_s(X)_{\ell^{2,1}}$. Then we have

$$\begin{aligned} \|X\|_{\ell^{2,1}} + \|P_\Delta^\perp V\|_{\ell^{2,1}} &= \|X\|_{\ell^{2,1}} + \|P_\Delta^\perp(Z - X)\|_{\ell^{2,1}} \\ &\leq \|P_\Delta X\|_{\ell^{2,1}} + 2\|P_\Delta^\perp X\|_{\ell^{2,1}} + \|P_\Delta^\perp Z\|_{\ell^{2,1}} \\ &= 2\|P_\Delta^\perp X\|_{\ell^{2,1}} + \|P_\Delta X\|_{\ell^{2,1}} + \|Z\|_{\ell^{2,1}} - \|P_\Delta Z\|_{\ell^{2,1}} \\ &\leq 2\sigma_s(X)_{\ell^{2,1}} + \|P_\Delta V\|_{\ell^{2,1}} + \|Z\|_{\ell^{2,1}}, \end{aligned}$$

which implies that

$$\|P_\Delta^\perp V\|_{\ell^{2,1}} \leq 2\sigma_s(X)_{\ell^{2,1}} + \|Z\|_{\ell^{2,1}} - \|X\|_{\ell^{2,1}} + \|P_\Delta V\|_{\ell^{2,1}}.$$

Now consider $\|P_\Delta V\|_{\ell^{2,1}}$. We have

$$\|P_\Delta V\|_{\ell^{2,1}} \leq \sqrt{s} \|P_\Delta V\|_{\ell^{2,2}} \leq \rho \|P_\Delta^\perp V\|_{\ell^{2,1}} + \sqrt{s} \gamma \|AV\|_{\ell^{2,2}}.$$

Hence

$$\|P_\Delta V\|_{\ell^{2,1}} \leq \rho(2\sigma_s(X)_{\ell^{2,1}} + \|Z\|_{\ell^{2,1}} - \|X\|_{\ell^{2,1}} + \|P_\Delta V\|_{\ell^{2,1}}) + \sqrt{s} \gamma \|AV\|_{\ell^{2,2}},$$

which gives

$$\|P_\Delta V\|_{\ell^{2,1}} \leq \frac{\rho}{1-\rho} (2\sigma_s(X)_{\ell^{2,1}} + \|Z\|_{\ell^{2,1}} - \|X\|_{\ell^{2,1}}) + \sqrt{s} \frac{\gamma}{1-\rho} \|AV\|_{\ell^{2,2}}.$$

Now we have

$$\begin{aligned} \|Z - X\|_{\ell^{2,1}} &\leq \|P_\Delta V\|_{\ell^{2,1}} + \|P_\Delta^\perp V\|_{\ell^{2,1}} \\ &\leq 2\sigma_s(X)_{\ell^{2,1}} + \|Z\|_{\ell^{2,1}} - \|X\|_{\ell^{2,1}} + 2\|P_\Delta V\|_{\ell^{2,1}} \\ &\leq \frac{1+\rho}{1-\rho} (2\sigma_s(X)_{\ell^{2,1}} + \|Z\|_{\ell^{2,1}} - \|X\|_{\ell^{2,1}}) + \frac{2\gamma}{1-\rho} \sqrt{s} \|A(Z - X)\|_{\ell^{2,2}}. \end{aligned}$$

This gives (A.5).

For (A.6), notice that it suffice show that

$$(A.7) \quad \|Z - X\|_{\ell^{2,2}} \leq \frac{3\rho + 1}{2} \frac{\|Z - X\|_{\ell^{2,1}}}{\sqrt{s}} + \frac{3\gamma}{2} \|A(Z - X)\|_{\ell^{2,2}}.$$

Once this is shown, then (A.6) follows immediately from (A.5). To show (A.7), let $V = Z - X$ and write $v_i \in \mathbb{C}^M$ for its i^{th} row. Let $\Delta \subseteq \{1, \dots, N\}$ be the index set of the largest s entries of $(\|v_i\|_{\ell^2})_{i=1}^N$. Then

$$\|P_{\Delta}V\|_{\ell^{2,2}} = \sqrt{\sum_{i \in \Delta} \|v_i\|_{\ell^2}^2} \geq \sqrt{s} \min_{i \in \Delta} \|v_i\|_{\ell^2} \geq \sqrt{s} \max_{i \notin \Delta} \|v_i\|_{\ell^2},$$

which implies that

$$\begin{aligned} \|P_{\Delta}^{\perp}V\|_{\ell^{2,2}}^2 &= \sum_{i \notin \Delta} \|v_i\|_{\ell^2}^2 \leq \sum_{i \notin \Delta} \|v_i\|_{\ell^2} \max_{i \notin \Delta} \|v_i\|_{\ell^2} \\ &\leq \sum_{i \notin \Delta} \|v_i\|_{\ell^2} \frac{\|P_{\Delta}V\|_{\ell^{2,2}}}{\sqrt{s}} = \frac{\|P_{\Delta}V\|_{\ell^{2,2}}}{\sqrt{s}} \|P_{\Delta}^{\perp}V\|_{\ell^{2,1}}. \end{aligned}$$

Now, applying Young's inequality, we deduce that

$$\|P_{\Delta}^{\perp}V\|_{\ell^{2,2}} \leq \frac{\|P_{\Delta}V\|_{\ell^{2,2}}}{2} + \frac{\|P_{\Delta}^{\perp}V\|_{\ell^{2,1}}}{2\sqrt{s}}.$$

Hence

$$\begin{aligned} \|V\|_{\ell^{2,2}} &\leq \|P_{\Delta}V\|_{\ell^{2,2}} + \|P_{\Delta}^{\perp}V\|_{\ell^{2,2}} \leq \frac{3}{2} \|P_{\Delta}V\|_{\ell^{2,2}} + \frac{\|P_{\Delta}^{\perp}V\|_{\ell^{2,1}}}{2\sqrt{s}} \\ &\leq \frac{3\rho + 1}{2\sqrt{s}} \|P_{\Delta}^{\perp}V\|_{\ell^{2,1}} + \frac{3\gamma}{2} \|AV\|_{\ell^{2,2}}. \end{aligned}$$

Since $\|P_{\Delta}^{\perp}V\|_{\ell^{2,1}} \leq \|V\|_{\ell^{2,1}}$ we obtain the desired result. ■

A.4. Miscellaneous results. The following is essentially [25, Prop. 3], although with a couple of minor modifications. Since the proof is identical, it is omitted.

Lemma A.7. *Let $\gamma \in \mathbb{N}$ and suppose that $A \in \mathbb{C}^{m \times N}$ has the RIP of order $5k\gamma^2$ with constant $\delta \leq 1/2$. Let $c \in \mathbb{C}^N$ and suppose that there is a set $\Delta \subseteq \{1, \dots, N\}$ with $|\Delta| \leq k$ such that*

$$\|P_{\Delta}^{\perp}c\|_{\ell^1} \leq \gamma \|P_{\Delta}c\|_{\ell^1} + \sigma,$$

for some $\sigma \geq 0$. Then

$$\|c\|_{\ell^2} \lesssim \frac{\sigma}{\gamma\sqrt{k}} + \|Ac\|_{\ell^2}.$$

Appendix B. Haar wavelets.

B.1. Definitions. The Haar scaling function and mother wavelet are defined by

$$\psi^{(0)}(t) = \begin{cases} 1 & 0 \leq t < 1 \\ 0 & \text{otherwise} \end{cases}, \quad \psi^{(1)}(t) = \begin{cases} 1 & 0 \leq t < 1/2 \\ -1 & 1/2 \leq t < 1 \\ 0 & \text{otherwise} \end{cases}$$

For $e \in \{0, 1\}$, $j, n \in \mathbb{Z}$, define $\psi_{j,n}^{(e)}(t) = 2^{j/2} \psi^{(e)}(2^j t - n)$. Then the set

$$\{\psi_{0,0}^{(0)}\} \cup \{\psi_{j,n}^{(1)} : n = 0, \dots, 2^j - 1, j = 0, 1, \dots\},$$

is an orthonormal basis of $L^2([0, 1])$.

Next, consider $d \geq 2$ and for $e = (e_1, \dots, e_d) \in \{0, 1\}^d$, $j \in \mathbb{Z}$ and $n = (n_1, \dots, n_d) \in \mathbb{Z}^d$ define the function

$$\psi_{j,n}^{(e)} = \psi_{j,n_1}^{(e_1)} \otimes \dots \otimes \psi_{j,n_d}^{(e_d)},$$

where \otimes denotes the tensor product. Then

$$\{\psi_{0,0}^{(0)}\} \cup \{\psi_{j,n}^{(e)} : e \in \{0, 1\}^d \setminus \{0\}, n = (n_1, \dots, n_d), n_1, \dots, n_d = 0, \dots, 2^j - 1, j = 0, 1, \dots\},$$

is an orthonormal basis of $L^2([0, 1]^d)$.

Given $f \in L^2([0, 1]^d)$, we may write

$$f = c_{0,0}^{(0)} \psi_{0,0}^{(0)} + \sum_{e \in \{0,1\}^d \setminus \{0\}} \sum_{j=0}^{\infty} \sum_{\substack{n=(n_1, \dots, n_d) \\ 0 \leq n_1, \dots, n_d < 2^j}} c_{j,n}^{(e)} \psi_{j,n}^{(e)},$$

where $c_{j,n}^{(e)} = \langle f, \psi_{j,n}^{(e)} \rangle$. For convenience, we define $c_{j,n} \in \mathbb{C}^{2^d - 1}$ for the vector containing the values $c_{j,n}^{(e)}$, $e \in \{0, 1\}^d \setminus \{0\}$.

Let $d \geq 1$, $N = 2^r$ and consider \mathbb{C}^{N^d} . Let

$$\phi_{j,n}^{(e)} = \text{vec}(\Phi_{j,n}^{(e)}) \in \mathbb{R}^{N^d}$$

where $\Phi_{j,n}^{(e)} \in \mathbb{R}^{N \times \dots \times N}$ with

$$(\Phi_{j,n}^{(e)})_i = N^{d/2} \psi_{j,n}^{(e)}(i_1/N, \dots, i_d/N), \quad i = (i_1, \dots, i_d) \in \{0, \dots, N-1\}^d,$$

is the (normalized) discretization of $\psi_{j,n}^{(e)}$ on an equispaced grid of N^d points on $[0, 1]^d$. Then the set

$$\{\phi_{0,0}^{(0)}\} \cup \{\phi_{j,n}^{(e)} : e \in \{0, 1\}^d \setminus \{0\}, n = (n_1, \dots, n_d), n_1, \dots, n_d = 0, \dots, 2^j - 1, j = 0, \dots, r-1\},$$

is an orthonormal basis for \mathbb{C}^{N^d} , the discrete Haar basis. After selecting an ordering for this basis, write $W \in \mathbb{R}^{N^d \times N^d}$ for the orthogonal matrix whose columns consist of these vectors, i.e. the discrete Haar wavelet transform.

B.2. Relation to the TV semi-norm. In the following two lemmas, $BV([0, 1]^d)$ is the space of functions of bounded variation on $[0, 1]^d$, and $|\cdot|_{BV}$ is the usual BV semi-norm, see, for example, [24]. The following can be found in [24, Lem. 7]:

Lemma B.1. *Let $x = \text{vec}(X) \in \mathbb{C}^{N^d}$, where $X \in \mathbb{C}^{N \times \dots \times N}$ and $f \in BV([0, 1]^d)$ be its isometric embedding as a piecewise constant function, i.e.*

$$f(i/N) = N^{d/2} X_i,$$

where $i = (i_1, \dots, i_d) \in \{0, \dots, N-1\}^d$. If $|f|_{BV}$ is the BV semi-norm of f , then

$$|f|_{BV} \leq N^{-d/2+1} \|x\|_{\text{TV}_a}.$$

The following result illustrates the relation between Haar coefficients and the BV semi-norm (see, for instance, [24, Prop. 8]):

Lemma B.2. *Let $d \geq 2$. There exists a constant $C > 0$ such that the following holds for all mean-zero $f \in BV([0, 1]^d)$. Let $c_{j,n}^{(e)}$ be the Haar wavelet coefficients of f and $c_{j,n} \in \mathbb{C}^{2^d-1}$ be the vector of values $c_{j,n}^{(e)}$, $e \in \{0, 1\}^d \setminus \{0\}$. Let $c_{(1)}, c_{(2)}, \dots$ be a reordering of these vectors so that $\|c_{(1)}\|_{\ell^2} \geq \|c_{(2)}\|_{\ell^2} \geq \dots$. Then*

$$|c_{(k)}| \lesssim \frac{|f|_{BV}}{k^{3/2}}, \quad d = 1,$$

and

$$\|c_{(k)}\|_{\ell^2} \lesssim \frac{2^{j_k(d-2)/2} |f|_{BV}}{k}, \quad d \geq 2,$$

where j_k is the scale corresponding to $c_{(k)}$.

B.3. The Fourier transform of a Haar wavelet. Finally, we also give the following:

Lemma B.3. *Let F be the one-dimensional DFT matrix, $\{\psi_{j,n}^{(e)}\}$ be the one-dimensional discrete Haar wavelet basis and ϱ be defined as in (2.1). Then for $j = 0, \dots, r-1$, $n = 0, \dots, 2^j-1$, $e \in \{0, 1\}$ and any $\omega \in \{-N/2+1, \dots, N/2\}$ we have*

$$(B.1) \quad \frac{1}{\sqrt{N}} |(F\psi_{j,n}^{(e)})_{\varrho^{-1}(\omega)}| \lesssim \frac{1}{\sqrt{\bar{\omega}}} \min \left\{ \left(\frac{2^j}{\bar{\omega}} \right)^{1/2}, \left(\frac{\bar{\omega}}{2^j} \right)^{1/2+e} \right\}.$$

In particular,

$$(B.2) \quad \frac{1}{\sqrt{N}} |(F\psi_{j,n}^{(e)})_{\varrho^{-1}(\omega)}| \lesssim \min \left\{ \frac{2^{j/2}}{\bar{\omega}}, \frac{\bar{\omega}^e}{2^{j(e+1/2)}} \right\} \lesssim \min \left\{ \frac{2^{j/2}}{\bar{\omega}}, \frac{1}{2^{j/2}} \right\} = \frac{2^{j/2}}{\max\{\bar{\omega}, 2^j\}}.$$

We recall here the definition $\bar{\omega} = \max\{1, |\omega|\}$, and that the rows of F are indexed over $\{1, \dots, N\}$; hence the use of the bijection ϱ . The calculations that lead to this lemma can be found in, for instance, [5, 20]. For completeness we give the proof:

Proof of Lemma B.3. We proceed by direct calculation. We have

$$\begin{aligned}
(F\psi_{j,n}^{(e)})_{\varrho^{-1}(\omega)} &= 2^{\frac{j-r}{2}} \sum_{n2^{r-j} < t \leq (n+1/2)2^{r-j}} e^{-2\pi i \omega(t-1)/N} \\
&\quad + (-1)^e 2^{\frac{j-r}{2}} \sum_{(n+1/2)2^{r-j} < t \leq (n+1)2^{r-j}} e^{-2\pi i \omega(t-1)/N} \\
&= 2^{\frac{j-r}{2}} e^{-2\pi i \omega n 2^{r-j}/N} \sum_{s=0}^{2^{r-j-1}-1} e^{-2\pi i \omega s/N} \\
&\quad + (-1)^e 2^{\frac{j-r}{2}} e^{-2\pi i \omega (n+1/2)2^{r-j}/N} \sum_{s=0}^{2^{r-j-1}-1} e^{-2\pi i \omega s/N}.
\end{aligned}$$

Hence

$$(F\psi_{j,n}^{(e)})_{\varrho^{-1}(0)} = \begin{cases} 2^{\frac{r-j}{2}} & e = 0 \\ 0 & \text{otherwise} \end{cases},$$

and for $\omega \in \{-N/2 + 1, \dots, N/2\} \setminus \{0\}$,

$$(B.3) \quad (F\psi_{j,n}^{(e)})_{\varrho^{-1}(\omega)} = 2^{j/2-r/2} e^{-2\pi i \omega n / 2^j} \left(1 + (-1)^e e^{-2\pi i \omega / 2^{j+1}} \right) \left(\frac{1 - e^{-2\pi i \omega / 2^{j+1}}}{1 - e^{-2\pi i \omega / 2^r}} \right).$$

Observe that (B.1) trivially holds when $\omega = 0$. Hence we now consider $\omega \neq 0$. By (B.3),

$$\frac{1}{\sqrt{N}} \left| (F\psi_{j,n}^{(e)})_{\varrho^{-1}(\omega)} \right| \leq 2^{j/2-r} \frac{|\sin(\pi \omega / 2^{j+1})|^{1+e}}{|\sin(\pi \omega / 2^r)|}.$$

Suppose first that $1 \leq |\omega| < 2^j$. Then, since $|\sin(\pi z)| \leq \pi |z|$, $\forall z \in \mathbb{R}$, and $|\sin(\pi z)| \geq 2|z|$ for $|z| \leq 1/2$, we have

$$\frac{1}{\sqrt{N}} \left| (F\psi_{j,n}^{(e)})_{\varrho^{-1}(\omega)} \right| \lesssim 2^{j/2-r} \frac{(|\omega|/2^j)^{1+e}}{|\omega|/2^r} = 2^{-j/2} (|\omega|/2^j)^e.$$

Conversely, if $2^j \leq |\omega| \leq 2^{r-1}$ then we use the bound $|\sin(\pi z)| \leq 1$, $\forall z \in \mathbb{R}$, to obtain

$$\frac{1}{\sqrt{N}} \left| (F\psi_{j,n}^{(e)})_{\varrho^{-1}(\omega)} \right| \lesssim 2^{j/2-r} \frac{1}{|\omega|/2^r} = 2^{-j/2} \frac{2^j}{|\omega|}.$$

This gives the result. ■

Appendix C. Proof of selected results from §4 and §5.

Proof of Lemma 4.1. Notice that $q_\omega = \overline{\omega_1}$ whenever $\omega = (\omega_1, 0, \dots, 0)$. Hence

$$\sum_{\omega} (q_\omega)^{-2} \geq \sum_{t=1}^{N/2} \frac{1}{t} \gtrsim \log(N).$$

Since $p = (p_\omega)$ is a probability distribution, i.e. $\sum_\omega p_\omega = 1$, we deduce that $\Gamma(p) \gtrsim \log(N)$.

Now consider the upper bound. Suppose first that d is even. Then there are $d!$ different nonincreasing rearrangements π . Hence

$$\sum_\omega (q_\omega)^{-2} \lesssim_d \sum_{t_1=1}^{N/2} \sum_{t_2=1}^{t_1} \cdots \sum_{t_d=1}^{t_{d-1}} \frac{1}{(t_1 \cdots t_{d/2})^2} \leq \sum_{t_1=1}^{N/2} \sum_{t_2=1}^{t_1} \cdots \sum_{t_{d/2}=1}^{t_{d/2-1}} \frac{(t_{d/2})^{d/2}}{(t_1 \cdots t_{d/2})^2} = F_{d/2}(N/2),$$

where

$$F_m(N) = \sum_{t_1=1}^N \sum_{t_2=1}^{t_1} \cdots \sum_{t_m=1}^{t_{m-1}} \frac{(t_m)^m}{(t_1 \cdots t_m)^2}.$$

Similarly, if d is odd we have

$$\begin{aligned} \sum_\omega (q_\omega)^{-2} &\lesssim_d \sum_{t_1=1}^{N/2} \sum_{t_2=1}^{t_1} \cdots \sum_{t_d=1}^{t_{d-1}} \frac{1}{(t_1 \cdots t_{(d-1)/2})^2 t_{(d+1)/2}} \\ &\leq \sum_{t_1=1}^{N/2} \sum_{t_2=1}^{t_1} \cdots \sum_{t_{(d+1)/2}=1}^{t_{(d-1)/2}} \frac{(t_{(d+1)/2})^{(d-1)/2}}{(t_1 \cdots t_{(d-1)/2})^2 t_{(d+1)/2}} \\ &= F_{(d+1)/2}(N/2). \end{aligned}$$

We now show that $F_m(N) \lesssim_m \log(N)$ for any $m \in \mathbb{N}$. When $m = 1$ the result is trivial. Now consider $m \geq 2$. We have

$$F_m(N) \lesssim_m \sum_{t_1=1}^N \sum_{t_2=1}^{t_1} \cdots \sum_{t_{m-1}=1}^{t_{m-2}} \frac{(t_{m-1})^{m-1}}{(t_1 \cdots t_{m-1})^2} = F_{m-1}(N).$$

Hence the result follows by induction. Therefore, for either even or odd d , we have shown that

$$\sum_\omega (q_\omega)^{-2} \lesssim_d \log(N).$$

Since $p = (p_\omega)$ is a probability distribution the result now follows. ■

Proof of Lemma 4.5. Since all norms are equivalent on \mathbb{R}^d , we may without loss of generality consider the ℓ^∞ -norm. We first estimate the constant $C_{N,d,\alpha}$. Hence

$$(C.1) \quad (C_{N,d,\alpha})^{-1} \asymp_d \sum_{t_1=1}^{N/2} \sum_{t_2=1}^{t_1} \cdots \sum_{t_d=1}^{t_{d-1}} \frac{1}{(t_d)^\alpha} \asymp_{d,\alpha} \sum_{t_1=1}^{N/2} (t_d)^{d-1-\alpha} \asymp_{d,\alpha} \begin{cases} N^{d-\alpha} & \alpha < d \\ \log(N) & \alpha = d \\ 1 & \alpha > d \end{cases}.$$

Next, observe that $\Gamma(p)$ is defined by

$$\Gamma(p) C_{N,d,\alpha} = \max_\omega \frac{(1 + \|\omega\|_{\ell^\infty})^\alpha}{(q_\omega)^2} \asymp_\alpha \max_\omega \frac{(\overline{\omega_{\pi(1)}})^\alpha}{(q_\omega)^2}.$$

We now split into two cases. Suppose first that $\alpha \geq 2$. Then, using the definition of q_ω , we see that the maximum is attained at $\omega = (N/2, 0, \dots, 0)$, giving

$$\Gamma(p)C_{N,d,\alpha} \asymp_\alpha N^{\alpha-2}.$$

Conversely, when $\alpha < 2$ the maximum is attained when $\omega = (0, \dots, 0)$, giving

$$\Gamma(p)C_{N,d,\alpha} \asymp_\alpha 1.$$

We now combine these two estimates with (C.1) to get the result. ■

Proof of Corollary 4.7. Observe that

$$(C_{N,d})^{-1} = \sum_{\omega} \frac{1}{\prod_{j=1}^d \bar{\omega}_j} \asymp_d \left(\sum_{t=1}^{N/2} \frac{1}{t} \right)^d \asymp_d \log^d(N).$$

Moreover, using the fact that $\bar{\omega}_{\pi(j)} \geq \sqrt{\bar{\omega}_{\pi(j)} \bar{\omega}_{\pi(d/2+j)}}$ when d is even, and similarly for d odd, we deduce that

$$(q_\omega)^2 \geq \bar{\omega}_1 \cdots \bar{\omega}_d = \frac{C_{N,d}}{p_\omega},$$

and therefore $\Gamma(p) \leq (C_{N,d})^{-1}$. Hence $\Gamma(p) \asymp_d \log^d(N)$, which gives the result. ■

Proof of Lemma 5.2. This follows immediately from (C.1). ■

REFERENCES

- [1] B. ADCOCK, V. ANTUN, R. BERGMAN, AND A. C. HANSEN, *Effective sampling strategies for compressive imaging*, In preparation, (2019).
- [2] B. ADCOCK, V. ANTUN, AND A. C. HANSEN, *Uniform recovery in infinite-dimensional compressed sensing and applications to structured binary sampling*, arXiv:1905.00126, (2019).
- [3] B. ADCOCK, S. BRUGIAPAGLIA, AND M. KING-ROSKAMP, *Do log factors matter? On optimal wavelet approximation and the foundations of compressed sensing*, arXiv:1905.10028, (2019).
- [4] B. ADCOCK, A. C. HANSEN, C. POON, AND B. ROMAN, *Breaking the coherence barrier: A new theory for compressed sensing*, Forum Math. Sigma, 5 (2017).
- [5] B. ADCOCK, A. C. HANSEN, AND B. ROMAN, *A note on compressed sensing of structured sparse wavelet coefficients from subsampled Fourier measurements*, IEEE Signal Process. Letters, 23 (2016), pp. 732–736.
- [6] V. ANTUN, *Coherence estimates between Hadamard matrices and Daubechies wavelets*, master's thesis, University of Oslo, 2016.
- [7] A. BASTOUNIS AND A. C. HANSEN, *On the absence of uniform recovery in many real-world applications of compressed sensing and the restricted isometry property and nullspace property in levels*, SIAM J. Imaging Sci., 10 (2017), pp. 335–371.
- [8] S. BECKER, J. BOBIN, AND E. J. CANDÈS, *NESTA: A Fast and Accurate First-Order Method for Sparse Recovery*, SIAM Journal on Imaging Sciences, 4 (2011), pp. 1–39.
- [9] J.-F. CAI AND W. XU, *Guarantees of total variation minimization for signal recovery*, Inf. Inference, 4 (2015), pp. 328–353.
- [10] E. J. CANDÈS, J. ROMBERG, AND T. TAO, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509.

- [11] A. CHAMBOLLE, V. DUVAL, G. PEYRÉ, AND C. POON, *Geometric properties of solutions to the total variation denoising problem*, Inverse Problems, 33 (2016), p. 015002.
- [12] A. CHAMBOLLE, M. NOVAGA, D. CREMERS, AND T. POCK, *An introduction to total variation for image analysis*, in Theoretical Foundations and Numerical Methods for Sparse Recovery, M. Fornasier, ed., vol. 9 of Radon Series in Computational and Applied Mathematics, de Gruyter, Berlin, 2010, pp. 263–340.
- [13] A. CHAMBOLLE AND T. POCK, *An introduction to continuous optimization for imaging*, Acta Numer., 25 (2016), pp. 161–319.
- [14] A. CHKIFA, N. DEXTER, H. TRAN, AND C. G. WEBSTER, *Polynomial approximation via compressed sensing of high-dimensional functions on lower sets*, Math. Comp., 87 (2018), pp. 1415–1450.
- [15] N. DEXTER, H. TRAN, AND C. WEBSTER, *A mixed ℓ_1 regularization approach for sparse simultaneous approximation of parameterized PDEs*, ESAIM Math. Model. Numer. Anal., 53 (2019), pp. 2025–2045.
- [16] S. FOUCART AND H. RAUHUT, *A Mathematical Introduction to Compressive Sensing*, Birkhauser, 2013.
- [17] E. GAUSS, *Walsh Funktionen für Ingenieure und Naturwissenschaftler*, Springer Fachmedien Wiesbaden, 1994.
- [18] B. GOLUBOV, A. EFIMOV, AND V. SKVORTSOV, *Walsh Series and Transforms: Theory and Applications*, Springer Netherlands, 1991.
- [19] F. KRAHMER, C. KRUSCHEL, AND M. SANDBICHLER, *Total variation minimization in compressed sensing*, in Compressed Sensing and Its Applications, Birkhäuser, 2018.
- [20] F. KRAHMER AND R. WARD, *Stable and robust sampling strategies for compressive imaging*, IEEE Trans. Image Process., 23 (2013), pp. 612–622.
- [21] C. LI AND B. ADCOCK, *Compressed sensing with local structure: uniform recovery guarantees for the sparsity in levels class*, Appl. Comput. Harmon. Anal., 46 (2019), pp. 453–477.
- [22] A. MOSHTAGHPOUR, *Computational Interferometry for Hyperspectral Imaging*, PhD thesis, Université catholique de Louvain, 2019.
- [23] A. MOSHTAGHPOUR, J. B. DIAS, AND L. JACQUES, *Close encounters of the binary kind: signal reconstruction guarantees for compressive Hadamard sampling with Haar wavelet basis*, IEEE Trans. Inf. Theory (in press), (2020).
- [24] D. NEEDELL AND R. WARD, *Near-optimal compressed sensing guarantees for total variation minimization*, IEEE Trans. Image Process., 22 (2013), pp. 3941–3949.
- [25] D. NEEDELL AND R. WARD, *Stable image reconstruction using total variation minimization*, SIAM J. Imaging Sci., 6 (2013), pp. 1035–1058.
- [26] C. POON, *On the role of total variation in compressed sensing*, SIAM J. Imaging Sci., 8 (2015), pp. 682–720.
- [27] B. ROMAN, A. C. HANSEN, AND B. ADCOCK, *On asymptotic structure in compressed sensing*, arXiv:1406.4178, (2014).
- [28] V. TEMLYAKOV, *Multivariate Approximation*, Cambridge University Press, 2018.