Do log factors matter? On optimal wavelet approximation and the foundations of compressed sensing

Ben Adcock, Simone Brugiapaglia and Matthew King–Roskamp
Department of Mathematics
Simon Fraser University
Canada
May 27, 2019

Abstract
A signature result in compressed sensing is that Gaussian random sampling achieves stable and robust recovery of sparse vectors under optimal conditions on the number of measurements. However, in the context of image reconstruction, it has been extensively documented that sampling strategies based on Fourier measurements outperform this purportedly optimal approach. Motivated by this seeming paradox, we investigate the problem of optimal sampling for compressed sensing. Rigorously combining the theories of wavelet approximation and infinite-dimensional compressed sensing, our analysis leads to new error bounds in terms of the total number of measurements \( m \) for the approximation of piecewise \( \alpha \)-Hölder functions. In this setting, we show the suboptimality of random Gaussian sampling, exhibit a provably optimal sampling strategy and prove that Fourier sampling outperforms random Gaussian sampling when the Hölder exponent \( \alpha \) is large enough. This resolves the claimed paradox, and provides a clear theoretical justification for the practical success of compressed sensing techniques in imaging problems.

1 Introduction
Compressed sensing asserts that a vector \( x \in \mathbb{C}^N \) with at most \( s \) nonzero components can be recovered from \( m \) suitably-chosen linear measurements \( y = Ax \), where \( A \in \mathbb{C}^{m \times N} \) and \( y \in \mathbb{C}^m \), with \( m \) satisfying

\[
m \geq c \cdot s \cdot \log(N/s).
\]  

(1.1)

This can be achieved, for instance, by using a random Gaussian matrix for \( A \) and by solving the basis pursuit problem

\[
\min_{z \in \mathbb{C}^N} \|z\|_{\ell^1} \text{ subject to } Az = y.
\]  

(1.2)

In practice, (1.1) can represent a significant saving in the number of measurements over classical approaches, and for this reason compressed sensing has found use in many different applications in science and engineering. In fact, compressed sensing is optimal for the recovery of sparse vectors. No stable method (that is, one which is robust to perturbations in \( x \)) can recover sparse vectors from asymptotically fewer than \( s \log(N/s) \) measurements.

Imaging lies at the foundation of compressed sensing, and has been one of its key beneficiaries. Natural images have approximately sparse wavelet coefficients, and compressed sensing
allows for the reconstruction of an image up to its best $s$-term approximation error from a few as $m \approx c \cdot s \cdot \log(N/s)$ measurements. Magnetic Resonance Imaging (MRI), for instance, was one of the original motivations for compressed sensing – indeed, it was considered in the seminal paper of Candès, Romberg & Tao [24] – and has been one of its most fruitful areas of application [49, 50]. Such methods were approved for commercial use in MRI by the US FDA in 2017 [35]. In the realm of optical imaging, the single-pixel camera [33] was one of the first empirical demonstrations of compressed sensing principles, and its various progenitors such as lensless imaging [16, 43] continue to be active areas of investigation [39]. Other modalities, including X-ray CT [40], infrared imaging [52], spectral imaging [11], light-field imaging [53], ghost imaging [44], STORM [69], holography [18], fluorescence microscopy [61], NMR [42, 45], radio interferometry [67], to name but a few, have all benefitted from compressed sensing approaches.

1.1 A paradox

Many imaging modalities such as MRI acquire Fourier samples of an image, and not measurements according to a random Gaussian matrix. The best known measurement condition for $s$-term recovery from Fourier measurements is $m \approx c \cdot s \cdot \log^3(s) \cdot \log^2(N)$ (see §1.7), which is in short of the optimal condition (1.1) for Gaussian measurements.

However, in practice, Fourier measurements outperform Gaussian measurements for recovering images. A typical example is shown in Fig. 1. With the same total number of measurements, reconstructing from a suitably-chosen set of Fourier samples gives a significantly better reconstruction. As is standard in imaging, in this figure a db4 wavelet basis is used as a sparsifying transform.

Motivated by this paradox, the focus of this paper is the wavelet approximation of piecewise smooth functions via compressed sensing. Specifically, we answer the following three questions:

(Q1) Is a random Gaussian sampling an optimal sampling strategy for wavelet approximation of piecewise smooth functions? Specifically, does it achieve optimal approximation rates in terms of $m$?

(Q2) If not, what is an optimal sampling strategy?

(Q3) How close to optimal is Fourier sampling? In particular, why is it that Fourier sampling often outperforms random Gaussian sampling, even though the latter is optimal for recovering sparse vectors?

The observation made in Fig. 1 that Fourier sampling outperforms random Gaussian sampling for imaging has been well-documented [9, 13, 59], but not yet rigorously explained. However, it arguably lies at the heart as to why compressed sensing has proved so effective for these applications. In particular, it explains why modalities such as MRI have benefitted significantly from compressed sensing principles, despite their measurements being seemingly suboptimal.¹

¹This also provides some explanation as to why attempts to modify devices such as MR scanners to produce Gaussian-like measurements (see, for example, [41, 56, 57]) have not been widely adopted.
1.2 Our contributions

Our answers to these questions use the language of nonlinear approximation theory of piecewise smooth functions. Consider a function $f : [0, 1] \to \mathbb{R}$ that is piecewise $\alpha$-Hölder continuous ($\alpha \geq 1$). We first show that random Gaussian sampling combined with a decoder based on orthonormal wavelets and basis pursuit (1.2) gives an approximation $\tilde{f}_m$ satisfying

$$
\|f - \tilde{f}_m\|_{L^2} \leq C(\log(m))^{\alpha} / m^\alpha.
$$

(1.3)

This result is optimal ‘up to a logarithmic factor’, and for small $\alpha$ one may not be inclined to worry. Yet this is hardly satisfactory for moderate to large $\alpha$. On the other hand, in answer to (Q2) we show that there exists a sampling strategy which, when combined with the same decoder (orthonormal wavelets and basis pursuit), achieves the optimal error bound

$$
\|f - \tilde{f}_m\|_{L^2} \leq C/m^\alpha.
$$

(1.4)

Unfortunately, this strategy does not employ Fourier measurements, rendering it inapplicable in many practical problems. Focusing on this case, we prove that it is possible to construct a Fourier sampling strategy that achieves an error bound

$$
\|f - \tilde{f}_m\|_{L^2} \leq C(\log(m))^{13/4+\delta} / m^\alpha,
$$

(1.5)
for any $0 < \delta < 1$. This result pertains to question (Q3). Its main implication is that for sufficiently regular functions (specifically, $\alpha > 13/4$) Fourier sampling is guaranteed to outperform random Gaussian sampling (1.3).

Let us highlight several key features of our analysis:

1. We, for arguably the first time, rigorously connect compressed sensing theory to the classical theory of nonlinear approximation using wavelets.

2. We work in the infinite-dimensional setting. Compressed sensing is customarily presented in the finite-dimensional setting of vectors and matrices, whereas the concern of wavelet approximation theory is, of course, functions in function spaces. Here we work directly with functions $f : [0, 1] \to \mathbb{R}$ using the framework of *infinite-dimensional* compressed sensing [5]. This setup avoids errors due to discretization (e.g. related the wavelet crime). Our analysis also takes care to incorporate all sources of approximation error, for instance, those due to truncation.\(^2\)

3. Our approach is essentially a *black box*. Given a smoothness parameter $\alpha$ and a budget of measurements $m$, our recipe determines the correct samples to acquire, and from them finds an approximation $\hat{f}_m$ satisfying the above bounds. No other inputs are required. In particular, unfeasible conditions such as bounds on the expansion error of $f$ (which are quite common in the compressed sensing literature) are not required.

4. In order to construct $\hat{f}_m$, our recipe uses Daubechies’ wavelets and solves a standard, finite-dimensional (weighted) $\ell^1$ minimization problem such as basis pursuit (1.2). This is very similar to standard implementations of compressed sensing in most practical applications. In particular, we do not make use of any exotic ‘structure-promoting’ decoders, these typically being difficult to implement in large-scale problems (see §1.5).

5. Our results (1.3) and (1.5) are uniform in $f$. That is, given $m$, we construct a sampling strategy and decoder that guarantees these error bounds (with high probability) for all piecewise $\alpha$-Hölder continuous functions. While (1.3) is nonuniform (that is, specific to the function $f$), we prove a uniform version with an additional $\sqrt{\log(m)}$ factor.

6. Our result for Fourier sampling relies on recent advances in compressed sensing theory based on local, as opposed to global, structure. Specifically, we use the principles of *local sparsity in levels* and *multilevel random subsampling* [8] to finely tune the sampling strategy to give the approximation result (1.5). To the best of our knowledge, this result cannot be achieved using standard, sparsity-based, compressed sensing theory.

### 1.3 Structure is key

How are the results (1.4) and (1.5) possible? The answer lies with the structure of wavelet coefficients. Random Gaussian measurements exploit the approximate sparsity of wavelet coefficients. In particular, a piecewise $\alpha$-Hölder continuous function has a best $s$-term wavelet approximation error decaying like $s^{-\alpha}$ (see Theorem 2.5). The bound (1.3) follows almost directly from this and the measurement condition (1.1).

\(^2\)In particular, by ‘Fourier measurements’ we mean samples of the continuous Fourier transform of $f$, not its discrete Fourier transform. Not only is this more convenient for the analysis, it is also more relevant in practice, since modalities such as MRI are based on the continuous Fourier transform [10].
By contrast, the Fourier and optimal sampling strategies exploit both the sparsity and the distribution of the wavelet coefficients. In particular, they exploit the following properties:

- Asymptotically, all but $O(\log(s))$ of the significant wavelet coefficients are located at coarse scales. That is, the sparsity at coarse scales satisfies $s_{\text{coarse}} = s - O(\log(s))$.
- The coarse scales are saturated. Coefficients at these scales are non-sparse, i.e. all coefficients contribute to the best $s$-term approximation.
- At fine scales, wavelet coefficients are sparse, with the number of significant coefficients being roughly $s_{\text{fine}} = O(\log(s))$.

Since the coarse scales are saturated, they can be recovered efficiently using ‘classical’ non-compressive measurements. Specifically, only $m_{\text{coarse}} = s_{\text{coarse}}$ suitably-chosen measurements are required to recover the corresponding coefficients. Conversely, the fine scales are recovered using compressive measurements, requiring $m_{\text{fine}} \approx c \cdot s_{\text{fine}} \cdot \log(N/s_{\text{fine}})$ measurements, where $N \approx s^{2\alpha+1}$ is the range of indices in which the largest $s$ wavelet coefficients live. Hence, as $s \to \infty$, we have

$$m = m_{\text{coarse}} + m_{\text{fine}} = s + o(s).$$

This, in combination with the $s^{-\alpha}$ approximation rate, leads to the optimal bound (1.3).

This works for the optimal sampling strategy since we have the luxury to choose the measurements. The situation is significantly more complicated for Fourier sampling. In this case, we exploit the following fundamental property of wavelets:

(P) Wavelets are concentrated in frequency. The Fourier transform of a wavelet at scale $k$ is essentially supported in a dyadic band $B_k$ in frequency.

Let $s_k$ be the sparsity of the wavelet coefficients at scale $k$. At coarse scales, $s_k = 2^k$, where $2^k$ is the size of the scale, and at fine scales $s_k = O(\log(s))$. The Fourier sampling strategy therefore proceeds as follows. At coarse scales, it fully samples the band $B_k$, using $m_k = |B_k| = 2^k = s_k$ measurements. At fine scales, it randomly subsamples from the band $B_k$, using $m_k = c \cdot s_k \cdot \text{polylog}(s_k, N)$ measurements. Hence, as $s \to \infty$, we have

$$m = m_1 + \ldots + m_r = s + o(s),$$

where $r \approx \log_2(N)$ is the maximum scale and where $N \approx m_{\text{max}}^{\max(2\alpha+1, \alpha-1/2)}$. As with the optimal strategy, the key here is that Fourier measurements can efficiently recover the coarse scale wavelet coefficients.

1.4 Is Fourier sampling optimal?

Having read the above argument, the reader may wonder why (1.5) involves a logarithmic factor at all? The challenge stems from the word ‘essentially’ appearing in property (P). While wavelets at scale $k$ are concentrated in the band $B_k$, they are not fully supported there. Hence, the pleasant scenario of simply choosing $m_k$ measurements per band according to the corresponding $s_k$ is not realized in practice. Dealing with these inter-scale interferences is a major technical hurdle, and culminates in the log term in (1.5). We discuss this term in more detail in §4.2 and, in particular, the prospect of reducing the exponent 13/4, in §11.

Nevertheless, our main result (1.5) states that Fourier sampling will, asymptotically, outperform Gaussian sampling for sufficiently large $\alpha$. In practice, this always appears to be
Figure 2: Comparison of different sampling and recovery strategies for the approximation of a piecewise \( \alpha \)-Hölder function with 10 discontinuities, defined as in (E.1), using Haar (left) and db4 wavelets (right). The strategies (Gauss, \( \ell^1 \)), (Optimal, \( \ell^1 \)) and (Fourier, \( \ell^1 \)) correspond to (1.3), (1.4) and (1.5), respectively. Recovery is performed via (1.2) in all cases. (Gauss, Tree) corresponds to random Gaussian sampling combined with tree-structured CoSaMP recovery, as proposed in [12], where \( c \) is a tuning parameter of the method. Further details are provided in Appendix E.

...Our focus in this paper is on proving theoretical statements such as (1.5). As part of this, we prescribe a specific Fourier sampling strategy (by this, we mean a specific set of frequencies to sample) suitable for the class of piecewise \( \alpha \)-Hölder functions. Unsurprisingly, in practice, further improvements can be achieved by empirically tuning the sampling strategy to better capture the structure of the images being reconstructed. We refer to [59] for a demonstration of the significant practical benefits of doing this in various different modalities.

1.5 On structure-promoting decoders

As discussed in §1.3, the methodology employed in this paper designs a sampling strategy based on the structured sparsity of the wavelet coefficients. There is a line of work in compressed sensing that seeks to exploit structured sparsity by changing the decoder, e.g. by replacing the \( \ell^1 \)-norm in (1.2) by a different convex penalty, or by using iterative or greedy algorithms [30, 34]. For example, [12] considers a modification of the CoSaMP algorithm that promotes the connected tree structure of wavelet coefficients. Theoretical analysis shows that for random Gaussian sampling this decoder achieves recovery using asymptotically fewer than the \( m \approx Cs \log(N/s) \) measurements required by (1.2). However, this analysis only applies to Gaussian measurements, and certainly not Fourier sampling, as is common in practice in imaging. Moreover, even with Gaussian measurements, this approach is typically outperformed by Fourier sampling with vanilla \( \ell^1 \)-minimization. Fig. 2 gives a typical example (see Appendix E for further examples). Computationally, the latter is also significantly faster.

From this, it was concluded [9] that the types of sparsity structures inherent to imaging are more effectively exploited in the sampling strategy than in the decoder. The results (1.4) and
support this conclusion. Further examples demonstrating this phenomenon in imaging scenarios are shown in [9, 59]. Whether one can do both, i.e. combine structured sampling strategies with decoders that promote structured sparsity, and witness even further benefits, is an interesting question for future work.

1.6 Extensions

The objects considered in this paper are piecewise $\alpha$- Hölder functions of one variable. Although wavelet approximation theory is more commonly studied in Besov spaces [32], we have chosen this class for the ease of presentation. Possible extensions, including to more than one dimension and other -let families such as shearlets, are discussed in §11.

1.7 Relation to previous work

The idea of sampling the coarse and fine wavelet scales of an image differently was arguably first considered by Tsaig & Donoho [63]. Romberg, in his seminal paper on compressed sensing for imaging, used DCT measurements to capture the coarse scales and noiselet measurements for the fine scales [60]. Later, Candès & Romberg [23] considered Fourier sampling in the dyadic bands $B_k$ according to structured wavelet sparsity. However, their analysis assumes direct sensing of individual wavelet scales, which is infeasible in practice. This idea was also used in [66] to design empirical Fourier sampling strategies for MRI. Similar ideas have also been pursued in [17].

In the context of Fourier sampling, it has long been known that fully sampling the low frequencies is crucial for high-quality compressed sensing reconstruction [49]. Partial theoretical justifications have been given in [26] and [58]. Furthermore, in [7] (one dimension) and [6] (two dimensions) it has been shown that low frequency Fourier samples optimally recover the coarse scale wavelet coefficients.

In terms of theoretical results, for two-dimensional discrete Fourier sampling with Haar wavelets, it was shown in [46] that

$$m \approx c \cdot s \cdot \log^3(s) \cdot \log^2(N),$$

measurements (chosen randomly according to an inverse square law) suffice for recovery of an approximately $s$-sparse vector of Haar wavelet coefficients. Unfortunately, even if this were extended to continuous Fourier measurements with higher-order wavelets, it would lead to a highly suboptimal error bound $\|f - \hat{f}_m\|_{L^2} \leq C(\log(m))^{5\alpha}/m^\alpha$.

Instead, our approach relies on the generalization of compressed sensing [8] based on local structure. This framework provides precise estimates relating local quantities (sparsities and numbers of measurements), which allows us to adjust the sampling strategy to the local sparsity structure. We also make use of ideas and theoretical results from [14], [48] and [62].

1.8 Outline

We commence in §2 with some requisite material, including nonlinear approximation using wavelets and some standard compressed sensing theory. Our main results are stated in §3 and discussed in detail in §4. §6 and §7 give the proofs corresponding to (1.3) and (1.4) respectively. The remainder of the paper is devoted to the proof of (1.5). In §8 we reformulate wavelet approximation from Fourier samples as a finite compressed sensing problem. Next in
§9 we recap the framework of [8]. Finally, §10 gives the proof of (1.5). We conclude in §11 by listing some open problems.

2 Preliminaries

2.1 Notation

We work primarily in the Hilbert space $L^2([0,1])$ of square-integrable functions on $[0,1]$. Write $\langle \cdot, \cdot \rangle_{L^2([0,1])}$ and $\| \cdot \|_{L^2([0,1])}$ (or, in short, $\langle \cdot, \cdot \rangle_{L^2}$ and $\| \cdot \|_{L^2}$) for the corresponding inner product or norm. We use $c, C$ to denote arbitrary numerical constants. We use $c^\alpha, C^\alpha$, and so forth for constants that depend on a parameter $\alpha$. By convention, capital letters are used in error estimates and lower case letters are used for assumptions on, for instance, the number of measurements $m$. Throughout, the value of such constants may change from line to line. We make no attempt to track constants.

If $\Omega \subseteq \{1, \ldots, N\}$, we use $P_\Omega$ to denote either the $N \times N$ matrix of the projection onto the space of vectors supported on $\Omega$, or the $m \times N$ matrix ($m = |\Omega|$) which selects the entries of a vector in $\Omega$. Its meaning will be clear from the context. If $\Omega = \{1, \ldots, M\}$ we write $P_M$, and if $\Omega = \{N_1 + 1, \ldots, N_2\}$ we write $P_{N_1}^{N_2}$.

2.2 Wavelets and nonlinear approximation

We consider Daubechies’ wavelets with $p \geq 1$ vanishing moments. Write $\varphi$ and $\psi$ for the corresponding scaling function and mother wavelet, respectively. Given such a wavelet, we define the smoothness parameter $q \geq 0$ as the largest number such that

$$
|\hat{\varphi}(\omega)| \lesssim (1 + |\omega|)^{-1-q}, \quad |\hat{\psi}(\omega)| \lesssim (1 + |\omega|)^{-1-q}, \quad \forall \omega \in \mathbb{R}.
$$

(2.1)

Here $\hat{\cdot}$ denotes the Fourier transform – see Appendix A. The exact values for $q$ can be found, for example, in [29, p. 226]. Note that asymptotically one has $q \to \infty$ as $p \to \infty$ [29, p. 226].

Since our interest lies with approximation on $[0,1]$, we use $P_\Omega$ to denote either the $N \times N$ matrix of the projection onto the space of vectors supported on $\Omega$, or the $m \times N$ matrix ($m = |\Omega|$) which selects the entries of a vector in $\Omega$. Its meaning will be clear from the context. If $\Omega = \{1, \ldots, M\}$ we write $P_M$, and if $\Omega = \{N_1 + 1, \ldots, N_2\}$ we write $P_{N_1}^{N_2}$.

Remark 2.1 There are several other strategies for constructing an orthonormal wavelet basis of $L^2([0,1])$ [51, Sec. 7.5]. Periodization is the simplest, and hence we use it throughout. Periodization treats the endpoints as additional discontinuities, which is often undesirable in practice. However, this is no limitation in our setting, since the concern of this paper lies with the asymptotic rate of nonlinear approximation of piecewise smooth functions, which is unaffected by the addition of a finite number of discontinuities. So-called boundary-adapted wavelets avoid this issue, and could be used in what follows. However, these require some rather intricate technical modifications to the proofs (for related work, see [2]).
Let \( \{\phi_n\}_{n=1}^\infty \) be an orthonormal basis of \( L^2([0,1]) \) (not necessarily of wavelet type). For \( s \geq 1 \), the \textit{linear \( s \)-term approximation} of a function \( f = \sum_{n=1}^\infty d_n \phi_n \in L^2([0,1]) \) is

\[
f_s^l = \sum_{k=1}^s d_k \phi_k,
\]

and the \textit{best \( s \)-term approximation} of \( f \) is

\[
f_s^{nl} = \sum_{n=1}^s d_{\pi(n)} \phi_{\pi(n)},
\]

where \( \pi : \mathbb{N} \rightarrow \mathbb{N} \) is a bijection that rearranges the coefficients \( d_n \) in nonincreasing order of absolute value, i.e. \( |d_{\pi(1)}| \geq |d_{\pi(2)}| \geq \ldots \). The \textit{linear} and \textit{best \( s \)-term approximation errors} are

\[
e_s(f)_{L^2} = \|f - f_s^l\|_{L^2} = \sqrt{\sum_{k>s} |d_k|^2}, \quad \sigma_s(f)_{L^2} = \|f - f_s^{nl}\|_{L^2} = \sqrt{\sum_{k>s} |d_{\pi(k)}|^2}, \tag{2.2}
\]

respectively.

The concern of this paper is nonlinear wavelet approximation in suitable spaces of piecewise regular functions. We now define these spaces:

**Definition 2.2.** Let \( 0 < \alpha \leq 1 \) and \(-\infty < a < b < \infty \). The H"older semi-norm of index \( \alpha \) of a function \( f \in C([a,b]) \) is

\[
|f|_{C^\alpha([a,b])} = \sup_{x,y \in [a,b], x \neq y} \left\{ \frac{|f(x) - f(y)|}{|x-y|^\alpha} \right\}.
\]

A function \( f \) is \( \alpha \)-H"older continuous if \( |f|_{C^\alpha([a,b])} < \infty \).

**Definition 2.3.** Let \(-\infty < a < b < \infty \), \( d \in \mathbb{N}_0 \), \( 0 < \beta \leq 1 \) and \( \alpha = d + \beta \). The H"older space \( C^\alpha([a,b]) \) consists of those functions \( f \) that are \( d \)-times continuously differentiable on \([a,b]\) and for which the \( d \)th derivative is \( \beta \)-H"older continuous. This is a Banach space with norm

\[
\|f\|_{C^\alpha([a,b])} = \sum_{j=0}^d \|f^{(j)}\|_{C([a,b])} + \|f^{(d)}\|_{C^\beta([a,b])},
\]

where \( \|g\|_{C([a,b])} = \sup_{x \in [a,b]} |g(x)| \).

**Definition 2.4.** For \( \alpha = d + \beta \), where \( d \in \mathbb{N}_0 \) and \( 0 < \beta \leq 1 \), the space \( PC^\alpha([a,b]) \) consists of all functions which are discontinuous at (at most) a finite number of points in \([a,b]\) and are \( \alpha \)-H"older continuous in between any two consecutive discontinuities.

We define the norm of such a function as the maximum of its \( C^\alpha \)-norm over all intervals of smoothness, and write \( \|f\|_{PC^\alpha([a,b])} \), or \( \|f\|_{PC^\alpha} \) when the domain of \( f \) is clear. For convenience, we also define

\[
\mathcal{N}(f) = |\{\text{discontinuities of } f \text{ in } (a,b)\}| + 1, \quad f \in PC^\alpha([a,b]),
\]

The following well-known result summarizes the effectiveness of wavelets for approximating piecewise smooth functions (see §5 for a proof):

9
Theorem 2.5. Suppose that \( f \in PC^\alpha([0, 1]) \), where \( \alpha = d + \beta \geq 1/2 \) for \( d \in \mathbb{N}_0 \) and \( 0 < \beta \leq 1 \). Consider the periodized Daubechies wavelet basis with \( p > d \) vanishing moments with coarsest scale \( j_0 \) given by (B.1). Then there exists a constant \( C_{p, \alpha} > 0 \) such that, for all \( s \geq 1 \),

\[
e_s(f)_{L^2} \leq C_{p, \alpha} \sqrt{N(f)} \| f \|_{PC^\alpha}^{-1} s^{-1/2},
\]

and, if \( s / \log_2(s) \geq 64 \alpha p^2 N(f) \),

\[
\sigma_s(f)_{L^2} \leq C_{p, \alpha} \| f \|_{PC^\alpha}^{-1} s^{-\alpha}.
\]

2.3 Standard compressed sensing theory

Recall that a vector \( x \in \mathbb{C}^N \) is \( s \)-sparse if it has at most \( 1 \leq s \leq N \) nonzero entries. Standard compressed sensing theory concerns the recovery of such a vector from measurements \( y = Ax \in \mathbb{C}^m \). Recovery can be carried out via a number of different procedures. However, in this paper we focus on convex optimization approaches, such as the basis pursuit (1.2). The overarching goal of compressed sensing is to derive conditions on the matrix \( A \in \mathbb{C}^{m \times N} \) under which any \( s \)-sparse \( x \) can be recovered via (1.2) from \( m \approx s \) measurements, up to log factors.

A particularly useful tool in this endeavour is the Restricted Isometry Property:

Definition 2.6. Let \( 1 \leq s \leq N \). The \( s \)th Restricted Isometry Constant (RIC) \( \delta_s \) of a matrix \( A \in \mathbb{C}^{m \times N} \) is the smallest \( \delta \geq 0 \) such that

\[
(1 - \delta) \| x \|_2^2 \leq \| Ax \|_2^2 \leq (1 + \delta) \| x \|_2^2, \quad \text{for all } s\text{-sparse } x. \tag{2.3}
\]

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

The following result is well known (see, for example, [21]):

Theorem 2.7. Suppose that \( A \in \mathbb{C}^{m \times N} \) satisfies the RIP of order \( 2s \) with constant

\[
\delta_{2s} < \sqrt{2} - 1. \tag{2.4}
\]

Let \( x \in \mathbb{C}^N \) and \( y = Ax \). Then any minimizer \( \hat{x} \in \mathbb{C}^N \) of (1.2) satisfies

\[
\| \hat{x} - x \|_{\ell^1} \leq C \frac{\sigma_s(x)_{\ell^1}}{\sqrt{s}}, \tag{2.5}
\]

where \( \sigma_s(x)_{\ell^1} = \min\{\| x - z \|_{\ell^1} : z \text{ is } s\text{-sparse}\} \) and the constant \( C > 0 \) depends on \( \delta_{2s} \) only.

We remark that the constant \( \sqrt{2} - 1 \) in (2.4) be improved (see [20]); however, this will be of little consequence for this paper. We also note that one can prove a recovery guarantee similar to this theorem for measurements that are also corrupted by noise. However, our focus in this paper is on the noise-free setting. Furthermore, it is well known random Gaussian matrices satisfy the RIP with high probability.

Theorem 2.8. Let \( 0 < \delta, \varepsilon < 1, 1 \leq s \leq N \) and

\[
m \geq c \cdot \left( s \cdot \log(eN/s) + \log(2\varepsilon^{-1}) \right). \tag{2.6}
\]

Let \( A \in \mathbb{R}^{m \times N} \) have i.i.d. entries drawn from the normal distribution with mean zero and variance \( 1/m \). Then, with probability at least \( 1 - \varepsilon \), \( A \) has the RIP of order \( s \) with \( \delta_s \leq \delta \).

The scaling \( m \geq c \cdot s \cdot \log(N/s) \) is essentially optimal. Any method which satisfies the error bound (2.5) must also have \( m \geq c \cdot s \cdot \log(N/s) \) [37, Chpt. 10].
3 Main results

We now present our main results. To state these, we use the language of encoders and decoders, terminology employed in the compressed sensing context, e.g., in [28].

3.1 Encoding and decoding

An $m$-term encoder is a linear mapping $\mathcal{E}_m : L^2([0,1]) \to \mathbb{C}^m$. An $m$-term decoder is a mapping $\mathcal{D}_m : \text{Ran}(\mathcal{E}_m) \to L^2([0,1])$. Note that $\mathcal{E}_m$ is by assumption linear, whereas $\mathcal{D}_m$ will typically be nonlinear. A Fourier encoder is an encoder of the form

$$\mathcal{E}_m(f) = (\hat{f}(\omega_i))_{i=1}^m,$$

where $\hat{f}$ is the Fourier transform of $f$ and $\omega_i \in \mathbb{R}$.

Random Gaussian measurements are popular in compressed sensing. Since in this paper we consider the approximation of functions (rather than finite vectors), we define a Gaussian encoder as any encoder of the form

$$\mathcal{E}_m(f) = A (\langle f, \phi_j \rangle)_{j=1}^N,$$

where $\{\phi_j\}_{j=1}^N$ is an orthonormal system in $L^2([0,1])$ and $A \in \mathbb{R}^{m \times N}$ is a random Gaussian matrix. Note that $\{\phi_j\}$ will typically coincide with the orthonormal basis that is used for decoding, i.e. a wavelet basis.

Given an encoder-decoder pair $(\mathcal{E}_m, \mathcal{D}_m)$ and $f \in L^2([0,1])$, we write $\tilde{f}_m = \mathcal{D}_m(\mathcal{E}_m(f))$.

3.2 Statements

We commence with Gaussian encoders:

**Theorem 3.1.** Let $0 < \varepsilon < 1$, $\alpha > 1/2$ and $N_* \in \mathbb{N}$. Then there exist constants $c, c_\alpha, C_\alpha > 0$ such that the following holds. For any $m \in \mathbb{N}$ satisfying

$$m \geq c \log(2/\varepsilon), \quad m/(\log(m))^2 \geq c_\alpha N_*,$$

the Gaussian encoder-decoder pair $(\mathcal{E}_m, \mathcal{D}_m)$ based on periodized Daubechies’ wavelets with $p = \lceil \alpha \rceil$ vanishing moments satisfies

$$\|f - \tilde{f}_m\|_{L^2} \leq C_\alpha \|f\|_{PC^\alpha} \left( \frac{\log(m)}{m} \right)^\alpha,$$

for all $f \in PC^\alpha([0,1])$ with $N(f) \leq N_*$, with probability at least $1 - \varepsilon$. The decoder requires the solution of a basis pursuit problem of size at most $m \times \lceil m^{2\alpha+1}/(\log(m))^{2\alpha+2} \rceil$.

We next demonstrate that the Gaussian encoder is far from optimal:

**Theorem 3.2.** Let $0 < \varepsilon < 1$, $\alpha > 1/2$ and $N_* \in \mathbb{N}$. Then there exist constants $c, c_\alpha, C_\alpha > 0$ such that the following holds. For any $m \in \mathbb{N}$ satisfying

$$m \geq c \log(2/\varepsilon), \quad m/(\log(m))^2 \geq c_\alpha N_*,$$
there exists an encoder-decoder pair \((E_m, D_m)\) such that, with probability at least \(1 - \varepsilon\),

\[
\|f - \hat{f}_m\|_{L^2} \leq C_\alpha \|f\|_{PC^\alpha} \frac{\sqrt{\log(m)}}{m^\alpha},
\]

(3.2)

for all \(f \in PC^\alpha([0, 1])\) with \(N(f) \leq N_*\). The decoder uses periodized Daubechies’ wavelets with \(p = \lceil \alpha \rceil\) vanishing moments and requires the solution of a basis pursuit problem of size at most \(m \times [m^{2\alpha+1}/(\log(m))^3]\).

In view of Theorem 2.5, this pair is optimal up to \(\sqrt{\log(m)}\); a significant improvement on the Gaussian case. This factor can be removed altogether at the price of a nonuniform recovery guarantee. This is discussed in §4.4.

**Remark 3.3** The above result ceases to hold for \(\alpha = 1/2\). If \(\alpha = 1/2\), then one can trivially find an encoder-decoder pair satisfying

\[
\|f - \hat{f}_m\|_{L^2} \leq C\sqrt{N(f)}\|f\|_{PC^{\alpha}} m^{-1/2}.
\]

Indeed, one merely directly senses the first \(m\) wavelet coefficients. This gives \(\|f - \hat{f}_m\|_{L^2} = \epsilon_m(f)_{L^2}\), and the bound follows from Theorem 2.5.

Next we present our main result for Fourier encoders:

**Theorem 3.4.** Let \(0 < \delta < 1\), \(\alpha > 1/2\), \(p \geq \lceil \alpha \rceil\), and \(N_* \geq 1\). Then, there exist constants \(c_{p,\alpha,\delta}, C_{p,\alpha} > 0\) such that the following holds. For any \(m \in \mathbb{N}\) such that

\[
m \geq c_{p,\alpha,\delta}(N_*)^2, \quad (\log(m))^5 \geq \log(\varepsilon^{-1}),
\]

there exists a Fourier encoder-decoder pair \((E_m, D_m)\) such that, with probability at least \(1 - \varepsilon\),

\[
\|f - \hat{f}_m\|_{L^2} \leq C_{p,\alpha} \|f\|_{PC^\alpha} \frac{(\log(m))^{1+(6+\delta)(q+\alpha+1/2)}}{m^{\alpha} (2q+1)^{\alpha}}.
\]

(3.3)

for all \(f \in PC^\alpha([0, 1])\) with \(N(f) \leq N_*\). The decoder uses periodized Daubechies’ wavelets with \(p\) vanishing moments, smoothness parameter \(q\) and requires solution of a weighted square-root LASSO problem of size no more than \(m \times [m^\sigma]\), where \(\sigma = \max\{\frac{\alpha}{\alpha-1/2}, 2\alpha+1\}\). In particular, if \(p\) is the smallest integer so that the smoothness parameter

\[
q \geq 6\alpha - 1/2 + \alpha - \frac{3}{2},
\]

(3.4)

then the encoder-decoder pair satisfies

\[
\|f - \hat{f}_m\|_{L^2} \leq C_{p,\alpha} \|f\|_{PC^\alpha} \frac{(\log(m))^{1+\delta}}{m^{\alpha} (2q+1)^{\alpha}}.
\]

(3.5)

for all \(f \in PC^\alpha([0, 1])\) with \(N(f) \leq N_*\).

The reader will notice that the decoder in the case solves a so-called weighted square-root LASSO problem. This problem takes the form

\[
\min_{z \in \mathbb{C}^N} \|z\|_{\ell^w} + \lambda \|Az - y\|_{\ell^2},
\]

(3.6)
where $\|z\|_{\ell^1} = \sum_{i=1}^{N} w_i |z_i|$ is a weighted $\ell^1$-norm and $\lambda > 0$ is a parameter. The reason for solving this problem instead of (1.2) is discussed in §4.3.

We stress that the encoder-decoder pairs in Theorems 3.1 and 3.2 are determined completely by $m$ and the smoothness $\alpha$. In Theorem 3.4 the Fourier encoder-decoder pair also depends on the number vanishing moments $p$ and the auxiliary parameter $\lambda$. In particular, neither $\mathcal{N}(f)$ nor $\mathcal{N}_s$ need to be known for any of these strategies.

Remark 3.5 These results are uniform in the sense that for each fixed (and sufficiently large) $m$, a single random draw of the matrix used in the encoding stage guarantees the corresponding approximation error bound for all $f \in PC^\alpha([0,1])$ with $\mathcal{N}(f) \leq \mathcal{N}_s$. The parameter $\mathcal{N}_s$ is included primarily for convenience. It, in combination with the condition on $m$, allows for a uniform error bound that is independent the number of discontinuities $\mathcal{N}(f)$.

4 Discussion

We now discuss these results in some more detail.

4.1 The idea of Theorems 3.1 and 3.2

Theorem 3.1 is proved as follows. To find a quasi-best $s$-term wavelet approximation one needs to search within the first $N$ wavelets, where $N \asymp s^{2\alpha+1}$. The measurement condition (2.6) now gives $m \asymp s \cdot \log(s)$, or equivalently, $s \asymp m / \log(m)$. After modifying the proof of Theorem 2.5, we show that $\sigma_s(P_N d)_{\ell^1} \lesssim s^{-(\alpha-1/2)}$, where $d \in \ell^2(\mathbb{N})$ is the vector of wavelet coefficients of $f$. The error bound now follows immediately from this and Theorem 2.7.

In contrast, Theorem 3.2 constructs a sampling strategy that directly measures the coarse wavelet scales and then randomly samples the fine scales. Specifically, the first $N_1 \approx m/2$ wavelet coefficients are directly sampled, and then for the wavelet coefficients in the range $\{N_1 + 1, \ldots, N_2\}$, where $N_2 \approx m^{2\alpha+1}$, we use random Gaussian sampling. By Theorem 2.7, the error is effectively determined by the term

$$\sigma_s(P_{N_1} d)_{\ell^1} \sqrt{s}, \tag{4.1}$$

where $s$ is any number satisfying $s \lesssim m / \log(m)$. As discussed in §1.3, there are very few significant coefficients in the range $\{N_1 + 1, \ldots, N_2\}$, roughly $O(\mathcal{N}(f) \log(m))$ in total. Due to the choice of $N_1$, this means that $\sigma_s(P_{N_1} d)_{\ell^1} \lesssim m^{-(\alpha-1/2)}$ whenever $\mathcal{N}(f) \log(m) \lesssim s \lesssim m / \log(m)$. However, in order for (4.1) to attain the optimal algebraic rate $m^{-\alpha}$, we need to make $s$ as large as possible, i.e. $s \approx m / \log(m)$. This leads directly to the $\sqrt{\log(m)}$ factor.

This approach improves significantly over the log term of the Gaussian case, but does not remove it completely. The reason can be traced to the $(\ell^2,\ell^1)$-instance optimality of Gaussian measurements [37, Chpt. 10]; that is, the fact that the bound (2.5) bounds the $\ell^2$-norm reconstruction error in terms of the $\ell^1$-norm best $s$-term error. As we show in §4.4, this log term can be removed by exploiting nonuniform $(\ell^2,\ell^2)$-instance optimality of Gaussian measurements. In this setting, (4.1) is replaced by $\sigma_s(P_{N_2} d)_{\ell^2}$ and the optimal rate $m^{-\alpha}$ follows by the same arguments.
4.2 The log factor in Theorem 3.4

We defer a more detailed explanation of the strategy behind this theorem to §10.1. However, let us briefly comment on how the log term arises. First, since we use Fourier measurements, the measurement condition that ensures a RIP-type property involves a significantly larger log factor than (2.6), roughly of size \( (\log(m))^6 \). As above, because of the lack of \((\ell^2, \ell^2)\)-instance optimality, we need to take a much larger sparsity to ensure the \( m^{-\alpha} \) rate than is strictly necessary to capture the large wavelet coefficients corresponding to the discontinuities. This effectively leads to a log term of the order \( \sqrt{(\log(m))^6} = (\log(m))^3 \) in the error bound.

Where do the remaining log terms come from? The source is the interference between wavelet scales and frequency bands discussed in §1.4. Coarse scale wavelets have small but nonzero components in the frequency bands corresponding to the fine scales. To control the effect of this interference, we need to fully sample a few more frequency bands than the number of saturated wavelet scales. The precise number of fully sampled bands corresponding to the saturated scales depends on the wavelet smoothness \( q \), and gives rise to the log factor seen in (3.3).

Nevertheless, we expect this log factor can be improved. In §11 we discuss how this might be achieved.

4.3 The decoder in Theorem 3.4

As noted, in Theorem 3.4 we solve (3.6) instead of basis pursuit. There are two aspects to this choice: the weights and the unconstrained formulation.

1) Weights. These are incorporated to control the effect of the interferences discussed above. The weights are constant on each wavelet scale, and in the \( k \)-th scale, are taken as roughly \( \sqrt{s/s_k} \), where \( s_k \) is the sparsity in that scale.

2) Unconstrained formulation. Unlike in Theorems 3.1 and 3.2, in Theorem 3.4 we have to deal with the effect of truncating the wavelet expansion \( \sum_{n=1}^{\infty} d_n \phi_n \) to its first \( N \) terms. This truncation introduces an error proportional to \( \| P_N^\perp d \|_{\ell^2} \), which can be considered as noise in the measurement vector \( y \). Such an error is, of course, unknown \textit{a priori}.

Unfortunately, standard decoders in compressed sensing for noisy measurements such as \textit{quadratically-constrained basis pursuit (QCBP)}

\[
\min_{z \in \mathbb{C}^M} \| z \|_{\ell^1} \text{ subject to } \| Az - y \|_{\ell^2} \leq \eta,
\]

typically require explicit upper bounds on the noise [19, 37].\(^3\) To avoid this, one may consider unconstrained optimization problems. One standard choice is the unconstrained LASSO

\[
\min_{z \in \mathbb{C}^N} \| z \|_{\ell^1} + \lambda \| Az - y \|_{\ell^2}^2.
\]

Unfortunately, it is well known that the optimal tuning parameter \( \lambda \) in LASSO depends on the \( \ell^2 \)-norm of the noise, rendering it unsuitable for this problem.

\(^3\)There are some theoretical results for QCBP in the presence of unknown noise [19, 31, 36, 68]. However, except in specific cases, these involve additional factors (so-called \textit{quotients}) which are difficult to estimate.
The square-root LASSO – a little known variant of its more famous cousin – was conceived specifically to overcome this issue [15]. It was introduced to the compressed sensing setting in [3]. Therein it was shown that the optimal tuning parameter is independent of the noise, thus rendering it suitable for our purposes. We note in passing that (3.6) can be solved efficiently via standard algorithms. For instance, Chambolle-Pock’s primal-dual algorithm [25].

It is common in Fourier imaging to construct a measurement matrix
\[ A = P_\Omega F \Phi^* , \]
where \( F \in \mathbb{C}^{N \times N} \) and \( \Phi \in \mathbb{C}^{N \times N} \) are the discrete Fourier and wavelet transforms respectively, and \( P_\Omega \in \mathbb{C}^{m \times N} \) restricts to the frequencies sampled. This formulation is readily amenable to fast computations. However, this approach commits a discretization error, since (as noted in §1.2) the measurements are of the continuous Fourier transform of \( f \). To avoid this error, we employ infinite-dimensional compressed sensing [5], and formulate \( A \) as \( P_\Omega U P_N \), where \( U \) is the cross-Grammian between the Fourier and wavelet bases of \( L^2([0,1]) \). We note in passing that this matrix also admits fast computations in FFT time [38].

### 4.4 An optimal nonuniform recovery guarantee

Our main results are uniform for each fixed \( m \) (see Remark 3.5). The log factor \( \sqrt{\log(m)} \) present in the error estimates of Theorem 3.2 can be removed at the price of having a nonuniform recovery guarantee, i.e. a guarantee that holds with high probability for a fixed function \( f \in PC^\alpha([0,1]) \) and not uniformly for every \( f \in PC^\alpha([0,1]) \).

The key element to prove this is a compressed sensing result concerning the so-called nonuniform \((\ell^2,\ell^2)\)-instance optimality of random Gaussian measurements. The following is a direct consequence of [37, Thm. 11.23].

**Theorem 4.1.** There exist constants \( 0 < c_1 < 1, c_2, C > 0 \) such that the following holds. Let \( 0 < \varepsilon < 1, x \in \mathbb{C}^N \) be a fixed vector, \( A \in \mathbb{R}^{m \times N} \) have i.i.d. entries drawn from the normal distribution with mean zero and variance \( 1/m \), and assume that \( c_1 m \leq N \) and

\[ m \geq c_2 (s \log(eN/m) + \log(5/\varepsilon)). \]

Then any minimizer \( \hat{x} \in \mathbb{C}^N \) of \((1.2)\) satisfies

\[ \|x - \hat{x}\|_{\ell^2} \leq C\sigma_s(x)_{\ell^2}, \]

with probability at least \( 1 - \varepsilon \), where \( \sigma_s(x)_{\ell^2} = \min\{\|x - z\|_{\ell^2} : z \text{ is } s\text{-sparse}\} \).

This result implies the following theorem, analogous to Theorem 3.2. The only differences with respect to the assumptions of Theorems 3.2 are the lack of uniformity with respect to \( f \in PC^\alpha([0,1]) \) and the slightly larger dimension of the basis pursuit problem.

**Theorem 4.2.** Let \( 0 < \varepsilon < 1 \) and \( \alpha > 1/2 \). Then there exist constants \( c, c_\alpha, C_\alpha > 0 \) such that the following holds. For each \( f \in PC^\alpha([0,1]) \) and \( m \in \mathbb{N} \) satisfying

\[ m \geq c \log(5/\varepsilon), \quad m/\log^2(m) \geq c_\alpha N(f), \]

there exists an encoder-decoder pair \((E_m, D_m)\) such that, with probability at least \( 1 - \varepsilon \),

\[ \|f - \tilde{f}_m\|_{L^2} \leq C\|f\|_{PC^\alpha m^{-\alpha}}, \]

(4.3)

The decoder uses periodized Daubechies’ wavelets with \( p = \lceil \alpha \rceil \) vanishing moments and requires the solution of a basis pursuit problem of size at most \( m \times \lfloor m^{2\alpha+1}/(\log(m))^2 \rfloor \).
4.5 Computational cost and the size of the optimization problems

It is important to consider the size of the optimization problem that needs to be solved in each case. This size is $m \times N$, where $N \leq \lfloor m^{2\alpha+1}/(\log(m))^{2\alpha+2} \rfloor$ in Theorem 3.1, $N \leq \lfloor m^{2\alpha+1}/(\log(m))^3 \rfloor$ in Theorem 3.2 and $N \leq \lfloor \max\{m^{2\alpha+1/2}, m^{2\alpha+1}\} \rfloor$ in Theorem 3.4. These choices have been made to balance the various error terms, and in particular, to remove any dependence on the number of discontinuities $\mathcal{N}(f)$.

Since the size of $N$ affects the memory and computational time required for decoding, it is worth dwelling on precisely how $N$ affects the approximation error. We have the following:

Theorem 4.3. Consider the setup of Theorem 3.1, except where the basis pursuit problem is of size at most $m \times N$ for some $m \leq N \leq m^{2\alpha+1}$. Then the error bound (3.1) is replaced by

$$
\|f - \tilde{f}_m\|_{L^2} \leq C_{\alpha}\|f\|_{PC^\alpha} \left( \frac{(\log(m))^\alpha}{m} + \sqrt{\frac{\mathcal{N}(f)}{N}} \right).
$$

For the optimal encoder-decoder, we have:

Theorem 4.4. Consider the setup of Theorem 3.2, except where the basis pursuit problem is of size at most $m \times N$ for some $m \leq N \leq m^{2\alpha+1}$. Then the error bound (3.2) is replaced by

$$
\|f - \tilde{f}_m\|_{L^2} \leq C_{\alpha}\|f\|_{PC^\alpha} \left( \frac{\sqrt{\log(m)}}{m^\alpha} + \sqrt{\frac{\mathcal{N}(f)}{N}} \right).
$$

Similarly, we have a nonuniform result analogous to Theorem 4.4.

Theorem 4.5. Consider the setup of Theorem 4.2, except where the basis pursuit problem is of size at most $m \times N$ for some $m \leq N \leq m^{2\alpha+1}$. Then the error bound (4.3) is replaced by

$$
\|f - \tilde{f}_m\|_{L^2} \leq C_{\alpha}\|f\|_{PC^\alpha} \left( m^{-\alpha} + \sqrt{\frac{\mathcal{N}(f)}{N}} \right).
$$

Finally, for the Fourier encoder-decoder, we have:

Theorem 4.6. Consider the setup of Theorem 3.4, except where the square-root LASSO problem is of size at most $m \times N$ for some $m \leq N \leq m^\sigma$, where $\sigma = \max\{\frac{\alpha}{2\alpha+1}, 2\alpha+1\}$. Then the error bounds (3.3) and (3.5) are replaced by

$$
\|f - \tilde{f}_m\|_{L^2} \leq C_{P,\alpha}\|f\|_{PC^\alpha} \left( m^{-\alpha} + N^{-(\alpha-1/2)} + \sqrt{\frac{\mathcal{N}(f)}{N}} \right) (\log(m))^{\frac{1}{4} + \frac{(\delta+\delta/2)(1+\alpha+1/2)}{2(\alpha+1)}},
$$

and

$$
\|f - \tilde{f}_m\|_{L^2} \leq C_{P,\alpha}\|f\|_{PC^\alpha} \left( m^{-\alpha} + N^{-(\alpha-1/2)} + \sqrt{\frac{\mathcal{N}(f)}{N}} \right) (\log(m))^{\frac{2}{4} + \delta},
$$

respectively.

Notice that Theorems 3.1, 3.2, 3.4, and 4.2 are just corollaries of Theorems 4.3, 4.4, 4.6, and 4.5 respectively, obtained by setting $N = \lfloor m^{2\alpha+1}/(\log(m))^{2\alpha+2} \rfloor$, $N = \lfloor m^{2\alpha+1}/(\log(m))^3 \rfloor$, 

16
\[ N = \max\{m^{\frac{\alpha}{2}}, m^{2\alpha+1}\} \text{ or } N = \lfloor m^{2\alpha}/(\log(m))^2 \rfloor. \]

For instance, in the case of Theorem 3.2 this choice of \( N \) gives
\[
\sqrt{\frac{\mathcal{N}(f)}{N}} \leq \frac{\sqrt{\log(m)}}{m^{\alpha}} \sqrt{\frac{\mathcal{N}(f)(\log(m))^2}{m}} \leq C_{\alpha} \frac{\sqrt{\log(m)}}{m^{\alpha}},
\]
which yields the corresponding error bound in Theorem 3.2 (note here we use the assumed condition \( m/ \log^{2}(m) \geq c_{\alpha} N_{*} \)).

However, the above results also show that the error bounds can be obtained using asymptotically smaller \( N \) – respectively \( N = \lfloor m^{\frac{2\alpha}{2}}/(\log(m))^2 \rfloor \), \( N = \lfloor m^{\frac{\alpha}{2}}/(\log(m))^2 \rfloor \) – at the expense of having the additional factor depending on \( N(f) \) in the error bound. For instance, making this choice in Theorem 4.4 results in the error bound
\[
\|f - \tilde{f}_{m}\|_{L^{2}} \leq C_{\alpha}\|f\|_{PC^{\alpha}} \left(1 + \sqrt{\frac{\mathcal{N}(f)}{m\alpha}}\right) \sqrt{\frac{\log(m)}{m^{\alpha}}}.
\]

5 Proof of Theorem 2.5

While numerous variants on Theorem 2.5 exist in the literature on wavelet approximation, we include a short proof in order to make the dependence on \( N(f) \) explicit, and also for exposition, since the key ideas in the proof will be used in subsequent arguments.

In what follows, we will make use of the periodic extension \( f^{\text{ext}} \) of a function \( f \), defined as
\[
f^{\text{ext}}(x + k) = f(x), \quad \forall x \in [0, 1), \forall k \in \mathbb{Z}.
\]
Notice that for every \( f \in L^{\infty}([0, 1]) \) and \( g \in L^{1}(\mathbb{R}) \), this satisfies
\[
\langle f^{\text{ext}}, g \rangle_{L^{2}(\mathbb{R})} = \langle f, g^{\text{per}} \rangle_{L^{2}([0,1])},
\]
where \( g^{\text{per}} \) is the periodization of \( g \), defined by
\[
g^{\text{per}}(x) = \sum_{k \in \mathbb{Z}} g(x + k) \tag{5.1}
\]

We first require the following lemma:

**Lemma 5.1.** Suppose that \( f \in PC^{\alpha}([0, 1]) \), where \( \alpha = d + \beta \) for some \( d \in \mathbb{N}_{0} \) and \( 0 < \beta \leq 1 \). Consider the periodized Daubechies’ wavelet basis with \( p > d \) vanishing moments and coarsest scale \( j_{0} \) given by (B.1). Then, there exists a constant \( C_{p,\alpha} > 0 \) such that
\[
|\langle f, \psi_{j,n}^{\text{per}} \rangle| \leq C_{p,\alpha}\|f\|_{PC^{\alpha}} 2^{-(\alpha+1)/j} ,
\]
whenever the interval \([(n - p + 1)/2^{j}, (n + p)/2^{j}] = \text{supp}(\psi_{j,n})\) contains no discontinuities of \( f^{\text{ext}} \) and
\[
|\langle f, \psi_{j,n}^{\text{per}} \rangle| \leq C_{p,\alpha}\|f\|_{PC^{\alpha}} 2^{-j/2},
\]
otherwise. The constant can be chosen as
\[
C_{p,\alpha} = \max \left\{ \frac{1}{d!} \int_{-p+1}^{p} |\psi(x)||x|^\alpha dx, \sqrt{2p-1} \right\}.
\]
Proof. First, we observe that for every interval $I \subseteq \mathbb{R}$ the periodic extension of $f$ satisfies $f_{\text{ext}} \in PC^\alpha(I)$ and $\|f_{\text{ext}}\|_{PC^\alpha(I)} \leq \|f\|_{PC^\alpha([0,1])}$.

Let us consider the case where $I := \text{supp}(\psi_{j,n})$ contains no discontinuities of $f_{\text{ext}}$. In that case, $f_{\text{ext}} \in C^\alpha(I)$ and, using Taylor’s theorem it is not difficult to show that there exists a polynomial $T$ of degree $d$ such that

$$|f_{\text{ext}}(x) - T(x)| \leq \frac{1}{d!}\|f_{\text{ext}}\|_{C^\alpha(I)}|x - n/2^j|^\alpha, \quad \forall x \in I.$$ 

As a consequence, by using the fact that $\psi$ has $p > d$ vanishing moments, we have

$$|\langle f, \psi_{j,n}^{\text{per}} \rangle| = |\langle f_{\text{ext}} - T, \psi_{j,n} \rangle| \leq \frac{1}{d!}\|f_{\text{ext}}\|_{C^\alpha(I)} 2^{-(\alpha + 1/2)} j \int_{-p+1}^{p} |\psi(x)||x|^\alpha dx,$$

which implies the first estimate.

When $I := \text{supp}(\psi_{j,n})$ contains at least one discontinuity of $f_{\text{ext}}$, we have

$$|\langle f, \psi_{j,n}^{\text{per}} \rangle| = |\langle f_{\text{ext}}, \psi_{j,n} \rangle| \leq \|f_{\text{ext}}\|_{C(I)}\|\psi_{j,n}\|_{L^1} \leq \|f\|_{C([0,1])} 2^{-j/2} \sqrt{2p - 1}.$$

This yields the second estimate and concludes the proof.

Proof of Theorem 2.5. Consider $e_s(f)_{L^2}$. Without loss of generality, we may assume that $s \geq 2^{j_0+1}$ (up to choosing $C_{p,\alpha}$ large enough). Hence there exists a $j_1 \geq 1$ such that $2^{j_0+j_1} < s < 2^{j_0+j_1+1}$. Due to the ordering (B.3), the linear approximation of $f$ of order $s$ includes at least all the scaling function coefficients $\langle f, \varphi_{j_0,n}^{\text{per}} \rangle_{L^2}$ and the wavelet coefficients $\langle f, \psi_{j,n}^{\text{per}} \rangle_{L^2}$ with $j = j_0, \ldots, j_0 + j_1 - 1$. Therefore

$$(e_s(f)_{L^2})^2 \leq \sum_{j \geq j_0 + j_1} \sum_{n=0}^{2^j-1} |\langle f, \psi_{j,n}^{\text{per}} \rangle|^2.$$ 

Since $f$ has $N(f) - 1$ discontinuities, the extension $f_{\text{ext}}$ has at most $2pN(f)$ discontinuities on the interval $[-p+1,p]$, and therefore there are at most $(2p)^2 N(f)$ wavelets $\psi_{j,n}$ at any fixed scale $j$ whose support contains a discontinuity of $f_{\text{ext}}$. Since there are no more than $2^j$ wavelets at each scale $j$ supported in smooth regions of $f$, we deduce that

$$(e_s(f)_{L^2})^2 \leq C_{p,\alpha}\|f\|^2_{PC^\alpha} \sum_{j \geq j_0 + j_1} (N(f)2^{-j} + 2^{-2\alpha j}) \leq C_{p,\alpha}\|f\|^2_{PC^\alpha}N(f)2^{-j_0-j_1},$$

where we have also used the fact that $\alpha \geq 1/2$. Since $s \geq 2^{j_0+j_1}$ the result now follows.

Now consider $\sigma_s(f)_{L^2}$. The idea is to judiciously choose $s$ wavelet coefficients so as to obtain the desired error decay rate. To this end, note that, by assumption, $s \geq 2^{j_0+1}$ and suppose that $\{I_j\}_{j \geq j_0}$ is a collection of index sets with $I_j \subseteq \{0, \ldots, 2^j - 1\}$ that satisfies

$$\left| \bigcup_{j \geq j_0} I_j \right| \leq s - 2^{j_0}, \quad (5.2)$$
(the subtraction of $2^j$ takes into account the inclusion of the scaling coefficients). Then

$$
(\sigma_s(f)_{L^2})^2 \leq \sum_{j=j_0}^{\infty} \sum_{\substack{0 \leq k < 2^j \setminus k \notin I_j}} |(f, \psi_{j,k}^{\text{per}})|^2.
$$

(5.3)

We now define the index sets $\{I_j\}_{j \geq j_0}$. First, for $j_1, j_2 \in \mathbb{N}$ such that $1 \leq j_1 \leq j_2$, which will be chosen in a moment, let

$$
I_j = \{0, \ldots, 2^j - 1\}, \quad j_0 \leq j < j_0 + j_1,
$$

$$
I_j = \emptyset, \quad j \geq j_0 + j_2,
$$

and for $j_0 + j_1 \leq j < j_0 + j_2$ let $I_j$ be the index set consisting of those values of $n$ where the support of $\psi_{j,n}$ contains a discontinuity of $f^{\text{ext}}$. Recall that, in this last case, $|I_j| \leq (2p)^2 \mathcal{N}(f)$ from earlier in the proof. Notice that

$$
\left| \bigcup_{j \geq j_0} I_j \right| = \sum_{j=j_0}^{j_0 + j_1 - 1} 2^j + \sum_{j=j_0 + j_1}^{j_0 + j_2 - 1} \left| I_j \right| \leq 2^{j_0 + j_1} + (2p)^2 \mathcal{N}(f)(j_2 - j_1).
$$

(5.5)

We now examine the error. From (5.3) and the definition of the $I_j$, we have

$$
(\sigma_s(f)_{L^2})^2 \leq \sum_{j=j_0}^{j_0 + j_2 - 1} \sum_{\substack{0 \leq k < 2^j \setminus k \notin I_j}} |(f, \psi_{j,k}^{\text{per}})|^2 + (\sigma_s(f)_{L^2})^2,
$$

where $\tilde{s} = 2^{j_0 + j_2}$. Hence, using Lemma 5.1 and the first part of the theorem, we deduce that

$$
(\sigma_s(f)_{L^2})^2 \leq C_{p,\alpha} \|f\|_{PC^\alpha}^2 \left( \sum_{j=j_0 + j_1}^{j_0 + j_2 - 1} 2^{-2\alpha j} + \mathcal{N}(f)2^{-(j_0 + j_2)} \right)
$$

$$
\leq C_{p,\alpha} \|f\|_{PC^\alpha}^2 \left( 2^{-2\alpha (j_0 + j_1)} + \mathcal{N}(f)2^{-(j_0 + j_2)} \right).
$$

Now note that, by assumption, $s \geq 2^{j_0 + 2}$. Set $j_0 + j_1 = \lfloor \log_2(s/2) \rfloor \geq 1$ so that $s/4 \leq 2^{j_0 + j_1} \leq s/2$, and let $j_0 + j_2 = \lfloor (2\alpha + 1) \log_2(s) \rfloor$ so that $s^{2\alpha + 1}/2 \leq 2^{j_0 + j_2} \leq s^{2\alpha + 1}$. This gives

$$
(\sigma_s(f)_{L^2})^2 \leq C_{p,\alpha} \|f\|_{PC^\alpha}^2 s^{-2\alpha}(1 + \mathcal{N}(f)s^{-1}).
$$

Notice that $j_0 + j_2 > j_0 + j_1$ since $\alpha \geq 1/2$ and $s \geq 2^{j_0 + 1} \geq 2$, hence this choice is valid.

It remains to verify that (5.2) holds for this choice of index set. Substituting the values of $j_1$ and $j_2$ into (5.5) and observing that $j_2 - j_1 \leq 2\alpha \log_2(s) + 2 \leq 4\alpha \log_2(s)$ (since $s \geq 2^{j_0 + 2} \geq 4$ and $\alpha \geq 1/2$), we see that

$$
\left| \bigcup_{j \geq j_0} I_j \right| \leq \frac{s}{2} + 16 \alpha p^2 \mathcal{N}(f) \log_2(s).
$$

Therefore, recalling again that $s \geq 2^{j_0 + 2}$, condition (5.2) holds since

$$
\frac{s}{\log_2(s)} \geq 64 \alpha p^2 \mathcal{N}(f) \implies \frac{s}{2} + 16 \alpha p^2 \mathcal{N}(f) \log_2(s) \leq \frac{s}{2} + \frac{s}{4} \leq s - 2^{j_0}.
$$

Moreover, the assumption on $s$ implies that $\mathcal{N}(f)s^{-1} \leq 1$. This completes the proof. \qed
6 Proof of Theorems 3.1 and 4.3

We now move on to the proofs of our main theorems. For each, we first give a recipe that describes the inputs (the number of measurements $m$ and the smoothness parameter $\alpha$), the various parameters for the encoder and decoder, and then the encoder and decoder themselves.

6.1 Recipe

The following recipe applies to Theorem 3.1:

Inputs: Number of measurements $m$, smoothness parameter $\alpha$.

Parameters: Let
- $p = \lceil \alpha \rceil$, $j_0$ be as in (B.1);
- $\{\varphi_{j_0,k}^{\text{per}}\} \cup \{\psi_{j,k}^{\text{per}}\}$ be the periodized Daubechies’ wavelet basis with $p$ vanishing moments;
- $r = \lfloor \log_2(m^{2\alpha+1}/(\log(m))^{2\alpha+2}) \rfloor - j_0$;
- $N = 2^{j_0+r}$;
- $A \in \mathbb{R}^{m \times N}$ be a random Gaussian matrix.

Encoder: Define $E_m(f) = A(\langle f, \varphi_{j_0+1} \rangle)$, where $\{\varphi_{j}^{\text{per}}\}_{j=1}^{N}$ are the periodized Daubechies’ wavelets up to scale $j_0 + r - 1$ (see (B.3)).

Decoder: Given measurements $y = E_m(f)$, define $D_m(y)$ as

$$D_m(y) = \sum_{j=1}^{N} \tilde{d}_j \varphi_j,$$

where $\tilde{d} = (\tilde{d}_j)_{j=1}^{N} \in \mathbb{C}^N$ is any solution of the basis pursuit problem

$$\min_{z \in \mathbb{C}^N} \|z\|_{\ell^1} \text{ subject to } Az = y.$$

The case of Theorem 4.3: We now assume $N$ is an input rather than a parameter, and change the definition of $r$ to $r = \lfloor \log_2(N) \rfloor - j_0$. Note that

$$N/2 \leq 2^{j_0+r} \leq N.$$

Hence, if necessary, we replace $N$ by $2^{j_0+r}$ so that the encoder-decoder pair includes all the wavelets up to scale $j_0 + r - 1$. This pair is then defined in exactly the same way as above.

6.2 Proof of Theorem 4.3

Notice that $y = E_m(f) = A P_N d$ for this encoder, where $d$ is the infinite vector of wavelet coefficients of $f$. Hence, $\tilde{d}$ is a minimizer of the problem

$$\min_{z \in \mathbb{C}^N} \|z\|_{\ell^1} \text{ subject to } Az = A P_N d.$$
Suppose that \( m \) satisfies
\[
m \geq c_1(s \log(eN/s) + \log(2/\varepsilon)),
\]
for a suitable \( 1 \leq s \leq N \), where \( c_1 > 0 \) is a constant. Then, by Theorems 2.7 and 2.8 we have
\[
\|P_N d - \tilde{d}\|_{L^2} \leq C \sigma_s(P_N d)_{L^1}/\sqrt{s},
\]
with probability at least \( 1 - \varepsilon \). Define \( s \) by
\[
s = \left\lfloor \frac{m}{2c_1(2\alpha + 1) \log(m)} \right\rfloor.
\]
Recall that \( N \leq m^{2\alpha + 1} \) by assumption and also that \( s \geq e \) for all \( m \geq c_\alpha \). Hence
\[
c_1(s \log(eN/s) + \log(2/\varepsilon)) \leq c_1 \frac{m}{2c_1(2\alpha + 1) \log(m)} \log(m^{2\alpha + 1}) + c_1 \log(2/\varepsilon) \leq \frac{m}{2} + \frac{m}{2} = m,
\]
where in the final step we recall that \( m \geq c \log(2/\varepsilon) \) by assumption.

We next estimate \( \sigma_s(P_N d)_{L^1} \) using arguments analogous to those employed in the proof of Theorem 2.5. First, we select all coefficients in scales \( j_0, \ldots, j_0 + \bar{r} - 1 \) where \( j_0 + \bar{r} = \left\lfloor \log_2(s/2) \right\rfloor \) (note that \( j_0 + \bar{r} \geq j_0 + 1 \) as soon as \( s \geq 2^{j_0 + 3} \), which is guaranteed by choosing \( m \geq c_\alpha \)). Next, in the remaining levels, we select all the coefficients corresponding to wavelets intersecting the discontinuities of \( f^{ext} \). Then there are at most \( (2p)^2 \mathcal{N}(f) \) of such coefficients at any given scale. Hence, the total number of such coefficients is bounded by
\[
(2p)^2 \mathcal{N}(f) (r - \bar{r}) \leq (2p)^2 \mathcal{N}(f) \log_2(N) \leq c_\alpha' \mathcal{N}(f) \log(m),
\]
which is at most \( s/2 \) for \( m/\log^2(m) \geq c_\alpha \mathcal{N}(f) \). Hence Lemma 5.1 gives
\[
\sigma_s(P_N d)_{L^1} \leq C_\alpha \|f\|_{PC^{\alpha}} \left( \sum_{j = j_0 + \bar{r}}^{j_0 + r} 2^{-(\alpha + 1/2)j} \cdot 2^j \right)
\leq C_\alpha \|f\|_{PC^{\alpha}} \cdot 2^{-(\alpha - 1/2)(j_0 + \bar{r})}
\leq C_\alpha \|f\|_{PC^{\alpha}} s^{-(\alpha - 1/2)}.
\]
In the last inequality we have used the fact that \( \alpha > 1/2 \). Therefore
\[
\|P_N d - \tilde{d}\|_{L^2} \leq C_\alpha \|f\|_{PC^{\alpha}} s^{-\alpha} \leq C_\alpha \|f\|_{PC^{\alpha}} \left( \frac{\log(m)}{m} \right)^\alpha.
\]
By Theorem 2.5, the linear approximation error satisfies
\[
e_N(f)_{L^2} \leq C_\alpha \|f\|_{PC^{\alpha}} \sqrt{\mathcal{N}(f)/N}
\]
Observing that
\[
\|f - \tilde{f}_m\|_{L^2} \leq \|P_N d - \tilde{d}\|_{L^2} + e_N(f)_{L^2}.
\]
concludes the proof.
6.3 Proof of Theorem 3.1

In this case \( N = 2^{j_0 + r} \), where \( r = \lfloor \log_2(m^{2\alpha + 1}/(\log(m))^{2\alpha + 2}) \rfloor - j_0 \). Observe that

\[
\frac{m^{2\alpha + 1}}{2(\log(m))^{2\alpha + 2}} \leq N \leq \frac{m^{2\alpha + 1}}{(\log(m))^{2\alpha + 2}}.
\]

We now apply Theorem 4.3 and use the fact that \( c_\alpha N(f) \leq \frac{m}{(\log(m))^2} \) by assumption.

7 Proof of Theorems 3.2, 4.2, 4.4 and 4.5

We now move on to the optimal sampling strategy. Recall from §4.1 that the key here is to directly sample up to a certain scale \((\bar{r} + j_0)\) below and then subsample the remaining scales with a random Gaussian matrix.

7.1 Recipe

We first consider Theorems 3.2 and 4.2 (the recipe is the same for both):

Inputs: number of measurements \( m \), smoothness parameter \( \alpha \).

Parameters: Let

- \( p = \lceil \alpha \rceil, j_0 \) be as in (B.1);
- \( \{\varphi_{j_0,k}^{\text{per}}\} \cup \{\psi_{j,k}^{\text{per}}\} \) be the periodized Daubechies’ wavelet basis with \( p \) vanishing moments;
- \( r = \lfloor (2\alpha + 1) \log_2(m) \rfloor - j_0; \)
- \( \bar{r} = \lfloor \log_2(m/2) \rfloor - j_0; \)
- \( N_1 = 2^{j_0 + \bar{r}}, N_2 = 2^{j_0 + r}; \)
- \( m_2 = m - m_1, \) where \( m_1 = N_1 = 2^{j_0 + \bar{r}}; \)
- \( A = (a_{il})_{i,l=1}^{m_2,N_2-N_1} \in \mathbb{R}^{m_2 \times (N_2-N_1)} \) be a random Gaussian matrix.

This choice of parameters requires that \( \bar{r} > 0 \) and that \( \bar{r} < r \). However, we note that the former holds for every \( m \geq 2^{j_0 + 2} \) and the latter holds for every \( \alpha > 1/2 \).

Encoder: Write \( E_m(f) = \left( E_{m_1}(f), E_{m_2}(f) \right), \) where

\[
E_{m_l}^{(l)}(f) = \left( e^{(l)}_{i_l}(f) \right)_{i_l=1}^{m_l} \in \mathbb{C}^{m_l}, \quad l = 1, 2.
\]

For \( l = 1 \), define

\[
e^{(1)}_{k+1}(f) = \langle f, \varphi_{j_0,k}^{\text{per}} \rangle_{L^2}, \quad k = 0, \ldots, 2^{j_0} - 1,
\]
\[
e^{(1)}_{2^{j_0}+k+1}(f) = \langle f, \psi_{j,k}^{\text{per}} \rangle_{L^2}, \quad k = 0, \ldots, 2^j - 1, \quad j = j_0, \ldots, j_0 + j_1 - 1.
\]

For \( l = 2 \), we let

\[
\chi_i = \sum_{j=\bar{r}}^{j_0} \sum_{k=0}^{2^j - 1} a_{i,2^j-r+k} \psi_{j,k}^{\text{per}}, \quad i = 1, \ldots, m_2,
\]
and define

\[ e_i^{(2)}(f) = \langle f, \chi_i \rangle_{L^2}, \quad i = 1, \ldots, m_2. \]

**Decoder:** Given measurements \( \mathcal{E}_m(f) \) as above, define

\[
D_m(f) = \sum_{k=0}^{2^{j_0}-1} e_{k+1}^{(1)}(f) \varphi_{j_0+k}^{\text{per}} + \sum_{j=j_0}^{\tilde{r}-1} \sum_{k=0}^{2^j-1} e_{2^j+k+1}^{(1)}(f) \psi_{j,k}^{\text{per}} + \sum_{j=\tilde{r}}^{r-1} \sum_{k=0}^{2^j-1} \tilde{d}_{2^j-r+k} \psi_{j,k}^{\text{per}},
\]

where the vector \( \tilde{d} = (\tilde{d}_i)_{i=1}^{N_2-N_1} \) is any minimizer of the basis pursuit problem

\[
\min_{z \in \mathbb{C}^{N_2-N_1}} \|z\|_{\ell^1} \text{ subject to } Az = \mathcal{E}_m^{(2)}(f). \tag{7.1}
\]

**The case of Theorems 4.4 and 4.5:** As previously, we let \( N \) be an input rather than a parameter, and change the definition of \( r \) to \( r = \lfloor \log_2(N) \rfloor - j_0 \). We now set \( N_2 = 2^{j_0+r} \) and note that

\[
N/2 \leq N_2 \leq N,
\]

so that the encoder-decoder pair includes all the wavelets up to the scale \( j_0 + r - 1 \). This pair is then defined in exactly the same way.

**7.2 Proof of Theorem 4.4**

Let \( d \in \ell^2(\mathbb{N}) \) be the wavelet coefficients of \( f \in L^2([0,1]) \). By Parseval’s identity, and the fact that \( \tilde{f}_m \) recovers the coarse scale coefficients exactly, we have

\[
\|f - \tilde{f}_m\|_{L^2} \leq \|P_{N_2}^N d - \tilde{d}\|_{\ell^2} + \|P_{N_2}^N \tilde{d}\|_{\ell^2}. \tag{7.2}
\]

Let \( \tilde{d} = (\tilde{d}_i)_{i=1}^{N_2-N_1} = P_{N_2}^N d \). Then,

\[
e_i^{(2)}(f) = \sum_{l=1}^{N_2-N_1} a_{i,l} \tilde{d}_l,
\]

and therefore \( \mathcal{E}_m^{(2)}(f) = A\tilde{d} \). Hence (7.1) is equivalent to

\[
\min_{z \in \mathbb{C}^{N_2-N_1}} \|z\|_{\ell^1} \text{ subject to } Az = A\tilde{d}.
\]

Since \( \tilde{d} \) is a minimizer, we deduce from Theorems 2.7 and 2.8 that there exists \( C > 0 \) such that

\[
\|P_{N_2}^N d - \tilde{d}\|_{\ell^2} \leq C \frac{\sigma_s \left( P_{N_2}^N d \right)_{\ell^2}}{\sqrt{s}}, \tag{7.3}
\]

with probability at least \( 1 - \varepsilon \), where \( s \in \mathbb{N} \) is any number such that

\[
m_2 \geq c_1 \left( s \log(e(N_2 - N_1)/s) + \log(2/\varepsilon) \right). \tag{7.4}
\]
We now choose $s$ as follows:

$$s = \left\lfloor \frac{m}{4c_1(2\alpha + 1) \log(m)} \right\rfloor. \tag{7.5}$$

Observe that $m_1 \leq m/2$ by construction and, consequently, $m_2 \geq m/2$. Also, since $N_2 \leq N \leq m^{2\alpha+1}$, we have

$$c_1 (s \log(e(N_2 - N_1)/s) + \log(2/\varepsilon)) \leq c_1 \frac{m}{4c_1(2\alpha + 1) \log(m)} \log(em^{2\alpha+1}/s) + c_1 \log(2/\varepsilon)$$

$$\leq \frac{m}{4} + \frac{m}{4} = m/2 \leq m_2,$$

hence this choice of $s$ is valid. Here, in the second step we used the facts that $s \geq e$ for $m/\log(m) \geq c_\alpha$ and $c_1 \log(2/\varepsilon) \leq m/4$ for $m \geq c \log(2/\varepsilon)$.

We now consider (7.3). Recall that $P_{j_0}^N d$ is the vector of wavelet coefficients at scales $j_0 + \bar{r}, \ldots, j_0 + r - 1$. As previously, we select all coefficients corresponding to wavelets intersecting the discontinuities of $f^{\text{ext}}$. The total number of such coefficients is at most

$$(2p)^2 N(f)(r - \bar{r}) \leq (2p)^2 N(f) \log_2(N) \leq 12(\alpha + 1)^3 N(f) \log(m),$$

where we used that $p = \lceil \alpha \rceil \leq \alpha + 1$ and the assumption $N \leq m^{2\alpha+1}$. Recalling (7.5) and the assumption on $m$, we note that the number of such coefficients does not exceed $s$. Indeed, under (7.5), we have

$$\frac{m}{\log^2(m)} \geq 48c_2(\alpha + 1)^4 N_\ast \implies 12(\alpha + 1)^3 N(f) \log(m) \leq \frac{1}{2}\frac{m}{c_2(2\alpha + 1) \log(m)} \leq s.$$

Hence, we may exclude all these slowly decaying coefficients, and, using Lemma 5.1, bound the best $s$-term approximation error by

$$\sigma_s\left(P_{N_2}^N d\right)_{L_1} \leq C_\alpha \|f\|_{PC^\alpha} \sum_{j=j_0+r}^{\infty} \sum_{k=0}^{2^j-1} 2^{-(\alpha+1/2)j}$$

$$\leq C_\alpha \|f\|_{PC^\alpha} 2^{-(\alpha-1/2)(j_0+r)} \leq C_\alpha \|f\|_{PC^\alpha} m^{-(\alpha-1/2)}.$$

In the penultimate step we have used the fact that $\alpha > 1/2$ and that $2j_0+r \geq m/4$. This gives

$$\|P_{N_2}^N d - \tilde{d}\|_{L_2} \leq C_\alpha \|f\|_{PC^\alpha} \frac{m^{-(\alpha-1/2)}}{\sqrt{s}} \leq C_\alpha \|f\|_{PC^\alpha} \sqrt{\log(m)m^{-\alpha}}.$$

To complete the proof, we now use (7.2), after noting that $\|P_{N_2}^N d\|_{L_2} = e_{N_2}(f)_{L^2}$, and therefore

$$\|P_{N_2}^N d\|_{L_2} \leq C_\alpha \|f\|_{PC^\alpha} \sqrt{N(f)/N_2} \leq C_\alpha \|f\|_{PC^\alpha} \sqrt{N(f)/N}, \tag{7.6}$$

by Theorem 2.5.

### 7.3 Proof of Theorem 4.5

The argument is similar to the proof of Theorem 4.4. First, we have

$$\|f - \tilde{f}_m\|_{L^2} \leq \|P_{N_2}^N d - \tilde{d}\|_{L_2} + \|P_{N_2}^N d\|_{L_2},$$

24
where \(d, \tilde{d} \in \mathbb{C}^{N_2 - N_1}\) are as defined therein. Theorem 4.1 implies that
\[
\|P_{N_1}^N d - \tilde{d}\|_2 \leq C\sigma_s(P_{N_1}^N d)_2, \tag{7.7}
\]
with probability at least \(1 - \varepsilon\), provided \(N_2 - N_1 \geq c_1m_2\) and
\[m_2 \geq c_2(s\log(e(N_1 - N_1)/m) + \log(5/\varepsilon)).\]
Note that \(N_2 - N_1 \geq m^{2\alpha + 1}/2 - m/2\). Hence, since \(m_1 \geq m/4\) by construction and, consequently, \(m_2 \leq 3m/4\), condition \(N_2 - N_1 \geq c_1m_2\) holds whenever \(m \geq c\), which is implied by the assumptions on \(m\). Now, we let
\[
s = \left\lfloor \frac{m}{4c_2(2\alpha + 1)\log(m)} \right\rfloor.
\]
Arguing as in Theorem 4.4, the assumptions \(m/\log(m) \geq c_\alpha\) and \(m \geq c\log(5/\varepsilon)\) give
\[c_2(s\log(e(N_1 - N_1)/m) + \log(5/\varepsilon)) \leq c_2(s\log(e(N_1 - N_1)/s) + \log(5/\varepsilon)) \leq m_2.
\]
Hence this choice is valid. Now, arguing as before, we deduce that
\[
\sigma_s(P_{N_1}^N d)_2 \leq C\alpha\|f\|_{PC^\alpha}m^{-\alpha}.
\]
Combining this with (7.6) now yields the result.

### 7.4 Proof of Theorem 3.2

We let \(N = 2^{j_0 + r}\), where \(r = [\log_2(m^{2\alpha + 1}/(\log(m))^3)] - j_0\). Then
\[
\frac{m^{2\alpha + 1}}{2(\log(m))^3} \leq N \leq \frac{m^{2\alpha + 1}}{(\log(m))^3}.
\]
We now apply Theorem 4.3 with this choice of \(N\) and using the fact that \(N(f) \leq c_\alpha m/(\log(m))^2\) by assumption.

### 7.5 Proof of Theorem 4.2

We let \(N = 2^{j_0 + r}\), where \(r = [\log_2(m^{2\alpha + 1}/(\log(m))^2)] - j_0\). With this choice, we have
\[
\frac{m^{2\alpha + 1}}{2\log^2(m)} \leq N \leq \frac{m^{2\alpha + 1}}{\log^2(m)}.
\]
We now apply Theorem 4.5 and use the fact that \(m/\log^2(m) \geq c_\alpha N(f)\).

### 8 Wavelet approximation from Fourier samples

The remainder of this paper is devoted to the proof of Theorems 3.4 and 4.6. This requires some significant additional work. In this section, we first formulate the approximation of wavelet coefficients from Fourier samples as a compressed sensing problem. Next, in §9 we present the framework for compressed sensing with local structure. Proofs of the two main theorems are presented in §10.

In order to align with notation used in previous works, we now make a minor change in notation. In particular, we will denote the dimension of the truncated wavelet space by \(M\) and the dimension of the truncated Fourier space by \(N\).
8.1 Formulation as a compressed sensing problem

We follow the approach of [5]. Let \( \{ \gamma_i \}_{i \in \mathbb{N}} \) be the Fourier basis \((A.1)\) and \( \{ \phi_j \}_{j \in \mathbb{N}} \) be the periodized Daubechies’ wavelet basis with \( p \) vanishing moments \((B.3)\). Define the infinite cross-Gramian matrix

\[
U = (\langle \phi_j, \gamma_i \rangle)_{i,j \in \mathbb{N}}.
\]  

(8.1)

Notice that \( U \) is a bounded, unitary operator on \( \ell^2(\mathbb{N}) \), since both sets of functions are orthonormal bases for \( L^2([0,1]) \). Recalling the notation introduced in \( \S 2.2 \), if \( f \in L^2([0,1]) \) is the function to recover, write \( d = (d_i)_{i \in \mathbb{N}} \) for its wavelet coefficients, so that \( f = \sum_{i \in \mathbb{N}} d_i \phi_i \).

Observe that \( b = Ud \),

(8.2)

where \( b = (\langle f, \gamma_i \rangle)_{i \in \mathbb{N}} \) is the infinite vector of Fourier coefficients of \( f \).

The Fourier encoder must use only \( m \) Fourier samples, or equivalently, select \( m \) rows of the infinite linear system \((8.2)\). Let \( P_\Omega \in \mathbb{C}^{m \times \infty} \) be the matrix that selects such rows, where \( \Omega \subset \mathbb{Z} \), \( |\Omega| = m \) is the set of frequencies. Then we consider the \( m \times \infty \) linear system

\[
P_\Omega Uz = P_\Omega b,
\]  

(8.3)

where \( z \in \ell^2(\mathbb{N}) \). This system is not suitable for computations, however, since the matrix \( P_\Omega U \) has infinitely-many columns. To handle this, we introduce an additional parameter \( M \geq 1 \) and replace \((8.3)\) by the \( m \times M \) linear system

\[
P_\Omega U P_M z = P_\Omega b,
\]  

(8.4)

where \( z \in \mathbb{C}^M \) and \( y = P_\Omega b \). Note that a solution \( z \in \mathbb{C}^M \) to this linear system is an approximation to the first \( M \) wavelet coefficients of \( f \), i.e. the vector \( P_M d \). Indeed,

\[
P_\Omega U P_M z = P_\Omega U P_M d + e, \quad \text{where } e = P_\Omega U P_M^\perp d,
\]

hence this problem is now a typical compressed sensing problem: the recovery of a vector \( P_M d \) from measurements taken according to a matrix \( P_\Omega U P_M \in \mathbb{C}^{m \times M} \) corrupted by noise \( e \).

8.2 Sampling and frequency bands

Having done this, we need to prescribe the set of samples \( \Omega \). For this, we follow the approach of [23] and divide frequency space \( \mathbb{Z} \) into dyadic bands \( B_k \). These are defined as follows:

\[
B_1 = \{-2^{j_0}, 1, \ldots, 2^{j_0}\},
\]

\[
B_{k+1} = \{-2^{j_0+k} + 1, \ldots, -2^{j_0+k-1}\} \cup \{2^{j_0+k-1} + 1, \ldots, 2^{j_0+k}\}, \quad k = 1, 2, \ldots.
\]

Here \( j_0 \) is as in \((B.1)\). Observe that \( \cup_{k \in \mathbb{N}} B_k = \mathbb{Z} \), and that

\[
|B_1| = 2^{j_0+1}, \quad |B_{k+1}| = 2^{j_0+k}, \quad k = 1, 2, \ldots.
\]

Since we typically consider the Fourier basis \((A.2)\) indexed over \( \mathbb{N} \) instead of \( \mathbb{Z} \) we now note that the \( B_k \) are equivalent to the partition of \( \mathbb{N} \) into subsets (subsequently referred to as levels)

\[
\{N_{k-1} + 1, \ldots, N_k\}, \quad k = 1, 2, \ldots,
\]
where

\[ N_0 = 0, \quad N_k = 2^{j_0 + k}, \quad k = 1, 2, \ldots \]  \quad (8.5)

With this in hand, we may now define the sampling scheme \( \Omega \). The idea is to select \( m_1 \) samples from the first level, \( m_2 \) samples from the second level, and so forth, up to some maximal level \( r \geq 1 \), where the numbers \( m_1, \ldots, m_r \) satisfy \( \sum_{k=1}^r m_k = m \). A judicious choice of these numbers is a key ingredient in the proof of Theorems 3.4 and 4.6. For \( k = 1, \ldots, r \), let \( \Omega_k \subseteq \{N_{k-1} + 1, \ldots, N_k\} \) be the subset of frequencies chosen in the \( k \)th level. Then we write

\[ \Omega = \Omega_1 \cup \ldots \cup \Omega_r. \]

In what follows, \( \Omega_k \) will be chosen randomly (see the next section).

## 9 Compressed sensing with local structure

As mentioned, a crucial part of the framework for wavelet approximation from Fourier samples is the notion of local structure in compressed sensing. The sampling scheme introduced previously is an instance of this principle, in the sense that different local numbers of samples can be chosen in different frequency bands. However, in order to perform wavelet approximation from Fourier samples we also need a notion of local sparsities.

In this section, we formalize these notions, and tie them together with a recovery guarantee which generalizes the classical compressed sensing result (Theorem 2.7). This follows the framework of [8]. Since it is unnecessary for the moment, in this section we do not assume the sampling/recovery is performed using the Fourier and wavelet bases, as in §8. We simply assume the existence of a unitary operator \( U : \ell^2(\mathbb{N}) \rightarrow \ell^2(\mathbb{N}) \), an infinite sequence \( d \in \ell^2(\mathbb{N}) \) to recover, and a collection of possible measurements \( b = Ud \) (recall (8.2)).

### 9.1 Definitions

We commence with a series of definitions:

**Definition 9.1.** Let \( r \geq \tilde{r} \geq 1 \), \( N = (N_1, \ldots, N_r) \), where \( 1 \leq N_1 < N_2 < \ldots < N_r < \infty \) and \( \mathbf{m} = (m_1, \ldots, m_r) \) where \( m_k = N_k - N_{k-1} \) for \( k = 1, \ldots, \tilde{r} \) and \( m_k = N_k - N_{k-1} \) for \( k = \tilde{r} + 1, \ldots, r \), with \( N_0 = 0 \). An \((N, \mathbf{m})\)-multilevel random subsampling pattern with saturation \( \tilde{r} \) is a subset \( \Omega \subset \mathbb{N} \) is of the form \( \Omega = \Omega_1 \cup \cdots \cup \Omega_r \), where

\[ \Omega_k = \{N_{k-1} + 1, \ldots, N_k\}, \quad k = 1, \ldots, \tilde{r}, \]

and, for each \( k = \tilde{r} + 1, \ldots, r \), \( \Omega_k = \{t_{k,1}, \ldots, t_{k,m_k}\} \) where the \( t_{k,i} \) are chosen independently and uniformly at random from the indices \( \{N_{k-1} + 1, \ldots, N_k\} \).

This formalizes the sampling strategy introduced in §8.2. Note that the \( N_k \) need not be given by (8.5) in general, although they will be whenever we consider the Fourier-wavelet problem. We refer to \( N = N_r \) as the sampling bandwidth, and the subsets \( \{N_{k-1} + 1, \ldots, N_k\} \) as levels. We also write \( m = m_1 + \ldots + m_r \) for the total number of measurements.

Given such a sampling pattern \( \Omega \), we now proceed similarly to §8.1, and write

\[ P_{\Omega} U P_M z = P_{\Omega} U P_M d + P_{\Omega} U P_M^\perp d. \]
For technical reasons, we also renormalize the rows of this linear system. Let
\[
D = \text{diag}(d_i)_{i=1}^\infty \in \mathbb{C}^{N \times \infty}, \quad d_i = \begin{cases} 
\sqrt{\frac{N_k - N_{k-1}}{m_k}} & N_{k-1} < i \leq N_k, \; k = 1, 2, \ldots, r \\
0 & i > N 
\end{cases}.
\]  \tag{9.1}

We then replace this system with
\[
Az = AP_Md + e,
\]  \tag{9.2}
where
\[
A = P_\Omega DU P_M \in \mathbb{C}^{m \times M}, \quad e = P_\Omega DU P_M^\perp d \in \mathbb{C}^m.
\]
The purpose of this normalization is to ensure that
\[
\mathbb{E}(A^*A) = P_MU^*P_NUP_M.
\]  \tag{9.3}

For suitable \( M \) and \( N \), the matrix on the right-hand side is well-conditioned, which is important for the proof. In order to ensure this, we define the following:

**Definition 9.2.** Let \( U : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N}) \) be unitary, \( 0 < \theta < 1 \) and \( N \geq M \geq 1 \). Then \( U \) has the balancing property with constant \( \theta \) if
\[
\|P_MU^*P_NUP_M - P_M\|_{\ell^2} \leq 1 - \theta.
\]  \tag{9.4}

We next define an appropriate local version of sparsity:

**Definition 9.3.** Let \( r \geq 1, M = (M_1, \ldots, M_r) \), where \( 1 \leq M_1 < M_2 < \ldots < M_r < \infty \) and \( s = (s_1, \ldots, s_r) \), where \( s_k \leq M_k - M_{k-1} \) for \( k = 1, \ldots, r \), with \( M_0 = 0 \). A vector \( x = (x_i)_{i=1}^M \in \mathbb{C}^M \) is \((s, M)\)-sparse if
\[
|\text{supp}(x) \cap \{M_{k-1} + 1, \ldots, M_k\}| \leq s_k, \; k = 1, \ldots, r.
\]

We write \( \Sigma_{s, M} \subseteq \mathbb{C}^M \) for the set of \((s, M)\)-sparse vectors.

Note that we take \( M_r = M \), where \( M \) is the truncation parameter introduced above. We refer to this as the sparsity bandwidth. We also write \( s = s_1 + \ldots + s_r \) for the total sparsity.

Although the sparsity levels can in general be arbitrary, when we consider the recovery of wavelet coefficients we will make the following specific choices:
\[
M_0 = 0, \quad M_k = 2^{j_0 + k}, \quad k = 1, 2, \ldots.
\]  \tag{9.5}

Since we assume the wavelets are ordered in the standard way (B.3), this means that the \( k \)th sparsity level (i.e. the indices \( \{M_{k-1} + 1, \ldots, M_k\} \)) corresponds exactly to the wavelet coefficients at scale \( j = j_0 + k - 1 \).

We require one additional concept. It is well known [22, 23] that recovery guarantees in compressed sensing from measurements taken according to a unitary matrix are determined by its so-called coherence, defined as follows:

**Definition 9.4.** The coherence of a matrix \( B = (b_{ij})_{i,j=1}^{N,M} \in \mathbb{C}^{N \times M} \) is
\[
\mu(B) = N \max_{i=1,\ldots,N} \max_{j=1,\ldots,M} |b_{ij}|^2.
\]
In our setting, we define so-called local coherences. Specifically, let $U$ be the infinite matrix defined in (8.1). Then the leading $N \times M$ section $P_N U P_M$ can be expressed in the block form

$$P_N U P_M = \begin{pmatrix} U^{(1,1)} & U^{(1,2)} & \ldots & U^{(1,r)} \\ U^{(2,1)} & U^{(2,2)} & \ldots & U^{(2,r)} \\ \vdots & \vdots & \ddots & \vdots \\ U^{(r,1)} & U^{(r,2)} & \ldots & U^{(r,r)} \end{pmatrix},$$

(9.6)

where, for $k, l = 1, \ldots, r$, the matrix $U^{(k,l)}$ is the block of $U$ defined by the $k$th sampling level and $l$th sparsity level:

$$U^{(k,l)} = P_{N_k}^{N_k-1} U P_{M_l}^{M_l-1} \in \mathbb{C}^{(N_k-N_{k-1}) \times (M_l-M_{l-1})}.$$  

We define $(k,l)$th local coherence of $U$ as the coherence of the corresponding block, i.e. $\mu \left( U^{(k,l)} \right)$.

### 9.2 The weighted-square root LASSO decoder

As discussed in §4.3 we will not employ QCBP (4.2) to solve (9.2). The reason for this is now clear. Standard recovery guarantees for (4.2) require the bound $\|e\|_2 \leq \eta$, where $e$ is the noise term. However, in (9.2) this term depends on the expansion tail $P_M d$, which is generally unknown. Instead, we consider the weighted square-root LASSO decoder

$$\min_{z \in \mathbb{C}^M} \lambda \|z\|_{\ell^1_w} + \|Az - y\|_2.$$  

(9.7)

This was introduced in [3] for compressed sensing in the context of high-dimensional function approximation. Here $y = AP_M d + e \in \mathbb{C}^m$, $w = (w_i)_{i=1}^M$ is a vector of positive weights, $\|z\|_{\ell^1_w} = \sum_{i=1}^M w_i |z_i|$ is the weighted $\ell^1$-norm, and $\lambda > 0$ is a parameter. As we show below, the optimal choice of $\lambda$ is independent of the noise level $\|e\|_2$, rendering this decoder suitable for our purposes.

These weights in (9.7) are used to control the off-diagonal blocks in $P_N U P_M$ – or, equivalently, the interferences between wavelet scales (see §1.4) – so as to give the best possible measurement condition. As it transpires, it is sufficient for the weights $w_i$ to be constant on the sparsity levels. Hence, from now on we assume that

$$w_i = w^{(k)}, \quad M_{k-1} < i \leq M_k, \quad k = 1, \ldots, r.$$  

(9.8)

We write $w = (w^{(1)}, \ldots, w^{(r)})$ for the vector of these weights.

### 9.3 A levels-based compressed sensing guarantee

We now present a recovery guarantee for (9.7). For this, we also define the $\ell^1_w$-norm best $(s, M)$-term approximation error:

$$\sigma_{s,M}(x)_{\ell^1_w} = \min \left\{ \|x - z\|_{\ell^1_w} : z \in \Sigma_{s,M} \right\}, \quad x \in \mathbb{C}^M,$$

Here $\Sigma_{s,M}$ is the set of $(s, M)$-sparse vectors.
Theorem 9.5. Let $0 < \varepsilon < 1$, $1 \leq \tilde{r} \leq r$, $U : \ell^2(N) \to \ell^2(N)$ be unitary, $N \geq M \geq 1$, $M = (M_1, \ldots, M_r)$ be sparsity levels, where $M_{\tilde{r}} = M$, and $s = (s_1, \ldots, s_r)$ be local sparsities with $s_l \geq 1$, $\forall l$, $N = (N_1, \ldots, N_r)$ be sampling levels, where $N_{\tilde{r}} = N$, $m = (m_1, \ldots, m_r)$ be local numbers of measurements, and $A = P_{\Omega} D U P_M \in \mathbb{C}^{m \times M}$ arise from the $(N, m)$-multilevel sampling scheme with saturation $\tilde{r}$ of Definition 9.1, where $D$ is as in (9.1). Let $x \in \ell^2(N)$, $y = Ax + e$ and suppose that

(A) $N$ and $M$ are such that the balancing property (9.4) holds with constant $\theta$,

(B) the weights $w$ are as in (9.8), with $w$ such that

$$c_1 \sqrt{s/s_k} \leq w^{(k)} \leq c_2 \sqrt{s/s_k},$$

for constants $c_1, c_2 > 0$,

(C) the vector $m$ of local numbers of measurements satisfies

$$m_k = N_k - N_{k-1}, \quad k = 1, \ldots, \tilde{r},$$

$$m_k \geq \theta^{-2} \cdot \left( \frac{c_2}{c_1} \right)^2 \cdot \sum_{l=1}^{r} s_l \mu(U^{(k,l)}) \cdot L, \quad k = \tilde{r} + 1, \ldots, r,$$

where $L = r^2 \cdot \log(m) \cdot \log^2(c_2^2 rs/(c_1^2 \theta)) \cdot \log(M) + r \cdot \log(\varepsilon^{-1})$,

(D) the parameter $\lambda$ satisfies

$$0 < \lambda \leq \frac{3\sqrt{\theta} - 1}{5\sqrt{2} c_2 \sqrt{rs}}.$$

Then, with probability at least $1 - \varepsilon$, any minimizer $\hat{x}$ of (9.7) satisfies

$$\| \hat{x} - x \|_2^2 \lesssim \frac{c_2^2}{c_1^2} \frac{1}{\varepsilon^{1/4} \sigma_s(x) \sigma_M(x) e_{\ell^2} + \frac{c_2^2}{c_1^2} \frac{1}{\varepsilon^{1/4} \sigma_s(x) \sigma_M(x) e_{\ell^2}}. \quad (9.10)$$

The proof of this theorem is given in Appendix C.

10 Proof of Theorems 3.4 and 4.6

10.1 The idea

Let $m$ and $\alpha$ be the inputs. The proof is based on making a judicious choice of the encoder parameters (number of levels $r$, the number of saturated levels $\tilde{r}$ and the local numbers of measurements $m = (m_1, \ldots, m_r)$) and the decoder parameters (weights $w = (w^{(1)}, \ldots, w^{(r)})$ and $\lambda$) in terms of $m$ and $\alpha$ so that there exists a vector of local sparsities $s = (s_1, \ldots, s_r)$ such that conditions (A)–(D) of Theorem 9.5 hold, and so that the right-hand side of (9.10) behaves like $m^{-\alpha} (\log(m))^t$, where $t$ is the exponent given in Theorem 3.4.

We first show that the choices made give that $\theta \asymp 1$, $c_1, c_2 \asymp 1$ and $r, \log(s), \log(M) \asymp \log(m)$. The next key is the measurement condition (C) of Theorem 9.5. By estimating the
local coherences and using the values of \( N_k \) in (8.5), we show that the \( m \) must satisfy

\[
m_1 = 2^{j_0 + 1} \]
\[
m_k = 2^{j_0 + k - 1}, \quad k = 2, \ldots, \tilde{r},
\]
\[
m_k \propto \left( \sum_{l=1}^{k} s_l 2^{-(2q+1)(k-l)} + \sum_{l=k+1}^{r} s_l 2^{-(2p+1)(l-k)} \right) \cdot (\log(m))^6, \quad k = \tilde{r} + 1, \ldots, r,
\]  

(10.1)

where \( p \geq 1 \) is the number of vanishing moments and \( q \geq 0 \) is the smoothness. Next, we make the following selection for the \( s \):

\[
s_1 = 2^{j_0 + 1} \]
\[
s_k = 2^{j_0 + k - 1}, \quad k = 2, \ldots, \tilde{r},
\]
\[
s_k = s_*, \quad k = \tilde{r} + 1, \ldots, r.
\]

(10.2)

In other words, \( s_k \) is equal to the size of the corresponding level for \( k \leq \tilde{r} \), and beyond this the \( s_k \)'s are constant. Note that we do not require \( \tilde{r} = \tilde{r} \). A judicious choice of \( \tilde{r} < \tilde{r} \) is crucial in the proof. As in previous proofs, \( s_* \) is chosen large enough to capture all the discontinuities of \( f \) at each scale. This gives

\[
\sigma_{s_* M}(x)_{\ell^1_{w_1}} \lesssim \|f\|_{PC^\alpha} \frac{2^{-(\alpha-1/2)\tilde{r}}}{r^{1/4} \sqrt{s_*}},
\]

(10.3)

(the other term in (9.10) is bounded in a similar fashion to previous results). We now relate this to the total number of measurements \( m \). Using the values (10.2) in (10.1), we see that

\[
m_k \propto \left( 2^{2(q+1)\tilde{r}-(2q+1)k} + s_* \right) \cdot (\log(m))^6, \quad k = \tilde{r} + 1, \ldots, r,
\]

and therefore

\[
m \propto 2^{\tilde{r}} + \left( 2^{2(q+1)\tilde{r}-(2q+1)\tilde{r}} + s_* \log(m) \right) \cdot (\log(m))^6.
\]

(10.4)

We now match terms, to get

\[
2^{\tilde{r}} \times m, \quad 2^{\tilde{r}} \times m(\log(m))^{-\frac{1}{q+1}}, \quad s_* \times m(\log(m))^{-7}.
\]

(10.5)

Substituting this into (10.3) now yields the desired \( m^{-\alpha}(\log(m))^\ell \) rate.

Let us make several remarks. First, the various bounds involving \( \lesssim \) and \( \gtrsim \) hide unknown constants depending on the wavelet. To avoid having such constants in the decoder, we introduce the parameter \( 0 < \delta < 1 \) and slightly increase the log factors in (10.5) by an amount depending on \( \delta \).

Second, note that it is critical that the number of full sampling levels \( \tilde{r} \) be allowed to exceed the number of saturated sparsity levels \( \tilde{r} \). If \( \tilde{r} = \tilde{r} \) then the \( \log(m) \) term in the error bound would not decrease with increasing \( q \), which is the key in Theorem 3.4. As discussed in §1.3, we are in effect heavily leveraging the ability of Fourier measurements to recover saturated wavelet scales efficiently, with a number of measurements equal to the size of the scale.
10.2 Recipe

We now provide the recipe. Henceforth, we assume that the $N_k$ and $M_k$ are given by (8.5) and (9.5) respectively. In particular, $N_k = M_k$, ∀k. The following applies to Theorem 3.4:

**Inputs:** number of measurements $m$, smoothness parameter $\alpha$, number of vanishing moments $p$, parameter $0 < \delta < 1$.

Note that $p$ only needs to be provided as an input for (3.3). For (3.5) it is chosen in the recipe so that (3.4) holds.

**Encoder parameters:** Let

- $q$ be the wavelet smoothness parameter, $j_0$ be as in (B.1);
- $r = \lceil \max\{2\alpha + 1, \frac{\alpha}{\alpha - 1/2}\} \log_2(m) \rceil - j_0$;
- $\tilde{r} = \lceil \log_2(m/2) \rceil - j_0$;
- $m_k = N_k - N_{k-1}$, $k = 1, \ldots, \tilde{r}$, and

$$m_k = \left\lfloor \frac{1}{4} \left( m - 2^{q+2} - (2q+1)(k+1) + \frac{m}{4(r - \tilde{r})} \right) \right\rfloor, \quad k = \tilde{r} + 1, \ldots, r.$$

Observe that $1 \leq \tilde{r} \leq \tilde{r} \leq r$ for all $m \geq c_{\alpha,p}$, and that

$$\sum_{k=1}^{r} m_k \leq 2\tilde{r} + \frac{1}{4} m^{2q+2} - (2q+1)(\tilde{r}+1) + \frac{m}{4} \leq m.$$

Hence these are valid parameters for the encoder.

**Decoder parameters:** Let

- $\{\phi_{j_0,n}\} \cup \{\psi_{j_0,n}\}$, the periodized Daubechies’ wavelet basis with $p$ vanishing moments;
- $L = (\log(m))^{6+\delta}$;
- $\tilde{r} = \lceil \log_2(m/L) \rceil - j_0$;
- $w = (w_i)_{i=1}^M$, where

$$w_i = \frac{m}{2^k L^{(q+1)/2}}, \quad M_{k-1} < i \leq M_k, \quad k = 1, \ldots, \tilde{r},$$

$$w_i = \frac{L^{2q+2}}{2^{q+1} r}, \quad M_{k-1} < i \leq M_k, \quad k = \tilde{r} + 1, \ldots, r,$$

and

$$\lambda = 1/\sqrt{rm};$$

- $\Omega$ be the corresponding $(N, m)$-multilevel random subsampling pattern with saturation $\tilde{r}$ (see Definition 9.1);
- $A \in \mathbb{C}^{m \times M}$ given by $A = P_{\Omega}DP_{M}$, where $U$ is the Fourier-wavelets matrix (8.1) and $D$ is as in (9.1).
Encoder: Let
\[ E_m(f) = \left( \hat{f}(\omega) \right)_{\omega \in \Omega} \in \mathbb{C}^{m_1 + \cdots + m_r}. \]

Decoder: Given measurements \( y = E_m(f) \), let \( \tilde{f}_m = \sum_{i=1}^{M} \tilde{d}_i \phi_i \), where \( \{ \phi_i \}_{i=1}^{\infty} \) is the wavelet basis and \( \tilde{d} = (d_i)_{i=1}^{M} \) is any minimizer of
\[ \min_{z \in \mathbb{C}^M} \lambda \| z \|_{\ell_1} + \| A z - y \|_{\ell_2}. \]

The case of Theorem 4.6: As in the previous section, we now assume \( N \) is an input rather than a parameter, and change the definition of \( r \) to \( r = \lfloor \log_2(N) \rfloor - j_0 \). We now set \( M = 2^{j_0 + r} \) and note that
\[ N/2 \leq M \leq N, \tag{10.6} \]
so that the encoder-decoder pair includes all the wavelets up to the scale \( j_0 + r - 1 \). The encoder and decoder are then defined in exactly the same way.

As before, we will prove Theorem 4.6 first, and then obtain Theorem 3.4 as a corollary.

10.3 Lower bounds on \( m \)
We first need several lower bounds on \( m \). Note that we assume the condition \( m \geq c_{p,\alpha,\delta} N(f)^2 \). In particular, this implies that
\[ m \geq c_{p,\alpha,\delta}, \]
and also that
\[ m \geq c_{p,\alpha,\delta} N(f) \left( \log(m) \right)^{7+\delta}. \]
The first inequality is immediate. The second follows after noting that \( \left( \log(m) \right)^{7+\delta} \leq \sqrt{m} \) for \( m \geq c \). We will use these inequalities repeatedly in what follows.

10.4 Technical lemmas
We first require the following three lemmas. Proofs are given in Appendix D.

Lemma 10.1. If \( N = M = N_r \), then \( U \) has the balancing property with constant \( \theta \) satisfying
\[ \theta \geq \inf_{|\omega| \leq \pi} |\hat{\phi}(\omega)|^2 = c_p > 0. \]
In particular, the balancing property holds with \( \theta = c_p \) depending only on the number of vanishing moments \( p \).

Lemma 10.2. The \( (k,l) \)th local coherence of \( U \) satisfies
\[ \mu(U^{(k,l)}) \leq c_p \left\{ \begin{array}{ll} 2^{-2q+1}(k-l) & k \geq l \\ 2^{-2p+1}(l-k) & k < l \end{array} \right. . \]

Lemma 10.3. Let \( d \in \ell^2(N) \), \( e = P_{\Omega} D U P_M^+ d \), where \( D \) is as in (9.1) and \( M = N \). Then
\[ \| e \|_{\ell^2} \leq C_p \| P_M^+ d \|_{\ell^1}. \]
The remainder of the proof is based on Theorem 9.5. We first verify that conditions (A)–(D) hold for the various choices of parameters. Then we estimate the recovery error in (9.10).
10.5 Step 1. Verification of conditions (A)–(D)

Lemma 10.4. Define local sparsities $s = (s_1, \ldots, s_r)$ by

$$s_k = M_k - M_{k-1}, \quad k = 1, \ldots, \bar{r},$$

and

$$s_k = s_* = \left\lfloor \frac{m}{L r} \right\rfloor, \quad k = \bar{r} + 1, \ldots, r.$$  

Then $1 \leq s_k \leq M_k - M_{k-1}$ for all $m \geq c_{p,\alpha,\delta}(\log(m))^{7+\delta}$, and conditions (A)–(D) of Theorem 9.5 hold for this choice of $s$ with constants $c_1 = c_{1,p}$ and $c_2 = c_{2,p}$ depending on $p$ only.

Proof. First, recalling the definition of $\bar{r}$, observe that $s_k \leq M_k - M_{k-1}$ provided

$$\bar{r} \leq \frac{m}{L r} \leq m_{\bar{r}+1} - M_{\bar{r}} = \frac{m}{L^{1/(q+1)}},$$

or equivalently $\bar{r} \geq 1$. This holds whenever $\sqrt{\bar{r}} \geq 1$, which, since $r \geq c_{p,\alpha}$, is implied by $m \geq c_{p,\alpha}$. Conversely, $s_* \geq 1$ provided $m \geq \bar{r} L r$, which, since $r \leq c_{p,\alpha}$ is implied by the condition $m \geq c_{p,\alpha}$.

Lemma 10.1 implies that (A) of Theorem 9.5 holds with $\theta \geq c_p$. Now consider (B). First, observe that

$$s = 2^{j_0 + \bar{r}} + (r - \bar{r})s_* \leq m \bar{r}^{-1/(q+1)} + m L^{-1} \leq 2m \bar{r}^{-1/(q+1)},$$

and conversely $s \geq \frac{1}{2} m \bar{r}^{-1/(q+1)}$, which gives

$$\frac{1}{2} m \bar{r}^{-1/(q+1)} \leq s \leq 2m \bar{r}^{-1/(q+1)}. \tag{10.7}$$

We also have

$$\frac{1}{2} m \bar{r}^{-1} r^{-1} \leq s_* \leq m \bar{r}^{-1} r^{-1},$$

for all $m \geq \bar{r} L$ (that holds thanks to the assumption on $m$), and therefore

$$\frac{1}{2} L^{2q+1} r \leq \frac{s}{s_k} \leq 4 L^{2q+1} r, \quad k = \bar{r} + 1, \ldots, r.$$  

Hence $\frac{1}{\sqrt{2}} w^{(k)} \leq \sqrt{s/s_k} \leq 2w^{(k)}$, which gives (B) for $k = \bar{r} + 1, \ldots, r$. For $k = 1, \ldots, \bar{r}$, we note that

$$2^k \leq s_k \leq c_p 2^k,$$

for some $c_p$, and therefore

$$c_p \frac{m}{2^k L^{1/(q+1)}} \leq \frac{s}{s_k} \leq \frac{m}{2^k L^{1/(q+1)}}, \quad k = 1, \ldots, \bar{r}.$$  

Hence $c_{1,p} w^{(k)} \leq \sqrt{s/s_k} \leq c_{2,p} w^{(k)}$ in this case as well.

In a similar manner we also note that (D) follows immediately from the definition of $\lambda$ and the fact that $m \geq c_p s$ for all $m \geq c_p$.
Finally, consider condition (C). By definition, the first $\tilde{r}$ levels are fully saturated. Note that $\tilde{r} \geq \bar{r}$ for all $m \geq c_p$. Now consider the unsaturated levels $\tilde{r} < k \leq r$. Then, by definition of the $s_k$ and Lemma 10.2, we have
\[
\sum_{l=1}^{r} s_l \mu \left( U^{(k,l)} \right) \leq c_p \left( \sum_{l=1}^{\bar{r}} 2^{l-(2q+1)(k-\bar{r})} + \sum_{l=\bar{r}+1}^{k} 2^{-(2q+1)(k-l)} s_s + \sum_{l=k+1}^{r} 2^{-(2p+1)(l-k)} s_s \right) 
\leq c_p \left( 2^{(2q+2)\bar{r}} 2^{-(2q+1)k} + s_s \right) 
= c_p \left( 2^{(2q+1)(\bar{r}-k)} 2^{\bar{r}} + s_s \right).
\]

Now consider the log factor $L$ in Theorem 9.5. Since $r \leq s \leq m$, $r \leq c_p \log(m)$, and $\log(s) \leq \log(m)$ we have
\[
L = r^2 \cdot \log(m) \cdot \log^2(c_p r \sigma/\theta) \cdot \log(M) + r \cdot \log(\epsilon^{-1}) 
\leq c_p \left( \log^6(m) + \log(m) \cdot \log(\epsilon^{-1}) \right) 
\leq c_p(\log(m))^6,
\]
where we have also used that $\log(c_p r^2/\theta) \leq c_p \log(s) \leq c_p \log(m)$, thanks to (A), and the condition $(\log(m))^5 \geq \log(\epsilon^{-1})$. Hence, writing $c' > 0$ for the universal constant understood in condition (C), we deduce that
\[
c' \left( \frac{c_{2,p}}{c_{1,p}} \right)^2 \cdot \theta^{-2} \cdot \left( \sum_{l=1}^{r} s_l \mu \left( U^{(k,l)} \right) \right) \cdot L 
\leq c_p \left( 2^{(2q+1)(\bar{r}-k)} 2^{\bar{r}} + s_s \right) \left( \log^6(m) + \log(m) \log(\epsilon^{-1}) \right) 
\leq c_p \left( m^{2q+2} 2^{-(2q+1)k} + m/r \right) \frac{\log^6(m) + \log(m) \log(\epsilon^{-1})}{L} 
\leq m_k,
\]
which is ensured for $m \geq c_p \delta$. Hence (C) holds.

\begin{flushright} \Box \end{flushright}

10.6 Step 2. Estimation of the approximation error

Having verified conditions (A)–(D) of Theorem 9.5, and using Lemma 10.3, we now get
\[
\| f - \bar{f} \|_{L^2} \leq \| P_M^\bot d \|_{\ell^2} + \| P_M d - \bar{d} \|_{\ell^2} 
\leq \| P_M^\bot d \|_{\ell^2} + C_p \frac{\sigma_{s,M}(P_M d)_{\ell^2}}{r^{1/4} \sqrt{s}} + C_p \frac{1}{r^{1/4} \sqrt{s}} \| P_M^\bot d \|_{\ell^1} 
= E_1 + E_2 + E_3. \tag{10.8}
\]

Note that $\| P_M^\bot d \|_{\ell^2} = \epsilon_M(f)_{L^2}$, where $\epsilon_M(f)_{L^2}$ is as in (2.2). Hence Theorem 2.5 gives
\[
E_1 \leq C_{p,\alpha} \sqrt{N(f)} \| f \|_{P_C^\alpha}/\sqrt{M}. \tag{10.9}
\]

Arguing similarly as in the proof of Theorem 2.5, we also see that
\[
\| P_M^\bot d \|_{\ell^1} \leq C_{p,\alpha} \| f \|_{P_C^\alpha} \left( M^{-(\alpha-1)/2} + N(f)/\sqrt{M} \right), \tag{10.10}
\]

35
where we use the fact that $\alpha > 1/2$. Hence, using the values of $\lambda$, $s$ and $r$, and the bound (10.7), we deduce that the final term of (10.8) satisfies

$$E_3 \leq C_p \frac{\sqrt{r^M \lambda}}{\sqrt{L^{\frac{1}{4(q+1)}}}} \| P_M e \|_{\ell^1} \leq C_{p,\alpha} (\log(m))^{\frac{1}{4} + \frac{\delta + \delta}{4(q+1)}} \| f \|_{PC^\alpha} \left( \frac{1}{M^{\alpha-1/2}} + \frac{N(f)^{1/2}}{\sqrt{M}} \right).$$

(10.11)

Next, consider $\sigma_{s,M}(P_M x)_{\ell^1_w}$. We first recall that $s_k = M_k - M_{k-1}$ for $k = 1, \ldots, r$. Moreover, $s_k \geq (2p)^2 N(f)$ for all $k > \bar{r}$ (recall that $(2p)^2 N(f)$ is the maximum number of the wavelets at any fixed scale $k$ whose support contains a discontinuity of $f^{\text{ext}}$; see the proof of Theorem 2.5) by construction and thanks to the assumption $m/(\log(m))^{q+\delta} \geq c_p N(f)$, which implies that $m \geq c_p N(f)Lr$. Therefore, the coefficients corresponding to the discontinuities of $f^{\text{ext}}$ are excluded from the approximation error. Applying Lemma 5.1 and the definition of the weights $w^{(k)}$, we deduce that

$$\frac{\sigma_{s,M}(P_M d)_{\ell^1_w}}{r^{1/4} \sqrt{s}} \leq C_{p,\alpha} \| f \|_{PC^\alpha} \sum_{j=r+1}^r \frac{1}{r^{1/4} \sqrt{s}} 2^{-(\alpha-1/2)j}.$$  

(10.12)

Recalling that $\alpha > 1/2$, we have

$$\sum_{j=r+1}^r \frac{1}{r^{1/4} \sqrt{s}} 2^{-(\alpha-1/2)j} \leq \frac{2^{-(\alpha-1/2)(r+1)}}{r^{1/4} \sqrt{s}}.$$

Using the definitions of $\bar{r}$, $r$ and $s$, we have

$$\frac{2^{-(\alpha-1/2)(r+1)}}{r^{1/4} \sqrt{s}} \leq C_{p,\alpha} m^{-\alpha(1/2)} L^{\frac{\alpha-1/2}{2(q+1)}} (\log(m))^{-1/4} \sqrt{m}/\sqrt{L}$$

$$= C_{p,\alpha} m^{-\alpha(1/2)} L^{\frac{\alpha-1/2}{2(q+1)}} \leq C_{p,\alpha} m^{-\alpha(1/2)} L^{\frac{\alpha-1/2}{2(q+1)}} \leq C_{p,\alpha} m^{-\alpha(1/2)} L^{\frac{\alpha-1/2}{2(q+1)}}.$$

Therefore, combining this with the previous estimate, we deduce that

$$E_2 \leq C_{p,\alpha} \| f \|_{PC^\alpha} m^{-\alpha(1/2)} L^{\frac{\alpha-1/2}{2(q+1)}}.$$  

(10.13)

### 10.7 Step 3. Concluding the proof

Substituting (10.13), (10.9) and (10.11) into (10.8), we finally deduce that

$$\| f - \hat{f} \|_2 \leq C_{p,\alpha} \| f \|_{PC^\alpha} \left[ \left( m^{-\alpha} + M^{-\alpha(1/2)} + N(f)/\sqrt{M} \right) (\log(m))^{1/4 + \frac{(6+\delta)(q+\alpha+1/2)}{2(q+1)}} \right].$$

To obtain (4.4), we merely use (10.6) and the fact that $N(f) \geq 1$. For (4.5), we note that if

$$q \geq \frac{6\alpha - 1/2}{\delta} + \alpha - \frac{3}{2}$$

Then

$$\frac{1}{4} + \frac{(6+\delta)(q+\alpha+1/2)}{2(q+1)} \leq \frac{13}{4} + \delta.$$
10.8 Proof of Theorem 3.4

We let $M = 2^{j_0 + r}$, where $r = \lfloor \max\{2\alpha + 1, \frac{\alpha}{\alpha - 1/2}\} \log_2(m) \rfloor - j_0$. This gives

$$M \geq \frac{1}{2} \max\left\{m^{2\alpha + 1}, m^{\frac{\alpha}{\alpha - 1/2}}\right\}.$$ 

This, and the condition $m \geq N(f)^2$, give

$$\| f - \tilde{f} \|_{L^2} \leq C_{p,\alpha}(\log(m))^{\frac{1}{2} + \frac{(\alpha + 1/2)(\alpha + 1/2)}{2(\alpha + 1)}} m^{-\alpha},$$

as required. For (3.5) we argue as above.

11 Conclusions and challenges

We conclude this paper by listing a number of open problems.

1. Reducing the log factor. The log factor in Theorem 3.4 has the potential to be decreased. There are several ways to do this. First, reducing the log factor $L$ in Theorem 9.5 (or, more specifically, in Theorem C.2). This is related to, although more general than, the question in compressed sensing of when a subsampled Fourier matrix has the RIP of order $s$. Theorem C.2 (specialized to $r = 1$ level) implies this whenever $m \gtrsim s \cdot \log(m) \cdot \log^2(s) \cdot \log(N)$. However, it is known [27] that this can be achieved under the weaker condition

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

(11.1)

If such arguments could be generalized to sparsity in levels, we could save one $\log(m)$ factor in Theorem C.2. This would decrease the $\log(m)$ exponent in Theorem 3.4 by 1/2. Another way to reduce this exponent would be to remove the dependence on $r$ in Theorem 9.5 (recall that $r \approx \log(m)$ in the setting of Theorem 3.4). Further improvements, however, would seemingly necessitate improving (11.1), which is a challenging open problem in compressed sensing theory.

2. Standard decoders. Since they are most commonly used in practice, it is desirable to have guarantees for the standard (unweighted) LASSO and QCBP decoders. Unfortunately, it is not clear how to avoid using a weighted $\ell^1$-norm without the recovery guarantee being ruined by the interferences (see §4.3). For some initial work in this direction, see [4].

3. Other function classes. We have chosen to study the class $PC^\alpha$ to avoid additional technical challenges. An interesting problem is to extend this work to Besov spaces. Another open problem is the extension to higher dimensions. While a direct extension using wavelets may be not be too challenging for suitable analogues of the class $PC^\alpha$, it is well known that wavelet are generally not optimal in higher dimensions. More interesting challenges involve extending this work to, for instance, shearlets. We note that sparse recovery guarantees of a similar flavour to (1.6) for Fourier sampling with shearlets have been shown in [47]. However, to extend our results, we require local recovery guarantees. See [55] for some work in this direction.

A further direction is to adapt this work to Total Variation (TV) minimization, or its various higher-order generalizations. For existing recovery guarantees for Fourier sampling with TV, see [54].
4. Binary measurements. While many imaging modalities employing Fourier sampling, others (in particular, optical imaging) are constrained to acquire binary measurements. In practice, one can design structure-exploiting binary measurements by replacing the Fourier transform with the Walsh (also known as Hadamard) transform [9, 2]. As with Fourier sampling, this performs significantly better than random Bernoulli sampling (the binary analogue of Gaussian sampling) [59]. Unfortunately, for Walsh sampling with wavelets, the corresponding local coherences \( \mu \left( U^{(k,l)} \right) \leq c_p 2^{-|k-l|} \) decay at a rate independent of the wavelet order [2], unlike in Fourier sampling (Lemma 10.2). In our proof of Theorem 3.4 fast decay of the coherences is critical. Hence it remains an open problem to extend our analysis to binary sampling.

5. Optimal algorithms. Finally, we remark that our decoders \( D_m \) are not algorithms per se, since they involve the exact solution of certain convex optimization problems. An open problem is to design an algorithm that takes inputs \( m, \alpha \) and the measurements \( E_m(f) \), and then computes the approximation \( \tilde{f}_m \) in polynomial time in \( m \).

Acknowledgements

The authors extend their thanks to Vegard Antun (University of Oslo), who performed the experiment in Fig. 1. They also would like to thank Anders C. Hansen, Bradley J. Lucier and Clarice Poon. S.B. acknowledges the support of the PIMS Postdoctoral Training Centre in Stochastics. This work was supported by the PIMS CRG in “High-dimensional Data Analysis” and by NSERC through grant R611675.

A Fourier transform and series

Given \( f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \) we define the Fourier transform as

\[
\hat{f}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t} \, dt.
\]

If \( f \in L^2([0,1]) \) then we can write \( f \) as its Fourier series

\[
f = \sum_{n \in \mathbb{Z}} \langle f, \gamma_n \rangle_{L^2} \gamma_n,
\]

where

\[
\gamma_n(t) = e^{2\pi i nt}, \quad n \in \mathbb{Z}, \tag{A.1}
\]

is the Fourier basis for \( L^2([0,1]) \). If we consider \( f \) as a function in \( L^2(\mathbb{R}) \) that is zero outside \([0,1]\), then \( \langle f, \gamma_n \rangle_{L^2} = \hat{f}(2\pi n) \). For convenience, we also re-index this basis over \( \mathbb{N} \) as follows:

\[
\gamma_{2n-1} = e^{-2\pi i (n-1)t}, \quad \gamma_{2n} = e^{2\pi i nt}, \quad n \in \mathbb{N}. \tag{A.2}
\]

B Orthogonal wavelet bases of \( L^2([0,1]) \)

Let \( \varphi \) and \( \psi \) be the scaling function and mother wavelet, respectively, of the Daubechies’ wavelet with \( p \geq 1 \) vanishing moments. Write

\[
\varphi_{j,n}(x) = 2^{j/2}\varphi(2^j x - k), \quad \psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad j, k \in \mathbb{Z}.
\]
Since we work with functions on the interval \([0, 1]\), we need an orthonormal wavelet basis of \(L^2([0, 1])\). We construct this via periodization (see (5.1) and [51, Sec. 7.5.1] for more details). Define the coarsest scale
\[
 j_0 = \begin{cases} 
 0 & p = 1 \\
 \lceil \log_2(2p) \rceil & p \geq 2
\end{cases},
\]  
(B.1)
(in general, one could allow any fixed \(j_0\) greater than or equal to the right-hand side. However, this does not affect any of the results in the paper, hence we simply specify \(j_0\) exactly). Then the set of functions
\[
\{\phi_{j_0,k}^{\text{per}} : k = 0, \ldots, 2^{j_0} - 1\} \cup \{\psi_{j,k}^{\text{per}} : k = 0, \ldots, 2^j - 1, j \geq j_0\},
\]  
(B.2)
is an orthonormal basis of \(L^2([0, 1])\), referred to as the periodized Daubechies wavelet basis. We note in passing that \(\psi_{j,k}^{\text{per}} = \psi_{j,k}, \phi_{j,k}^{\text{per}} = \phi_{j,k}, k = p - 1, \ldots, 2^j - p\), that is, wavelets that are fully supported in \([0, 1]\) are unchanged, and
\[
\begin{align*}
\phi_{j,k}^{\text{per}} &= \phi_{j,k} + \phi_{j,2^j + k}, \\
\psi_{j,k}^{\text{per}} &= \psi_{j,k} + \psi_{j,2^j + k}, & k &= 0, \ldots, p - 2, \\
\phi_{j,k}^{\text{per}} &= \phi_{j,k} + \phi_{j,2^j - p - k}, \\
\psi_{j,k}^{\text{per}} &= \psi_{j,k} + \psi_{j,2^j - p - k}, & k &= 2^j - p + 1, \ldots, 2^j - 1,
\end{align*}
\]
where the functions in the right-hand sides are implicitly restricted to \([0, 1]\). As needed, we order the basis (B.2) in the usual way, rewriting it as \(\{\phi_n\}_{n \in \mathbb{N}}\), where
\[
\begin{align*}
\phi_{n+1} &= \phi_{j_0,n}^{\text{per}}, & n &= 0, \ldots, 2^{j_0} - 1, \\
\phi_{2j+n+1} &= \phi_{j,n}^{\text{per}}, & n &= 0, \ldots, 2^j - 1, j \geq j_0.
\end{align*}
\]  
(B.3)

C  Proof of Theorem 9.5

The technical tools we need to prove this theorem were introduced in [2], where a similar result was proven for the weighted quadratically-constrained basis pursuit decoder.

We require several concepts from [2]. First, we introduce several additional pieces of notation. Given sparsity levels \(M = (M_1, \ldots, M_r)\) and local sparsities \(s = (s_1, \ldots, s_r)\), let
\[
D_{s,M} = \{\Delta \subseteq \{1, \ldots, N\} : |\Delta \cap \{M_{k-1} + 1, \ldots, M_k\}| \leq s_k\},
\]
be the set of all possible supports of an \((s,M)\)-sparse vector. Given positive weights \(w = (w_i)_{i=1}^M \in \mathbb{C}^M\) and a set \(\Delta \subseteq \{1, \ldots, M\}\), we define its weighted cardinality as follows:
\[
|\Delta|_w = \sum_{i \in \Delta} (w_i)^2.
\]
The conventional tool in compressed sensing for establishing recovery guarantees is the so-called Restricted Isometry Property (RIP). In our case, we require an generalized version of the RIP. This takes into account the sparsity in levels structure, and the fact that the measurement matrix \(A\) satisfies (9.3), rather than the more standard condition \(\mathbb{E}(A^*A) = I\).
Definition C.1 (G-adjusted RIP in Levels). Let $M = (M_1, \ldots, M_r)$ be sparsity levels, $s = (s_1, \ldots, s_r)$ be local sparsities and $G \in \mathbb{C}^{M \times M}$ be invertible, where $M = M_r$ is the sparsity bandwidth. The $(s, M)$-th $G$-adjusted Restricted Isometry Constant in Levels (G-RIPL) $\delta_{s,M,G}$ of a matrix $A \in \mathbb{C}^{m \times M}$ is the smallest $\delta \geq 0$ such that

$$(1 - \delta)\|Gx\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta)\|Gx\|_2^2, \quad \forall x \in \Sigma_{s,M}.$$  

If $0 < \delta_{s,M,G} < 1$ then the matrix is said to have the $G$-adjusted Restricted Isometry Property in levels (G-RIPL) of order $(s, M)$.

In our setting, if $N, M$ are such that $P_N U P_M$ is full rank (in particular, if the balancing property holds), then $G$ will be taken as the unique positive definite square-root of the positive definite matrix $P_M U^* P_N U P_M$. We write $G = \sqrt{P_M U^* P_N U P_M}$ in this case.

The following result [2, Thm. 3.6] gives conditions under which the matrix $A$ satisfies the G-RIPL:

**Theorem C.2.** Let $0 < \delta, \varepsilon < 1$, $M \geq 2$, $1 \leq \tilde{r} \leq r \leq N$ and $M = (M_1, \ldots, M_r)$ and $s = (s_1, \ldots, s_r)$ be sparsity levels and local sparsities respectively, where $s = s_1 + \ldots + s_r \geq 2$ and $M_r = M$. Let $\Omega$ be an $(N, m)$-multilevel random subsampling pattern with $r$ levels and saturation $\tilde{r}$, and $N = N_r$. Suppose that $N, M$ are such that $P_N U P_M$ is full rank, where $U$ is as in (8.1) and consider the matrix $A$ given by (8.4). If

$$m_k \gtrsim \delta^{-2} \cdot \|G^{-1}\|_{\ell^2}^2 \cdot \left(\sum_{k=1}^r s_k \mu(U^{(k,l)})\right) \cdot L, \quad k = \tilde{r} + 1, \ldots, r,$$

where $L = r \cdot \log(m) \cdot \log^2(s) \cdot \log(N) + \log(\varepsilon^{-1})$, then, with probability at least $1 - \varepsilon$, $A$ satisfies the G-RIPL of order $(s, M)$ with constant $\delta_{s,M,G} \leq \delta$ and $G$ given by $G = \sqrt{P_M U^* P_N U P_M}$.

In order to establish Theorem 9.5, we next show that the G-RIPL implies stable and robust recovery. To do so, we first introduce the following generalization of the so-called robust Null Space Property (rNSP):

Definition C.3. Let $M = (M_1, \ldots, M_r)$ be sparsity levels, $s = (s_1, \ldots, s_r)$ be local sparsities and $w \in \mathbb{C}^M$ be positive weights, where $M = M_r$. A matrix $A \in \mathbb{C}^{m \times M}$ has the weighted robust null space property in levels (weighted rNSPL) of order $(s, M)$ with constants $0 < \rho < 1$ and $\gamma > 0$ if

$$\|P_\Delta x\|_{\ell^2} \leq \frac{\rho \|P_\Delta x\|_{\ell^2}}{\sqrt{|\Delta|}_w} + \gamma \|Ax\|_{\ell^2},$$

for all $x \in \mathbb{C}^M$ and $\Delta \in D_{s,M}$.

Suppose the weights $w = (w_i)_{i=1}^M$ are of the form (9.8), i.e. constant on the sparsity levels, and define

$$\xi = \xi(s, w) = \sum_{k=1}^r (w^{(k)})^2 s_k, \quad \zeta = \zeta(s, w) = \min_{k=1, \ldots, r} \left\{ (w^{(k)})^2 s_k \right\}.$$  

The following combines Lemmas 5.2 and 5.3 of [2]:

40
Lemma C.4. Suppose that $A$ has the weighted rNSPL of order $(s, M)$ with constants $0 < \rho < 1$ and $\gamma > 0$. Let $x, z \in \mathbb{C}^M$. Then
\[
\|z - x\|_{\ell_2} \leq \frac{1 + \rho}{1 - \rho} \left( 2\|s \|_{\ell_2} (x) \|_{\ell_2} + \|z\|_{\ell_2} - \|x\|_{\ell_2} \right) + \frac{2\gamma}{1 - \rho} \sqrt{\xi} \|A(z - x)\|_{\ell_2},
\]
and
\[
\|z - x\|_{\ell_2} \leq \left( \rho + (1 + \rho)(\xi/\zeta)^{1/4}/2 \right) \frac{\|z - x\|_{\ell_2}}{\sqrt{\xi}} + \left( 1 + (\xi/\zeta)^{1/4}/2 \right) \gamma \|A(z - x)\|_{\ell_2}.
\]

The G-RIPL implies the weighted rNSPL (see [2, Thm. 5.5]):

Theorem C.5. Let $A \in \mathbb{C}^{m \times M}$ and $G \in \mathbb{C}^{M \times M}$ be invertible. Let $M = (M_1, \ldots, M_r)$ and $\mathbf{s} = (s_1, \ldots, s_r)$ be sparsity levels and local sparsities respectively, and $\mathbf{w}$ be positive weights of the form (9.8). Let $0 < \rho < 1$, and suppose that $A$ has the G-RIPL of order $(\mathbf{t}, M)$ and constant $1/2$, where $\mathbf{t} = (t_1, \ldots, t_r)$ satisfies
\[
t_l = \min \left\{ 2 \left[ 3 \frac{\kappa(G)^2}{\rho^2} \frac{\xi(s, w)}{w(l)} \right], M_l - M_{l-1} \right\}, \quad l = 1, \ldots, r,
\]
and $\kappa(G) = \|G\|_{\ell_2}\|G^{-1}\|_{\ell_2}$ is the condition number of $G$ with respect to the $\ell_2$-norm. Then, there exists $0 < \gamma \leq \sqrt{2}\|G^{-1}\|_{\ell_2}$ such that $A$ has the weighted rNSPL of order $(\mathbf{s}, M)$ with constants $\rho$ and $\gamma$.

Finally, we are now ready to prove Theorem 9.5:

Proof of Theorem 9.5. Recall that $G^2 = P_MU^*P_NUP_M$. Hence $G$ is invertible since $U$ has the balancing property (9.4), and moreover, we have
\[
\|G^{-1}\|_{\ell_2} \leq 1/\sqrt{\theta}.
\]
We also have $\|G\|_{\ell_2} \leq 1$ since $U$ is unitary, and therefore $\kappa(G) \leq 1/\sqrt{\theta}$.

Let $t_l$ be given by (C.2) with $\rho = 1/2$. Recalling (9.9) and (C.1), observe that
\[
t_l \leq 48 \frac{c_2^2 r s_l}{c_1^2 \theta},
\]
Therefore
\[
t = t_1 + \ldots + t_r \leq 48 \frac{c_2^2 r}{c_1^2 \theta} s,
\]
and
\[
\|G^{-1}\|_{\ell_2}^2 \cdot \sum_{k=1}^r t_k \mu(U^{(k,l)}) \cdot (r \cdot \log(m) \cdot \log^2(t) \cdot \log(M) + \log(\varepsilon^{-1})) \leq \theta^{-2} \frac{c_2^2 r}{c_1} \cdot \sum_{k=1}^r s_k \mu(U^{(k,l)}) \cdot (r \cdot \log(m) \cdot \log^2(c_2^2 r s/c_1^2 \theta) \cdot \log(M) + \log(\varepsilon^{-1})).
\]
Hence, condition (C) and Theorem C.2 imply that the matrix $A$ has the G-RIPL of order $(\mathbf{t}, M)$ with constant $\delta_{t, M, G} \leq 1/2$. It now follows from Theorem C.5 that $A$ has the weighted rNSPL of order $(\mathbf{s}, M)$ with constants $\rho = 1/2$ and $\gamma \leq \sqrt{2}\|G^{-1}\|_{\ell_2} \leq \sqrt{2/\theta}$.
To complete the proof we use Lemma C.4 with $z = \hat{x}$. Using this, (C.3) and the bounds

\[
c_1^2rs \leq \xi \leq c_2^2rs, \quad c_1^2s \leq \zeta \leq c_2^2s. \tag{C.4}
\]

we see that

\[
\|\hat{x} - x\|_{\ell^2} \leq \left(1 + (c_2^2r/c_1^2)^{1/4}\right) \left[\frac{\|\hat{x} - x\|_{\ell^2}}{c_1\sqrt{rs}} + \sqrt{2/\theta}\|A(\hat{x} - x)\|_{\ell^2} + 5\sqrt{2/\theta}(c_2/c_1)\|A(\hat{x} - x)\|_{\ell^2}\right].
\]

We now use the fact that $\hat{x}$ is a minimizer, and therefore

\[
\|\hat{x}\|_{\ell^1} - \|x\|_{\ell^1} \leq \frac{1}{\lambda} (\|Ax - y\|_{\ell^2} - \|Ax - y\|_{\ell^2}),
\]

Writing $\|A(\hat{x} - x)\|_{\ell^2} \leq \|\hat{x} - y\|_{\ell^2} + \|Ax - y\|_{\ell^2}$ and combining with the previous inequality now yields

\[
\|\hat{x} - x\|_{\ell^2} \leq \left(1 + (c_2^2r/c_1^2)^{1/4}\right) \left[\frac{6\sigma_{s,M}(x)\ell^2}{c_1\sqrt{rs}} + \left(5\sqrt{2/\theta}(c_2/c_1) + \frac{3}{c_1\sqrt{rs}}\right)\|Ax - y\|_{\ell^2}
\]

\[
+ \left(5\sqrt{2/\theta}(c_2/c_1) - \frac{3}{c_1\sqrt{rs}}\right)\|A\hat{x} - y\|_{\ell^2}\right].
\]

The result now follows from the bound (D) on $\lambda$ and the fact that $e = y - Ax$. \hfill \Box

## D Proofs of Lemmas 10.1, 10.2 and 10.3

**Proof of Lemma 10.1.** We first observe that $\theta = \inf_{|\omega| \leq \pi} |\hat{\varphi}(\omega)|^2 > 0$ for the Daubechies wavelet basis [7, Remark 7.1]. Now let $x = (x_n)_{n=1}^N \in \mathbb{C}^N$ with $\|x\|_{\ell^2} = 1$ and write $g = \sum_{n=1}^N x_n \varphi_n$ for the corresponding finite wavelet expansion. Observe that $\|g\|_{L^2([0,1])}^2 = \|x\|_{\ell^2}^2 = 1$. Let $V_j^\text{per} = \text{span}\{\varphi_{j,n} : n = 0, \ldots, 2^j - 1\}$ and $W_j^\text{per} = \text{span}\{\psi_{j,n} : n = 0, \ldots, 2^j - 1\}$. Then

\[
g \in V_{j_0}^\text{per} \oplus W_{j_0}^\text{per} \oplus \cdots \oplus W_{j_0+r-1}^\text{per} = V_{j_0+r}^\text{per},
\]

and conversely every $g \in V_{j_0+r}^\text{per}$ with $\|g\|_{L^2([0,1])}^2 = 1$ is equivalent to a vector of coefficients $x \in \mathbb{C}^M$ with $\|x\|_{\ell^2} = 1$. Note also that

\[
\|P_N U P_N x\|_{\ell^2}^2 = \sum_{n=1}^N |\langle g, \gamma_n \rangle|^2.
\]

Hence

\[
\inf_{\|x\|_{\ell^2} = 1} \|P_N U P_N x\|_{\ell^2}^2 = \inf \left\{ \sum_{n=1}^N |\langle g, \gamma_n \rangle|^2 : g \in V_{j_0+r}^\text{per}, \|g\|_{L^2([0,1])} = 1 \right\}. \tag{D.1}
\]

42
Fix a \( g \in V_{\text{per}}^{j_0} \), with \( \|g\|_{L^2([0,1])} = 1 \) and write
\[
g = \sum_{k=0}^{N-1} z_k \omega_{r+j_0,k}^{\text{per}},
\]
where \( \|z\|_2 = \|g\|_{L^2(0,1)} = 1 \) and \( z = (z_k)_{k=0}^{N-1} \). Then, for any integer \( n \),
\[
\hat{g}(2\pi n) = N^{-1/2} \hat{\phi}(2\pi n/N) \sum_{k=0}^{N-1} z_k e^{-2\pi i k n/N} = N^{-1/2} \hat{\phi}(2\pi n/N) G(n/N),
\]
where \( G(x) = \sum_{k=0}^{N-1} z_k e^{-2\pi i k x} \) is a 1-periodic function. In the first equality, we have used that
\[
\hat{\omega}_{r+j_0,k}(\omega) = \hat{\phi}_{r+j_0,k}(\omega) = 2^{-j/2} \hat{\phi}(\omega/2^j) e^{-i\omega k/2^j}, \quad \forall j,k \in \mathbb{Z}, \forall \omega \in 2\pi \mathbb{Z},
\]
and that \( N = 2^{j_0} + r \). Hence,
\[
\sum_{n=1}^{N} |\langle g, \gamma_n \rangle|^2 = \sum_{n=-N/2+1}^{N/2} |\hat{g}(2\pi n)|^2 = N^{-1} \sum_{n=-N/2+1}^{N/2} |\hat{\phi}(2\pi n/N)|^2 |G(n/N)|^2. \tag{D.3}
\]
Using the fact that \( G \) is 1-periodic we deduce that
\[
\sum_{n=1}^{N} |\langle g, \gamma_n \rangle|^2 \geq \inf_{|\omega| \leq \pi} |\hat{\phi}(\omega)|^2 N^{-1} \sum_{n=0}^{N-1} |G(n/N)|^2.
\]
Now, since \( G \) is a trigonometric polynomial, it follows that
\[
N^{-1} \sum_{n=0}^{N-1} |G(n/N)|^2 = \|G\|_{L^2([0,1])}^2 = \|g\|_{L^2([0,1])}^2 = 1.
\]
Therefore
\[
\sum_{n=1}^{N} |\langle g, \gamma_n \rangle|^2 \geq \inf_{|\omega| \leq \pi} |\hat{\phi}(\omega)|^2 = \theta > 0.
\]
Since \( g \) was arbitrary, we deduce that
\[
\inf_{\|x\|_{\ell^2} = 1} \|P_N U P_N x\|_{\ell^2}^2 \geq \theta.
\]
To complete the proof, we first recall that \( P_N - P_N U^* P_N U P_N \) is positive semidefinite (since \( U \) is unitary), and therefore
\[
\|P_N - P_N U P_N\|_{\ell^2} = \sup_{\|x\|_{\ell^2} = 1} \langle (P_N - P_N U^* P_N U P_N)x, x \rangle
\]
\[
= 1 - \inf_{\|x\|_{\ell^2} = 1} \|P_N U P_N x\|_{\ell^2}^2
\]
\[
\leq 1 - \theta,
\]
as required.
For Lemma 10.2, we first require the following:

**Lemma D.1.** The \((k,l)\)th local coherence satisfies

\[
\mu \left( U^{(k,l)} \right) \leq 2^{1+k-l} \max_{\omega \in B_k} \left| \hat{\psi}(2\pi \omega / 2^l) \right|^2, \quad l > 1,
\]

and

\[
\mu \left( U^{(k,1)} \right) \leq 2^k \max \left\{ \max_{\omega \in B_k} \left| \hat{\psi}(2\pi \omega / 2^j) \right|^2, \max_{\omega \in B_k} \left| \hat{\varphi}(2\pi \omega / 2^j) \right|^2 \right\}.
\]

**Proof.** By definition,

\[
\mu \left( U^{(k,l)} \right) = |B_k| \max_{\omega \in B_k} \max_{0 \leq n < 2^j} \left| \hat{\psi}_{j_0,n}(2\pi \omega) \right|^2, \quad l > 1,
\]

and

\[
\mu \left( U^{(k,1)} \right) = |B_k| \max \left\{ \max_{\omega \in B_k} \left| \hat{\psi}_{j_0,n}(2\pi \omega) \right|^2, \max_{\omega \in B_k} \left| \hat{\varphi}_{j_0,n}(2\pi \omega) \right|^2 \right\}.
\]

Recall that \(|B_k| \leq 2^{j_0+k}\). Moreover, recall relation (D.2) and note that an analogous formula holds for \(\hat{\psi}_{j,k}^{\text{per}}\). Since \(B_k\) is a set of integers, the result now follows immediately.

\[\Box\]

**Proof of Lemma 10.2.** By the previous lemma, it suffices to estimate the Fourier transform of the wavelet and scaling function in different regions of frequency space. First, suppose that \(k \geq l \geq 1\). Then \(|\omega| \geq 2^{j_0+k-1}\) for \(\omega \in B_k\), and the smoothness conditions (2.1) give

\[
|\hat{\psi}(2\pi \omega / 2^l) \hat{\psi}(2\pi \omega / 2^{j_0-1}) | \leq 2^{-(q+1)(k-l)}, \quad |\hat{\varphi}(2\pi \omega / 2^l) \hat{\varphi}(2\pi \omega / 2^{j_0-1}) | \leq 2^{-(q+1)(k-l)}.
\]

The first estimate now follows from Lemma D.1.

For the second estimate, we need to bound \(|\hat{\psi}(2\pi \omega)|\) for \(|\omega| \ll 1\). For this, we recall that \(\hat{\psi}(z) = (-iz)^p \hat{\chi}_p(z)\) for some bounded function \(\hat{\chi}_p(z)\) [51, Thm. 7.4]. Hence

\[
|\hat{\psi}(2\pi \omega)|^2 \leq c_p |\omega|^{2p}.
\]

If \(l > k \geq 1\) then this and the previous lemma give

\[
\mu \left( U^{(k,l)} \right) \leq 2^{1+k-l} \max_{|\omega| \leq 2^{j_0+k}} |\hat{\psi}(2\pi \omega / 2^l) |^2 \leq c_p 2^{k-l} 2^{2p(k-l)}.
\]

The result now follows immediately.

\[\Box\]

**Proof of Lemma 10.3.** By direct calculation

\[
\left\| P_{\Omega} DU P_{M}^\perp d \right\|_{\ell^2}^2 \leq \sum_{k=1}^r \left( \frac{N_k - N_{k-1}}{m_k} \right) \max_{N_{k-1} \leq i \leq N_k} |\langle u_i, P_{M}^\perp d \rangle|^2,
\]

where \(u_i = U^* e_i\) is the \(i\)th row of \(U\). Observe that

\[
|\langle u_i, P_{M}^\perp d \rangle|^2 = \left\| \sum_{j > M} u_{ij} d_j \right\|_2^2 \leq \max_{j > M} |u_{ij}|^2 \left\| P_{M}^\perp d \right\|_{\ell^2}^2.
\]

44
Figure 3: The function $f_K$ defined as in (E.1) for $K = 1, 10, 20$.  

Hence

$$\|P_\Omega D U P_M^\perp d\|_2^2 \leq \sum_{k=1}^r (N_k - N_{k-1}) \max_{N_{k-1} < j \leq N_k} |u_{ij}|^2 \||P_M^\perp d\||_{\ell^1}^2 = \sum_{k=1}^r \mu \left( P_{N_k}^{N_k-1} U P_M^\perp \right) \||P_M^\perp d\||_{\ell^1}^2,$$

which gives

$$\|P_\Omega D U P_M^\perp d\|_2 \leq \left( \sum_{k=1}^r \mu \left( P_{N_k}^{N_k-1} U P_M^\perp \right) \right)^{1/2} \||P_M^\perp d\||_{\ell^1}.$$

Since $M = M_r$, we now apply Lemma 10.2 to get

$$\mu \left( P_{N_k}^{N_k-1} U P_M^\perp \right) = \sup_{l > r} \mu \left( U^{(k,l)} \right) \leq c_p 2^{-(2p+1)(r-k)}.$$

Hence

$$\sum_{k=1}^r \mu \left( P_{N_k}^{N_k-1} U P_M^\perp \right) \leq c_p \sum_{k=1}^r 2^{-(2p+1)(r-k)} \leq c_p.$$  

The result now follows.

E  Numerical experiments

In this section, we discuss some technical details behind Fig. 2. Moreover, we provide further numerical evidence to support the comparison shown therein. We consider the function

$$f_K(x) = \sum_{i=1}^K (-1)^{\text{mod}(i,5)} x^{\text{mod}(i,3)} \text{sign}(x - (1.3)^{i-9}), \quad 0 \leq x \leq 1. \quad (E.1)$$

This function has $K$ discontinuities in $(0, 1)$ and its plot is shown in Fig. 3. We approximate $f_K$ for $K = 1, 10, 20$ using the four different encoder-decoder pairs described below.

(Fourier, $\ell^1$): This strategy corresponds to the setting of Theorems 3.4 and 4.6 and to the error bound (1.5), up to a few minor technical modifications. The Fourier sampling strategy is
Figure 4: Comparison of different encoder-decoder pairs for the approximation of the function $f_K$ defined in (E.1), using Haar (left) and db4 wavelets (right) and for $K = 1$ (top), $K = 10$ (center), and $K = 20$ (bottom).

as follows. We divide the frequency space into dyadic bands and consider a sampling scheme analogous to the $(\mathbf{N}, \mathbf{m})$-multilevel random subsampling strategy with saturation $\tilde{r}$ described in Definition 9.1, where symmetry of the samples is enforced in every frequency band. In particular, $\mathbf{N}$ is defined as in (8.5), the saturation level is $\tilde{r} = \text{round}(\log_2(m/2))$, and the local numbers of measurements are

$$m_k = 2 \left\lfloor \frac{m}{4(r-r_0)} \right\rfloor, \quad k = \tilde{r} + 1, \ldots, r - 1,$$

where, in the last frequency band, we let $m_r = m - (m_1 + \cdots + m_{r-1})$ in order to reach a total budget of $m$ measurements exactly. The samples are then computed as follows. The first $\tilde{r}$
dyadic bands are saturated. For every \( k > \tilde{r} \), we pick \( m_k/2 \) samples uniformly at random from the \( k \)-th frequency semiband (corresponding to positive frequencies) and we choose frequencies in the opposite semiband (corresponding to negative frequencies) in a symmetric way. The wavelet coefficients of \( f \) are recovered via basis pursuit (1.2). Numerically, (1.2) is solved using the Matlab package SPGL1 (see \cite{64, 65}) with parameters \( \text{bpTol} = 1e-6, \text{optTol} = 1e-6 \), and a maximum of 10000 iterations.\(^4\)

**(Gauss, \( \ell^1 \)):** This is the standard encoder-decoder pair of compressed sensing with random Gaussian measurements, corresponding to the setting of Theorems 3.1 and 4.3 and to the error bound (1.3). The vector \( d \in \mathbb{R}^N \) of wavelet coefficients of \( f \) is explicitly computed and then encoded as \( y = Ad \), where \( A \in \mathbb{R}^{m \times N} \) has i.i.d. entries drawn from the normal distribution with mean zero and variance \( 1/m \). The function is recovered by means of the basis pursuit decoder (1.2), numerically solved using SPGL1 as in the previous case.\(^5\)

**((Optimal, \( \ell^1 \)):** This strategy corresponds to the setting of Theorems 3.1 and 4.3 and to the optimal error bound (1.3). As in the previous case, we compute the vector \( d \in \mathbb{R}^N \) of wavelet coefficients of \( f \). Then, the first \( m_1 = \text{round}(m/2) \) entries of \( d \) are directly encoded into \( y^{(1)} \in \mathbb{R}^{m_1} \). The remaining \( m_2 = m - m_1 \) measurements are computed as \( y^{(2)} = A(d_{n})_{n=m_1+1}^{N} \), where \( A \in \mathbb{R}^{m_2 \times (N-m_1)} \) has i.i.d. entries drawn from the normal distribution with mean zero and variance \( 1/m_2 \). We consider the basis pursuit decoder (1.2), numerically solved using SPGL1 as in the previous cases.

**((Gauss, Tree)):** This encoder-decoder pair corresponds to the model-based compressive sensing strategy proposed in \cite{12}. The encoder identical to (Gauss, \( \ell^1 \)), and the decoder explicitly promotes tree-structured sparsity in the recovered function using the model-based CoSaMP algorithm \cite{12}. This strategy requires tuning a parameter \( c \), which links \( m \) to the desired tree-sparsity level \( s \) as \( m = cs \). In the numerical tests, we consider \( c = 3, 4, 5, 6, 7 \). We employ the Model-based Compressive Sensing Toolbox v1.1 provided by the authors of \cite{12}. The maximum number of iterations for the outer loop of CoSaMP is set to 100.

These four encoder-decoder pairs are compared with \( N = 2^{15} = 32768 \) and values of \( m \) ranging from \( 2^3 = 8 \) to \( 2^9 = 512 \). We employ Haar and db4 wavelets, having \( p = 1 \) and \( p = 2 \) vanishing moments, respectively. The relative \( L^2 \) error is computed using the wavelet coefficients of \( f \), approximated as in the strategies (Gauss, \( \ell^1 \)), (Optimal, \( \ell^1 \)), and (Gauss, Tree).

In Fig. 4 the encoder-decoder pair (Fourier, \( \ell^1 \)) consistently outperforms all the other strategies, with only a few exceptions. Moreover, this behaviour is independent of the number of discontinuities \( K \). It is remarkable that (Fourier, \( \ell^1 \)) is able to numerically outperform even the theoretically-optimal pair (Optimal, \( \ell^1 \)). Although our theory prescribes the use of weighted square-root LASSO decoder in the Fourier case, the numerics show that employing basis pursuit (1.2) is enough to numerically outperform the other strategies. Using weighted

\(^4\)The entries of the cross-Gramian matrix \( U \) (8.1) used in this sampling strategy are computed by applying the inverse discrete wavelet and Fourier transforms to the first \( N \) elements of the canonical basis of the augmented space \( \mathbb{R}^{16N} \). Then, only the \( N \) entries corresponding to the frequencies of interest are kept. This augmentation makes the computation of \( U \) more accurate.

\(^5\)In order to avoid discretization effects related to the wavelet crime, the vector \( d \) of wavelet coefficients is computed by sampling the function \( f \) on a uniform grid of \( 16N \) points, applying the discrete wavelet transform, and then keeping the first \( N \) entries of the resulting vector.
\textit{\ell}^1\textit{ minimization could lead to further improvements.}

\textbf{References}

[1] http://www3.gehealthcare.in/~media/images/product/product-categories/magnetic-resonance-imaging/optima-mr450w-1-5t-with-gem-suite/1-clinical/optima_mr450w_with_gem_suite_brainpropt2_clinical.jpg.

[2] B. Adcock, V. Antun, and A. C. Hansen. Uniform recovery in infinite-dimensional compressed sensing and applications to structured binary sampling. \textit{arXiv:1905.00126}, 2019.

[3] B. Adcock, A. Bao, and S. Brugiapaglia. Correcting for unknown errors in sparse high-dimensional function approximation. \textit{Numer. Math. (to appear)}, 2019.

[4] B. Adcock, C. Boyer, and S. Brugiapaglia. On the gap between local recovery guarantees in compressed sensing and oracle estimates. \textit{arXiv:1806.03789}, 2018.

[5] B. Adcock and A. C. Hansen. Generalized sampling and infinite-dimensional compressed sensing. \textit{Found. Comput. Math.}, 16(5):1263–1323, 2016.

[6] B. Adcock, A. C. Hansen, G. Kutyniok, and J. Ma. Linear stable sampling rate: Optimality of 2D wavelet reconstructions from Fourier measurements. \textit{SIAM J. Math. Anal.}, 47(2):1196–1233, 2015.

[7] B. Adcock, A. C. Hansen, and C. Poon. On optimal wavelet reconstructions from Fourier samples: linearity and universality of the stable sampling rate. \textit{Appl. Comput. Harmon. Anal.}, 36(3):387–415, 2014.

[8] B. Adcock, A. C. Hansen, C. Poon, and B. Roman. Breaking the coherence barrier: A new theory for compressed sensing. \textit{Forum Math. Sigma}, 5, 2017.

[9] B. Adcock, A. C. Hansen, and B. Roman. The quest for optimal sampling: computationally efficient, structure-exploiting measurements for compressed sensing. In \textit{Compressed Sensing and Its Applications}. Springer, 2015.

[10] B. Adcock, A. C. Hansen, B. Roman, and G. Teschke. Generalized sampling: stable reconstructions, inverse problems and compressed sensing over the continuum. \textit{Advances in Imaging and Electron Physics}, 182:187–279, 2014.

[11] G. R. Arce, D. J. Brady, L. Carin, H. Arguello, and D. Kittle. Compressive coded aperture spectral imaging: An introduction. \textit{IEEE Signal Process. Mag.}, 31(1):105–115, 2014.

[12] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hedge. Model-based compressive sensing. \textit{IEEE Trans. Inform. Theory}, 56(4):1982–2001, 2010.

[13] A. Bastounis, B. Adcock, and A. C. Hansen. From global to local: Getting more from compressed sensing. \textit{SIAM News}, 2017.

[14] A. Bastounis and A. C. Hansen. On the absence of the RIP in real-world applications of compressed sensing and the RIP in levels. \textit{SIAM J. Imaging Sci.}, 2017 (to appear).

[15] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. \textit{Biometrika}, 98(4):791–806, 2011.

[16] V. Boominathan, J. K. Adams, M. S. Asif, B. W. Avants, J. T. Robinson, R. G. Baraniuk, A. C. Sankaranarayanan, and A. Veeraraghavan. Lensless imaging: A computational renaissance. \textit{IEEE Signal Process. Mag.}, 33(5):23–35, 2016.

[17] C. Boyer, J. Bigot, and P. Weiss. Compressed sensing with structured sparsity and structured acquisition. \textit{Appl. Comput. Harm. Anal.}, 46(2):312–350, 2017.
[18] D. J. Brady, K. Choi, D. L. Marks, R. Horisaki, and S. Lim. Compressive holography. *Opt. Express*, 17:13040–13049, 2009.

[19] S. Brugiapaglia and B. Adcock. Robustness to unknown error in sparse regularization. *IEEE Trans. Inform. Theory*, 64(10):6638–6661, 2018.

[20] T. Cai and A. Zhang. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE Trans. Inform. Theory*, 60(1):122–132, 2014.

[21] E. Candes. The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris*, 346(9-10):589–592, 2008.

[22] E. J. Candès and Y. Plan. A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inform. Theory*, 57(11):7235–7254, 2011.

[23] E. J. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, 2007.

[24] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[25] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145, 2011.

[26] N. Chauffert, P. Ciuciu, J. Kahn, and P. Weiss. Variable density sampling with continuous trajectories. *SIAM J. Imaging Sci.*, 7(4):1962–1992, 2014.

[27] A. Chkifa, N. Dexter, H. Tran, and C. G. Webster. Polynomial approximation via compressed sensing of high-dimensional functions on lower sets. *Math. Comp.*, 87:1415–1450, 2018.

[28] A. Cohen, W. Dahmen, and R. A. DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, 22(1):211–231, 2009.

[29] I. Daubechies. *Ten Lectures on Wavelets*, volume 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.

[30] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok. Introduction to compressed sensing. In *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2011.

[31] R. DeVore, G. Petrova, and P. Wojtaszczyk. Instance-optimality in probability with an \(\ell_1\)-minimization decoder. *Appl. Comput. Harmon. Anal.*, 27(3):275–288, 2009.

[32] R. A. DeVore. Nonlinear approximation. *Acta Numer.*, 7:51–150, 1998.

[33] M. F. Duarte, M. A. Davenport, D. Takhar, J. Laska, K. Kelly, and R. G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Process. Mag.*, 25(2):83–91, 2008.

[34] M. F. Duarte and Y. C. Eldar. Structured compressed sensing: from theory to applications. *IEEE Trans. Signal Process.*, 59(9):4053–4085, 2011.

[35] J. A. Fessler. Optimization methods for MR image reconstruction. *arXiv:1903.05510*, 2019.

[36] S. Foucart. Stability and robustness of \(\ell_1\)-minimizations with Weibull matrices and redundant dictionaries. *Linear Algebra Appl.*, 441:4–21, 2014.

[37] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhauser, 2013.

[38] M. Gataric and C. Poon. A practical guide to the recovery of wavelet coefficients from Fourier measurements. *SIAM J. Sci. Comput.*, 38(2):A1075–A1099, 2016.
[39] M. E. Gehm and D. J. Brady. Compressive sensing in the EO/IR. *Applied Optics*, 54(8):C14–C22, 2015.
[40] C. G. Graff and E. Y. Sidky. Compressive sensing in medical imaging. *Appl. Opt.*, 54:C23–C44, 2015.
[41] J. Haldar, D. Hernando, and Z. Liang. Compressed-sensing MRI with random encoding. *IEEE Trans. Med. Imaging*, 30(4):893–903, 2011.
[42] D. J. Holland, M. J. Bostock, L. F. Gladden, and D. Nietlispach. Fast multidimensional NMR spectroscopy using compressed sensing. *Angew. Chem. Int. Ed.*, 50(29), 2011.
[43] G. Huang, H. Jiang, K. Matthews, and P. Wilford. Lensless imaging by compressive sensing. In *20th IEEE International Conference on Image Processing*, 2013.
[44] O. Katz, Y. Bromberg, and Y. Silberberg. Compressive ghost imaging. *Appl. Phys. Lett.*, 95:131110, 2009.
[45] K. Kazimierczuk and V. Y. Orekhov. Accelerated NMR spectroscopy by using compressed sensing. *Angew. Chem. Int. Ed.*, 50(24), 2011.
[46] F. Krahmer and R. Ward. Stable and robust recovery from variable density frequency samples. *IEEE Trans. Image Proc.*, 23(2):612–622, 2013.
[47] G. Kutyniok and W.-Q. Lim. Optimal compressive imaging of Fourier data. *SIAM J. Imaging Sci.*, 11(1):507–546, 2018.
[48] C. Li and B. Adcock. Compressed sensing with local structure: uniform recovery guarantees for the sparsity in levels class. *Appl. Comput. Harmon. Anal.*, 46(3):453–477, 2019.
[49] M. Lustig, D. L. Donoho, and J. M. Pauly. Sparse MRI: the application of compressed sensing for rapid MRI imaging. *Magn. Reson. Med.*, 58(6):1182–1195, 2007.
[50] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly. Compressed Sensing MRI. *IEEE Signal Process. Mag.*, 25(2):72–82, March 2008.
[51] S. G. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, 3 edition, 2009.
[52] R. F. Marcia, R. M. Willett, and Z. T. Harmany. Compressive optical imaging: Architectures and algorithms. In G. Cristobal, P. Schelken, and H. Thienpont, editors, *Optical and Digital Image Processing: Fundamentals and Applications*, pages 485–505. Wiley New York, 2011.
[53] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar. Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Trans. Graph.*, 32(46), 2013.
[54] C. Poon. On the role of total variation in compressed sensing. *SIAM J. Imaging Sci.*, 8(1):682–720, 2015.
[55] C. Poon. Structure dependent sampling in compressed sensing: theoretical guarantees for tight frames. *Appl. Comput. Harm. Anal.*, 42(3):402–451, 2017.
[56] G. Puy, J. P. Marques, R. Gruetter, J. Thiran, D. Van De Ville, P. Vanderheynst, and Y. Wiaux. Spread spectrum Magnetic Resonance Imaging. *IEEE Trans. Med. Imaging*, 31(3):586–598, 2012.
[57] X. Qu, Y. Chen, X. Zhuang, Z. Yan, D. Guo, and Z. Chen. Spread spectrum compressed sensing MRI using chirp radio frequency pulses. *arXiv:1301.5451*, 2013.
[58] B. Roman, A. Bastounis, B. Adcock, and A. C. Hansen. On fundamentals of models and sampling in compressed sensing. *Preprint*, 2015.
[59] B. Roman, A. C. Hansen, and B. Adcock. On asymptotic structure in compressed sensing. *arXiv:1406.4178*, 2014.
[60] J. Romberg. Imaging via compressive sampling. *IEEE Signal Process. Mag.*, 25(2):14–20, 2008.

[61] V. Studer, J. Bobin, M. Chahid, H. Moussavi, E. Candès, and M. Dahan. Compressive fluorescence microscopy for biological and hyperspectral imaging. *Proc. Natl Acad. Sci. USA*, 109(26):1679—1687, 2011.

[62] Y. Traonmilin and R. Gribonval. Stable recovery of low-dimensional cones in Hilbert spaces: One RIP to rule them all. *Appl. Comput. Harm. Anal.*, 45(1):170–205, 2018.

[63] Y. Tsaig and D. L. Donoho. Extensions of compressed sensing. *Signal Process.*, 86(3):549–571, 2006.

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

[65] E. van den Berg and M. P. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comput.*, 31(2):890–912, 2008.

[66] Z. Wang and G. R. Arce. Variable density compressed image sampling. *IEEE Trans. Image Proc.*, 19(1):264–270, 2010.

[67] Y. Wiaux, L. Jacques, G. Puy, A. M. M. Scaife, and P. Vandergheynst. Compressed sensing imaging techniques for radio interferometry. *Mon. Not. R. Astron. Soc.*, 395(3):1733–1742, 2009.

[68] P. Wojtaszczyk. Stability and instance optimality for Gaussian measurements in compressed sensing. *Found. Comput. Math.*, 10(1):1–13, 2010.

[69] L. Zhu, W. Zhang, D. Elnatan, and B. Huang. Faster STORM using compressed sensing. *Nature Methods*, 9:721—723, 2012.