3-13-2013

Stable Image Reconstruction Using Total Variation Minimization

Deanna Needell
Claremont McKenna College

Rachel Ward
University of Texas at Austin

Recommended Citation

Needell, D., Ward, R., "Stable Image Reconstruction Using Total Variation Minimization," SIAM J. Imaging Sci., 6(2), 1035–1058, 2013. doi: 10.1137/120868281

SIAM J. IMAGING SCIENCES
Vol. 6, No. 2, pp. 1035–1058

Stable Image Reconstruction Using Total Variation Minimization[*]

Deanna Needell[†] and Rachel Ward[‡]

Abstract. This paper presents near-optimal guarantees for stable and robust image recovery from undersampled noisy measurements using total variation minimization. In particular, we show that from $O(s\log(N))$ nonadaptive linear measurements, an image can be reconstructed to within the best s-term approximation of its gradient up to a logarithmic factor, and this factor can be removed by taking slightly more measurements. Along the way, we prove a strengthened Sobolev inequality for functions lying in the null space of a suitably incoherent matrix.

Key words. compressed sensing, stability, restricted isometry property, Sobolev inequality, total variation minimization

AMS subject classifications. 41A46, 68Q25, 68W20, 90C27

DOI. 10.1137/120868281

1. Introduction.

Compressed sensing (CS) provides the technology to exploit sparsity when acquiring signals of general interest, allowing for accurate and robust signal acquisition from surprisingly few measurements. Rather than acquiring an entire signal and then later compressing, CS proposes a mechanism to collect measurements in compressed form, skipping the often costly step of complete acquisition. The applications are numerous and range from image and signal processing to remote sensing and error correction [20].

In compressed sensing one acquires a signal $\boldsymbol{x} \in \mathbb{C}^d$ via $m \ll d$ linear measurements of the form $y_k = \langle \boldsymbol{\phi_k}, \boldsymbol{x}\rangle + z_k$. The vectors $\boldsymbol{\phi_k}$ form the rows of the *measurement matrix* $\boldsymbol{\Phi}$, and the measurement vector $\boldsymbol{y} \in \mathbb{C}^m$ can thus be viewed in matrix notation as

$$\boldsymbol{y} = \boldsymbol{\Phi x} + \boldsymbol{z},$$

where $\boldsymbol{z}$ is the noise vector modeling measurement error. We then ask to recover the signal of interest $\boldsymbol{x}$ from the noisy measurements $\boldsymbol{y}$. Since $m \ll d$, this problem is ill-posed without further assumptions. However, signals of interest in applications contain far less information than their dimension d would suggest, often in the form of sparsity or compressibility in a given basis. We call a vector $\boldsymbol{x}$ s-sparse when

$$\|\boldsymbol{x}\|_0 \overset{\text{def}}{=} |\operatorname{supp}(\boldsymbol{x})| \le s \ll d. \tag{1}$$

Compressible vectors are those which are approximated well by sparse vectors.

[*] Received by the editors March 1, 2012; accepted for publication (in revised form) March 4, 2013; published electronically June 4, 2013.
http://www.siam.org/journals/siims/6-2/86828.html

[†] Department of Mathematics and Computer Science, Claremont McKenna College, Claremont, CA 91711 (dneedell@cmc.edu).

[‡] Mathematics Department, University of Texas at Austin, Austin, TX 78712 (rward@math.utexas.edu). This author's research was supported in part by a Donald D. Harrington Faculty Fellowship, Alfred P. Sloan Research Fellowship, and DOD-Navy grant N00014-12-1-0743.

In the simplest case, if we know that $\boldsymbol{x}$ is s-sparse and the measurements are free of noise, then the inverse problem $\boldsymbol{y} = \boldsymbol{\Phi x}$ is well-posed if the measurement matrix $\boldsymbol{\Phi}$ is one-to-one on sparse vectors. To recover $\boldsymbol{x} \in \mathbb{C}^d$ from $\boldsymbol{y} \in \mathbb{C}^m$ we solve the optimization problem

$$(L_0) \qquad \hat{\boldsymbol{x}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \|\boldsymbol{w}\|_0 \quad \text{such that} \quad \boldsymbol{\Phi w} = \boldsymbol{y}.$$

If $\boldsymbol{\Phi}$ is one-to-one on s-sparse vectors, then (L_0) recovers exactly any s-sparse signal $\hat{\boldsymbol{x}} = \boldsymbol{x}$. The optimization problem (L_0), however, is in general NP-hard [39], so we instead consider its relaxation to the ℓ_1-norm,

$$(L_1) \qquad \hat{\boldsymbol{x}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \|\boldsymbol{w}\|_1 \quad \text{such that} \quad \|\boldsymbol{\Phi w} - \boldsymbol{y}\|_2 \le \varepsilon,$$

where $\|\boldsymbol{w}\|_1 = \sum_i |w_i|$, $\|\boldsymbol{w}\|_2 = \left(\sum_i w_i^2\right)^{1/2}$ denotes the standard Euclidean norm, and ε bounds the noise level $\|\boldsymbol{z}\|_2 \le \varepsilon$. The problem (L_1) may be cast as a second order cone program (SOCP) and can thus be solved efficiently using modern convex programming methods [17, 21].

If we require that the measurement matrix be not only one-to-one on s-sparse vectors, but moreover an approximate *isometry* on s-sparse vectors, then remarkably (L_1) will still recover any s-sparse signal exactly. Candès and Tao introduced the *restricted isometry property* (RIP) and showed that this requirement on the measurement matrix $\boldsymbol{\Phi}$ guarantees robust recovery of compressible signals via (L_1) [12].

Definition 1. *A matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m \times d}$ is said to have the RIP of order s and level $\delta \in (0,1)$ if*

$$(2) \qquad (1-\delta)\|\boldsymbol{x}\|_2^2 \le \|\boldsymbol{\Phi x}\|_2^2 \le (1+\delta)\|\boldsymbol{x}\|_2^2 \qquad \forall\, s\text{-sparse } \boldsymbol{x} \in \mathbb{C}^d.$$

The smallest such δ for which this holds is denoted by δ_s and called the restricted isometry constant *for the matrix $\boldsymbol{\Phi}$.*

When $\delta_{2s} < 1$, the RIP guarantees that no $2s$-sparse vectors reside in the null space of $\boldsymbol{\Phi}$. When a matrix has a small restricted isometry constant, $\boldsymbol{\Phi}$ acts as a near-isometry over the subset of s-sparse signals.

Many classes of random matrices can be used to generate matrices having small RIP constants. With probability exceeding $1 - e^{-Cm}$, a matrix whose entries are independent and identically distributed appropriately normalized Gaussian random variables has a small RIP constant $\delta_s < c$ when $m \gtrsim c^{-2} s \log(d/s)$. This number of measurements is also shown to be necessary for the RIP [30]. More generally, the RIP holds with high probability for any matrix generated by a sub-Gaussian random variable [13, 36, 48, 2]. One can also construct matrices with the RIP using fewer random bits. For example, if $m \gtrsim s \log^4(d)$, then the RIP holds with high probability for the *random subsampled Fourier matrix* $\boldsymbol{F}_\Omega \in \mathbb{C}^{m \times d}$, formed by restricting the $d \times d$ discrete Fourier matrix to a random subset of m rows and renormalizing [48]. The RIP also holds for randomly subsampled bounded orthonormal systems [47, 45] and randomly generated circulant matrices [46].

Candès, Romberg, and Tao showed that when the measurement matrix $\boldsymbol{\Phi}$ satisfies the RIP with sufficiently small restricted isometry constant, (L_1) produces an estimation $\hat{\boldsymbol{x}}$ to $\boldsymbol{x}$ with error [11],

$$(3) \qquad \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \le C\left(\frac{\|\boldsymbol{x} - \boldsymbol{x}_s\|_1}{\sqrt{s}} + \varepsilon\right).$$

This error rate is optimal on account of classical results about the Gel'fand widths of the ℓ_1 ball due to Kashin [28] and Garnaev and Gluskin [25].

Here and throughout, $\boldsymbol{x_s}$ denotes the vector consisting of the largest s coefficients of $\boldsymbol{x}$ in magnitude. Similarly, for a set S, $\boldsymbol{x}_S$ denotes the vector (or matrix, appropriately) of $\boldsymbol{x}$ restricted to the entries indexed by S. The bound (3) then says that the recovery error is proportional to the noise level and the norm of the tail of the signal, $\boldsymbol{x} - \boldsymbol{x_s}$. In particular, when the signal is exactly sparse and there is no noise in the measurements, (L_1) recovers $\boldsymbol{x}$ *exactly*. We note that for simplicity, we restrict focus to CS decoding via the program (L_1), but acknowledge that other approaches in compressed sensing such as compressive sampling matching pursuit [40] and iterative hard thresholding [4] yield analogous recovery guarantees.

Signals of interest are often compressible with respect to bases other than the canonical basis. We consider a vector $\boldsymbol{x}$ to be s-sparse with respect to the basis $\boldsymbol{B}$ if

$$\boldsymbol{x} = \boldsymbol{B}\boldsymbol{z} \quad \text{for some } s\text{-sparse } \boldsymbol{z},$$

and $\boldsymbol{x}$ is compressible with respect to this basis when it is well approximated by a sparse representation. In this case one may recover $\boldsymbol{x}$ from CS measurements $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{\xi}$ using the modified ℓ_1 minimization problem

$$(BL_1) \qquad \hat{\boldsymbol{x}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \|\boldsymbol{B}^*\boldsymbol{w}\|_1 \quad \text{such that} \quad \|\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{y}\|_2 \leq \varepsilon.$$

As before, the recovery error $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2$ is proportional to the noise level and the norm of the tail of the signal if the composite matrix $\boldsymbol{\Psi} = \boldsymbol{\Phi}\boldsymbol{B}$ satisfies the RIP. If $\boldsymbol{B}$ is a fixed orthonormal matrix and $\boldsymbol{\Phi}$ is a random matrix generated by a sub-Gaussian random variable, then $\boldsymbol{\Psi} = \boldsymbol{\Phi}\boldsymbol{B}$ has the RIP with high probability with $m \gtrsim s\log(d/s)$ due to the invariance of norm-preservation for sub-Gaussian matrices [2]. More generally, following the approach of [2] and applying Proposition 3.2 in [30], this rotation-invariance holds for any $\boldsymbol{\Phi}$ with the RIP and randomized column signs. The rotational-invariant RIP also extends to the classic ℓ_1-analysis problem which solves (BL_1) when $\boldsymbol{B}^*$ is a tight frame [8].

1.1. Imaging with CS.

Grayscale digital images do not fill the entire space of $N \times N$ blocks of pixel values, consisting primarily of slowly varying pixel intensities except along edges. In other words, digital images are compressible with respect to their discrete gradient. Concretely, we denote an $N \times N$ block of pixels by $\boldsymbol{X} \in \mathbb{C}^{N\times N}$, and we write $X_{j,k}$ to denote any particular pixel. The discrete directional derivatives of $\boldsymbol{X} \in \mathbb{C}^{N\times N}$ are defined pixelwise as

$$(4) \qquad \boldsymbol{X}_x : \mathbb{C}^{N\times N} \to \mathbb{C}^{(N-1)\times N}, \qquad (\boldsymbol{X}_x)_{j,k} = \boldsymbol{X}_{j+1,k} - \boldsymbol{X}_{j,k},$$

$$(5) \qquad \boldsymbol{X}_y : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times (N-1)}, \qquad (\boldsymbol{X}_y)_{j,k} = \boldsymbol{X}_{j,k+1} - \boldsymbol{X}_{j,k}.$$

The discrete gradient transform $\boldsymbol{\nabla} : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N\times 2}$ is defined in terms of the directional derivatives and in matrix form,

$$[\boldsymbol{\nabla}\boldsymbol{X}]_{j,k} \stackrel{\text{def}}{=} \begin{cases} \big((\boldsymbol{X}_x)_{j,k}, (\boldsymbol{X}_y)_{j,k}\big), & 1 \leq j \leq N-1, \quad 1 \leq k \leq N-1, \\ \big(0, (\boldsymbol{X}_y)_{j,k}\big), & j = N, \quad 1 \leq k \leq N-1, \\ \big((\boldsymbol{X}_x)_{j,k}, 0\big), & k = N, \quad 1 \leq j \leq N-1, \\ \big(0, 0\big), & j = k = N. \end{cases}$$

Finally, the *total variation* seminorm is just the sum of the magnitudes of its discrete gradient,

$$\|\boldsymbol{X}\|_{TV} \overset{\text{def}}{=} \|\boldsymbol{\nabla X}\|_1. \tag{6}$$

We note here that the choice of $((\boldsymbol{X}_x)_{j,k}, (\boldsymbol{X}_y)_{j,k})$ in the definition of $\left[\boldsymbol{\nabla X}\right]_{j,k}$ leads to the *anisotropic* version of the total variation norm. The *isotropic* version of the total variation norm instead stems from the choice of $(\boldsymbol{X}_x)_{j,k} + i(\boldsymbol{X}_y)_{j,k}$ in the definition of the discrete gradient. In the isotropic case, $\|\boldsymbol{X}\|_{TV}$ becomes the sum of terms

$$\left|(\boldsymbol{X}_x)_{j,k} + i(\boldsymbol{X}_y)_{j,k}\right| = \left((\boldsymbol{X}_x)_{j,k}^2 + (\boldsymbol{X}_y)_{j,k}^2\right)^{1/2}.$$

The isotropic and anisotropic induced total variation norms are thus equivalent up to a factor of $\sqrt{2}$. We emphasize here that our method applies to both anisotropic and isotropic total variation. However, we will consider only the anisotropic case for simplicity because the treatment of the isotropic case is analogous.

Natural images have small total variation due to the low-dimensionality of their subset of pixels representing edges. As such, searching for the image with smallest total variation that matches a set of measurements, the convex relaxation of searching for the image with fewest edges, is a natural choice for image reconstruction. In the context of CS, the measurements $\boldsymbol{y} \in \mathbb{C}^m$ from an image $\boldsymbol{X}$ are of the form $\boldsymbol{y} = \mathcal{M}(\boldsymbol{X}) + \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is a noise term and $\mathcal{M} : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ is a linear operator defined via its components by

$$[\mathcal{M}(\boldsymbol{X})]_j \overset{\text{def}}{=} \langle \boldsymbol{M_j}, \boldsymbol{X}\rangle = \operatorname{trace}(\boldsymbol{M_j}\boldsymbol{X}^*)$$

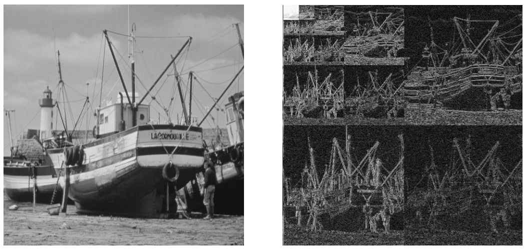
for suitable matrices $\boldsymbol{M_j}$. Here and throughout, $\boldsymbol{M}^*$ denotes the adjoint of the matrix $\boldsymbol{M}$. Total variation minimization refers to the convex optimization

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{Z}}{\operatorname{argmin}} \|\boldsymbol{Z}\|_{TV} \quad \text{such that} \quad \|\mathcal{M}(\boldsymbol{Z}) - \boldsymbol{y}\|_2 \le \varepsilon. \tag{TV}$$

The standard theory of compressed sensing does not apply to total variation minimization. In fact, the gradient transform $\boldsymbol{Z} \to \boldsymbol{\nabla Z}$ not only fails to be orthonormal, but, viewed as an invertible operator over mean-zero images, the Frobenius operator norm of its inverse grows *linearly with* N. This poor conditioning would lead to magnification of error even if the usual CS techniques could be applied.

Despite this, total variation minimization (TV) is widely used in applications and exhibits accurate image reconstruction empirically (see, e.g., [11, 14, 10, 43, 16, 33, 34, 32, 42, 35, 27, 29]). However, to the best of our knowledge there have been no provable guarantees that (TV) recovery is robust.

Images are also compressible with respect to wavelet transforms. For example, in Figure 1 we display the image of boats alongside its (bivariate) Haar wavelet transform. The Haar transform (like wavelet transforms more generally) is multiscale, collecting information not only about local differences in pixel intensity, but also about differences in average pixel intensity on all dyadic scales. Therefore, the level of compressibility in the wavelet domain is controlled by the total variation seminorm [22]. We will use in particular that the rate of decay of the bivariate Haar coefficients of an image can be bounded by the total variation (see Proposition 8 in section 4).



(a) (b)

Figure 1. (a) *Original boats image and* (b) *its bivariate Haar coefficients.*

Recall that the (univariate) Haar wavelet system constitutes a complete orthonormal system for square-integrable functions on the unit interval, consisting of the constant function

$$H^0(t) = \begin{cases} 1, & 0 \le t < 1, \\ 0 & \text{otherwise,} \end{cases}$$

the mother wavelet

$$H^1(t) = \begin{cases} 1, & 0 \le t < 1/2, \\ -1, & 1/2 \le t < 1, \end{cases}$$

and dyadic dilations and translates of the mother wavelet

$$H_{n,k}(t) = 2^{n/2} H^1(2^n t - k), \quad n \in \mathbb{N}, \quad 0 \le k < 2^n. \tag{7}$$

The *bivariate* Haar system comprises an orthonormal system for $L_2(Q)$, the space of square-integrable functions on the unit square $Q = [0,1)^2$, and is derived from the univariate Haar system by the usual tensor-product construction. In particular, starting from the multivariate functions

$$H^e(u,v) = H^{e_1}(u) H^{e_2}(v), \quad e = (e_1, e_2) \in V = \big\{\{0,1\},\{1,0\},\{1,1\}\big\},$$

the bivariate Haar system consists of the constant function and all functions

$$x = (u,v), \qquad H^e_{j,k}(x) = 2^j H^e(2^j x - k), \quad e \in V, \quad j \ge 0, \quad k \in \mathbb{Z}^2 \cap 2^j Q. \tag{8}$$

Discrete images are isometric to the space $\Sigma_N \subset L_2(Q)$ of piecewise-constant functions

$$\Sigma_N = \left\{ f \in L_2(Q), \quad f(u,v) = c_{j,k}, \quad \frac{j-1}{N} \le u < \frac{j}{N}, \quad \frac{k-1}{N} \le v < \frac{k}{N} \right\} \tag{9}$$

via the identification $c_{j,k} = N X_{j,k}$. Letting $N = 2^n$, the bivariate Haar basis restricted to the N^2 basis functions $\{H^e_{j,k} : j \le n-1\}$ and identified via (9) as discrete images $\boldsymbol{h}^e_{j,k}$, forms an orthonormal basis for $\mathbb{C}^{N \times N}$. We denote by $\mathcal{H}(\boldsymbol{X})$ the matrix product that computes the *discrete bivariate Haar transform* $\boldsymbol{X} \to (\langle \boldsymbol{X}, \boldsymbol{h}^e_{j,k} \rangle)_{j,k,e}$.

Because the bivariate Haar transform is orthonormal, standard CS results guarantee that images can be reconstructed up to a factor of their best approximation by s Haar basis

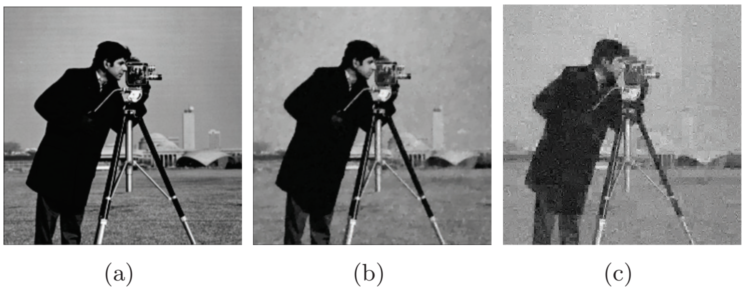


Figure 2. (a) *Original* 256×256 *cameraman image and its reconstruction from* 20% *randomly selected Fourier coefficients using* (b) *total variation minimization and* (c) ℓ_1*-minimization of its bivariate Haar coefficients.*

functions using $m \gtrsim s \log(N)$ measurements. One might then consider ℓ_1-minimization of the Haar coefficients as an alternative to total variation minimization. However, total variation minimization (TV) gives better empirical image reconstruction results than ℓ_1-Haar wavelet coefficient minimization, despite not being fully justified by CS theory. For details, see [10, 11, 23] and references therein. For example, Figure 2 shows the reconstruction of the cameraman image using 20% of its discrete Fourier coefficients (selected at random). The image recovered via total variation minimization (TV) is not only more pleasing to the eye but has a much lower recovery error than that of the image recovered via Haar minimization, e.g., (BL_1) with orthonormal transform $\boldsymbol{B} = \mathcal{H}$.

In the case of noise, Figure 3 displays the original Fabio image, corrupted with additive Gaussian noise. Again, we compare the performance of (TV) and (BL_1) at reconstruction using 20% Fourier measurements. As is evident, total variation minimization outperforms Haar minimization in the presence of noise as well. Another type of measurement noise is a consequence of round-off or quantization error. This type of error may stem from the inability to take measurements with arbitrary precision and differs from Gaussian noise since it depends on the signal itself. Figure 4 displays the lake image with quantization error along with the recovered images. As in the case of Gaussian noise, total variation minimization outperforms Haar minimization. All experiments here and throughout used the software ℓ_1-magic to solve the minimization programs [24].

We note that the use of total variation regularization in image processing predates the theory of CS. The seminal paper of Rudin, Osher, and Fatemi introduced total variation regularization in imaging [49], and subsequently total variation has become a regularizer of choice for image denoising, deblurring, inpainting, and segmentation [9, 43, 50, 16, 15]. For more details on the connections between total variation minimization and wavelet frame-based methods in image analysis, we refer the reader to [6].

1.2. Contribution of this paper.

Although theoretical guarantees have been obtained guaranteeing recovery via (TV) of images with exactly sparse gradients without noise [11, 14], to the best of our knowledge no results have shown that this recovery is robust, despite strong suggestions by numerical evidence. In this paper, we prove precisely this. *We show that* (TV) *robustly recovers images* from a few RIP measurements. The error guarantees are analogous to those of (3) up to a logarithmic factor, which can be removed by taking slightly more measurements (see Theorem 6 below).

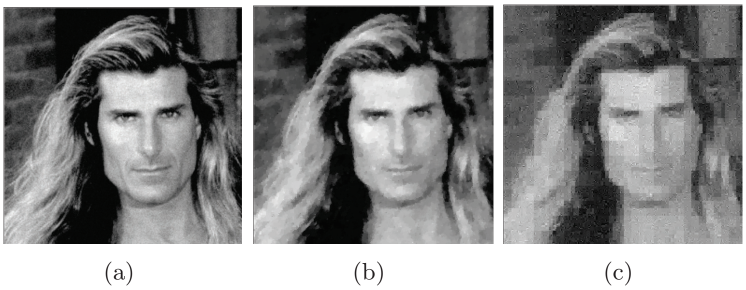


Figure 3. (a) *Original* 256 × 256 *Fabio image corrupted with Gaussian noise and its reconstruction from* 20% *randomly selected Fourier coefficients using* (b) *total variation minimization and* (c) ℓ_1*-minimization of its bivariate Haar coefficients. Original image* (a) *used with the kind permission of Fabio, Inc., Santa Monica, CA.*

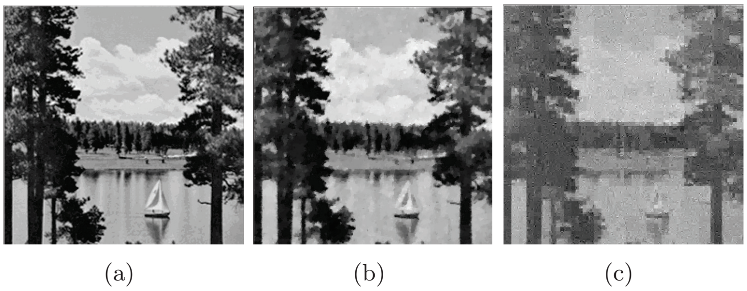


Figure 4. (a) *Original* 256 × 256 *lake image corrupted with quantization noise and its reconstruction from* 20% *randomly selected Fourier coefficients using* (b) *total variation minimization and* (c) ℓ_1*-minimization of its bivariate Haar coefficients.*

Theorem 2. *Let $\boldsymbol{X} \in \mathbb{C}^{N\times N}$ be an image with discrete gradient $\boldsymbol{\nabla X}$. Suppose we observe noisy measurements $\boldsymbol{y} = \mathcal{M}(\boldsymbol{X}) + \boldsymbol{\xi}$ constructed from a matrix satisfying the RIP of order s, with noise level $\|\boldsymbol{\xi}\|_2 \le \varepsilon$. Then the solution*

$$\hat{\boldsymbol{X}} = \operatorname*{argmin}_{\boldsymbol{Z}} \|\boldsymbol{Z}\|_{TV} \quad \textit{such that} \quad \|\mathcal{M}(\boldsymbol{Z}) - \boldsymbol{y}\|_2 \le \varepsilon \tag{10}$$

satisfies

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2 \le C \log\left(\frac{N^2}{s}\right)\left(\frac{\|\boldsymbol{\nabla X} - (\boldsymbol{\nabla X})_s\|_1}{\sqrt{s}} + \varepsilon\right). \tag{11}$$

This error bound is optimal up to the logarithmic factor, as we discuss below. To the best of our knowledge, this is the first near-optimal result on stable image recovery by total variation minimization. For details about the construction of the measurements, see Theorem 5 and the remarks following.

1.3. Previous theory for total variation minimization. The last few years have witnessed numerous algorithmic advances that allow the efficient implementation of total variation minimization (TV). The recent split Bregman algorithm proposed by [26], based on the Bregman distance [5], is very efficient. Several algorithms are designed to exploit the structure of

Fourier measurements for further speed-up; see for example [52, 3]. Image reconstruction via independent minimization of the partial derivatives $\boldsymbol{X}_x$ and $\boldsymbol{X}_y$ was observed in [19] to give superior empirical results.

With respect to theory, it was shown in [14] that if an image $\boldsymbol{X}$ has an exactly sparse gradient, then (TV) recovers the image exactly from a small number of partial Fourier measurements. Moreover, using that the discrete Fourier transform commutes with the discrete gradient operator, one may change coordinates in this case and recast (TV) as an ℓ_1 program (L_1) with respect to the discrete gradient image [44] to derive stable *gradient* recovery results. In this paper, we extend these Fourier-specific stable gradient recovery results to general RIP measurement ensembles.

However, *robust recovery of the gradient need not imply robust recovery of the image itself.* To see this, suppose the error $\boldsymbol{\nabla X} - \boldsymbol{\nabla}\hat{\boldsymbol{X}}$ in the recovery of the gradient has a single nonzero component, of size α, located at pixel $(1, 1)$. That is, the gradient is recovered perfectly except at one pixel location, namely the upper left corner. Then based on this alone, it is possible that every pixel in $\hat{\boldsymbol{X}}$ differs from that in $\boldsymbol{X}$ by the amount α. This accumulation of error means that even when the reconstructed gradient is close to the gradient of $\boldsymbol{X}$, the images $\hat{\boldsymbol{X}}$ and $\boldsymbol{X}$ may be drastically different, magnified by a factor of N^2! Even for mean-zero images, the error may be magnified by a factor of N, as for images $\boldsymbol{X}$ with pixels $X_{j,k} = j$. We show that due to properties of the null space of RIP matrices, the (TV) reconstruction error $\boldsymbol{X} - \hat{\boldsymbol{X}}$ in Theorem 2 cannot propagate as such.

Recent work in [38] presents an *analysis co-sparse model* which considers signals sparse in the analysis domain. A series of theoretical and numerical tools are developed to solve the analysis problem (BL_1) in a general framework. In particular, the analysis operator may be the finite difference operator, which concatenates the vertical and horizontal derivatives into a single vector and is thus closely linked with the total variation operator. Effective pursuit methods are also proposed to solve such problems under the analysis co-sparse prior assumption. We refer the reader to [38] for details.

Finally, we note that our robustness recovery results for (TV) are specific to two-dimensional images, as the embedding theorems we rely on do not hold for one-dimensional arrays. Thus, our results do not imply robust recovery for one-dimensional piecewise-constant signals. Robustness for the recovery of the gradient support for piecewise-constant signals was studied in [51]. While the results for higher dimensional signals, $\boldsymbol{X} \in \mathbb{C}^{N^d}$ for $d > 2$, also do not immediately follow from the results in this article, we have recently extended these results to higher dimensional signals [41].

1.4. Organization.

The paper is organized as follows. Section 2 contains the statement of our main results about robust total variation recovery. The proof of our main results will occupy most of the remainder of the paper. We first prove robust recovery of the image gradient in section 3. In section 4 we derive a strong Sobolev inequality for discrete images lying in the null space of an RIP matrix which will bound the image recovery error by its total variation. Our result relies on a result by Cohen et al. [19] that the compressibility of the bivariate Haar wavelet transform is controlled by the total variation of an image. We prove Theorem 2 by way of Theorem 5 in section 4.1. We prove Theorem 6 in section 5, showing that the logarithmic factor of Theorem 2 can be removed by taking slightly more measurements.

We conclude in section 6 with some brief discussion. Proofs of intermediate propositions are included in the appendix.

2. Main results.

Our main results use the following proposition which generalizes the results used implicitly in the recovery of sparse signals using ℓ_1-minimization. It allows us to bound the norm of an entire signal when the signal (a) is close to the null space of an RIP matrix and (b) obeys an ℓ_1 cone constraint. In particular, (13) is just a generalization of results in [14], while (14) follows from (13) and the cone constraint (12). The proof of Proposition 3 is contained in the appendix.

Proposition 3. *Fix parameters $\gamma \geq 1$ and $\delta < 1/3$. Suppose that $\mathcal{A}$ satisfies the RIP of order $5k\gamma^2$ and level δ, and suppose that the image $\boldsymbol{D}$ satisfies a tube constraint*

$$\|\mathcal{A}(\boldsymbol{D})\|_2 \lesssim \varepsilon.$$

Suppose further that for a subset S of cardinality $|S| \leq k$, $\boldsymbol{D}$ satisfies the cone constraint

$$\|\boldsymbol{D}_{S^c}\|_1 \leq \gamma\|\boldsymbol{D}_S\|_1 + \sigma. \tag{12}$$

Then

$$\|\boldsymbol{D}\|_2 \lesssim \frac{\sigma}{\gamma\sqrt{k}} + \varepsilon \tag{13}$$

and

$$\|\boldsymbol{D}\|_1 \lesssim \sigma + \gamma\sqrt{k}\varepsilon. \tag{14}$$

Neither the RIP level of $5k\gamma^2$ nor the restricted isometry constant $\delta < 1/3$ are sharp; for instance, an RIP level of $2s$ and restricted isometry constant $\delta_{2s} \approx .4931$ are sufficient for Proposition 3 with $\gamma = 1$ [37, 7].

Our main results show robust recovery of images via the total variation minimization program (TV). For simplicity of presentation, we say that a linear operator $\mathcal{A} : \mathbb{C}^{N_1\times N_2} \to \mathbb{C}^m$ has the RIP of order s and level $\delta \in (0,1)$ if

$$(1-\delta)\|\boldsymbol{X}\|_2^2 \leq \|\mathcal{A}(\boldsymbol{X})\|_2^2 \leq (1+\delta)\|\boldsymbol{X}\|_2^2 \qquad \forall\, s\text{-sparse } \boldsymbol{X} \in \mathbb{C}^{N_1\times N_2}. \tag{15}$$

Here and throughout, $\|\boldsymbol{X}\|_p = \left(\sum_{j,k} |\boldsymbol{X}_{j,k}|^p\right)^{1/p}$ denotes the entrywise ℓ_p-norm of the image $\boldsymbol{X}$, treating the image as a vector. In particular, $p = 2$ is the Frobenius norm

$$\|\boldsymbol{X}\|_2 = \sqrt{\sum_{j,k} |X_{j,k}|^2} = \sqrt{\mathrm{tr}(\boldsymbol{X}\boldsymbol{X}^*)}.$$

This norm is generated by the image inner product

$$\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \mathrm{trace}(\boldsymbol{X}\boldsymbol{Y}^*). \tag{16}$$

Note that if the linear operator $\mathcal{A}$ is given by

$$(\mathcal{A}(\boldsymbol{X}))_j = \langle \boldsymbol{A_j}, \boldsymbol{X}\rangle,$$

then $\mathcal{A}$ satisfies this RIP precisely when the matrix whose rows consist of $\boldsymbol{A_j}$ unraveled into vectors satisfies the standard RIP as defined in (1). Thus there is clearly a one-to-one correspondence between the RIP for linear operators $\mathcal{A} : \mathbb{C}^{N_1\times N_2} \to \mathbb{C}^m$ and the RIP for matrices $\boldsymbol{\Phi} \in \mathbb{C}^{m\times(N_1N_2)}$, and we treat these notions as equivalent.

Since we are considering images $\boldsymbol{X} \in \mathbb{C}^{N\times N}$ rather than vectors, it will be helpful to first determine what form an optimal error recovery bound takes in the setting of images. In standard CS, the optimal minimax error rate from $m \gtrsim s\log(N^2/s)$ nonadaptive linear measurements is

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \le C\left(\frac{\|\boldsymbol{x} - \boldsymbol{x_s}\|_1}{\sqrt{s}} + \varepsilon\right). \tag{17}$$

In the setting of images, this implies that the best possible error rate from $m \gtrsim s\log(N^2/s)$ linear measurements is at best

$$\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_2 \le C\left(\frac{\|\boldsymbol{\nabla X} - (\boldsymbol{\nabla X})_s\|_1}{\sqrt{s}} + \varepsilon\right). \tag{18}$$

Above, $(\boldsymbol{\nabla X})_s$ is the best s-sparse approximation to the discrete gradient $\boldsymbol{\nabla X}$. To see that we could not possibly hope for a better error rate, observe that if we could, we would reach a contradiction in light of the norm of the discrete gradient operator: $\|\boldsymbol{\nabla Z}\|_2 \le 4\|\boldsymbol{Z}\|_2$.

Theorem 5 guarantees a recovery error proportional to (18) up to a single logarithmic factor $\log(N^2/s)$. That is, the recovery error of Theorem 5 is optimal up to at most a logarithmic factor. We see in Theorem 6 that by taking more measurements, we obtain the optimal recovery error, without the logarithmic term. Here and throughout we use the notation $u \gtrsim v$ to indicate that there exists some absolute constant $C > 0$ such that $u \ge Cv$. We use the notation $u \lesssim v$ accordingly. In this paper, $C > 0$ will always denote a universal constant that might be different in each occurrence.

To change coordinates from the image domain to the gradient domain, it will be useful for us to consider matrices $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Phi}^0$ obtained from a matrix $\boldsymbol{\Phi}$ by concatenating a row of zeros to the bottom and top of $\boldsymbol{\Phi}$, respectively. Concretely, for a matrix $\boldsymbol{\Phi} \in \mathbb{C}^{(N-1)\times N}$, we denote by $\boldsymbol{\Phi}^0 \in \mathbb{C}^{N\times N}$ the augmented matrix $\boldsymbol{\Phi}^0$ with entries

$$(\boldsymbol{\Phi^0})_{j,k} = \begin{cases} 0, & j = 1, \\ \Phi_{j-1,k}, & 2 \le j \le N. \end{cases} \tag{19}$$

We denote similarly by $\boldsymbol{\Phi_0}$ the matrix resulting by adding an additional row of zeros to the bottom of $\boldsymbol{\Phi}$.

We can relate measurements using the padded matrices (19) of the entire image to measurements of its directional gradients, as defined in (4). This relation can be verified by direct algebraic manipulation, and so the proof is omitted.

Lemma 4. *Given $\boldsymbol{X} \in \mathbb{C}^{N\times N}$ and $\boldsymbol{\Phi} \in \mathbb{C}^{(N-1)\times N}$,*

$$\langle \boldsymbol{\Phi}, \boldsymbol{X}_x \rangle = \left\langle \boldsymbol{\Phi}^0, \boldsymbol{X} \right\rangle - \langle \boldsymbol{\Phi}_0, \boldsymbol{X} \rangle$$

and

$$\left\langle \boldsymbol{\Phi}, \boldsymbol{X}_y^T \right\rangle = \left\langle \boldsymbol{\Phi}^0, \boldsymbol{X}^T \right\rangle - \left\langle \boldsymbol{\Phi}_0, \boldsymbol{X}^T \right\rangle,$$

where $\boldsymbol{X}^T$ denotes the (nonconjugate) transpose of the matrix $\boldsymbol{X}$.

For a linear operator $\mathcal{A} : \mathbb{C}^{(N-1)\times N} \to \mathbb{C}^m$ with component measurements $\mathcal{A}(\boldsymbol{X})_j = \langle \boldsymbol{A}_j, \boldsymbol{X}\rangle$ we denote by $\mathcal{A}^0 : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ the linear operator with components $[\mathcal{A}^0(\boldsymbol{X})]_j = \big\langle (\boldsymbol{A}^0)_j, \boldsymbol{X}\big\rangle$. We define $\mathcal{A}_0 : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ similarly.

We are now prepared to state our main results which guarantee stable image recovery by total variation minimization using RIP measurements.

Theorem 5. *Let $N = 2^n$ be a power of two. Let $\mathcal{A} : \mathbb{C}^{(N-1)\times N} \to \mathbb{C}^{m_1}$ and $\mathcal{A}' : \mathbb{C}^{(N-1)\times N} \to \mathbb{C}^{m_1}$ be such that the concatenated operator $[\mathcal{A}\,\mathcal{A}']$ has the RIP of order $5s$ and level $\delta < 1/3$. Recall the bivariate Haar transform $\mathcal{H} : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ as defined in (8), and let $\mathcal{B} : \mathbb{C}^{N\times N} \to \mathbb{C}^{m_2}$ be such that the composite operator $\mathcal{B}\mathcal{H}^{-1} : \mathbb{C}^{N\times N} \to \mathbb{C}^{m_2}$ has the RIP of order $2s$ and level $\delta < 1$.*

Let $m = 4m_1 + m_2$, and consider the linear operator $\mathcal{M}(\boldsymbol{X}) : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ with components

$$\mathcal{M}(\boldsymbol{X}) = \Big(\mathcal{A}^0(\boldsymbol{X}), \mathcal{A}_0(\boldsymbol{X}), \mathcal{A}'^0(\boldsymbol{X}^T), \mathcal{A}'_0(\boldsymbol{X}^T), \mathcal{B}(\boldsymbol{X})\Big). \tag{20}$$

If $\boldsymbol{X} \in \mathbb{C}^{N\times N}$ has discrete gradient $\boldsymbol{\nabla X}$ and noisy measurements $\boldsymbol{y} = \mathcal{M}(\boldsymbol{X}) + \boldsymbol{\xi}$ are observed with noise level $\|\boldsymbol{\xi}\|_2 \le \varepsilon$, then

$$\hat{\boldsymbol{X}} = \operatorname*{argmin}_{\boldsymbol{Z}} \|\boldsymbol{Z}\|_{TV} \quad \textit{such that} \quad \|\mathcal{M}(\boldsymbol{Z}) - \boldsymbol{y}\|_2 \le \varepsilon \tag{21}$$

satisfies

$$\|\boldsymbol{\nabla X} - \boldsymbol{\nabla}\hat{\boldsymbol{X}}\|_2 \lesssim \frac{\|\boldsymbol{\nabla X} - (\boldsymbol{\nabla X})_s\|_1}{\sqrt{s}} + \varepsilon, \tag{22}$$

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_{TV} \lesssim \|\boldsymbol{\nabla X} - (\boldsymbol{\nabla X})_s\|_1 + \sqrt{s}\varepsilon, \tag{23}$$

and

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2 \lesssim \log\left(\frac{N^2}{s}\right)\left(\frac{\|\boldsymbol{\nabla X} - (\boldsymbol{\nabla X})_s\|_1}{\sqrt{s}} + \varepsilon\right). \tag{24}$$

To the best of our knowledge, Theorem 5 provides the first provable guarantee of robust recovery for images from compressed measurements via total variation minimization. Since we require the RIP only on the measurements generating the operator $\mathcal{M}$, Theorem 5 implies Theorem 2.

Our second main result shows that, by allowing for more measurements, one obtains stable and robust recovery guarantees as in Theorem 5 but without the additional log factor. Moreover, the following theorem holds for general sensing matrices such that, composed with the inverse bivariate Haar transform, they have the RIP.

Theorem 6. *Let $N = 2^n$ be a power of two. Let $\mathcal{H}$ be the bivariate Haar transform, and let $\mathcal{A} : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ be such that the composite operator $\mathcal{A}\mathcal{H}^{-1} : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ has the RIP of order $Cs\log^3(N)$ and level $\delta < 1/3$. Then the following holds for any $\boldsymbol{X} \in \mathbb{C}^{N\times N}$: if noisy measurements $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}) + \boldsymbol{\xi}$ are observed with noise level $\|\boldsymbol{\xi}\|_2 \le \varepsilon$, then*

$$\hat{\boldsymbol{X}} = \operatorname*{argmin}_{\boldsymbol{Z}} \|\boldsymbol{Z}\|_{TV} \quad \textit{such that} \quad \|\mathcal{A}(\boldsymbol{Z}) - \boldsymbol{y}\|_2 \le \varepsilon$$

satisfies

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2 \lesssim \frac{\|\nabla \boldsymbol{X} - (\nabla \boldsymbol{X})_s\|_1}{\sqrt{s}} + \varepsilon. \tag{25}$$

Remark 1. In light of (18), the gradient error guarantees (22) and (23) provided by Theorem 5 are optimal, and the image error guarantee (24) is optimal up to a logarithmic factor, which we conjecture to be an artifact of the proof which relies on the Haar wavelet transform. We also believe that the $4m_1$ measurements derived from $\mathcal{A}$ in the theorem, which are used only to prove stable gradient recovery, are not necessary and can be removed. Theorem 6 provides optimal error recovery guarantees, at the expense of requiring an additional factor of $\log^3(N)$ measurements.

Remark 2. The RIP requirements in Theorem 5 mean that the linear measurements can be generated from standard RIP matrix ensembles that are incoherent with the Haar wavelet basis. For example, they can be generated from a sub-Gaussian random matrix $\boldsymbol{\Phi} \in \mathbb{R}^{m \times N^2}$ with $m \gtrsim s \log(N^2/s)$ or a partial Fourier matrix $\boldsymbol{F}_\Omega \in \mathbb{C}^{m \times N^2}$ with $m \gtrsim s \log^5(N)$ and randomized column signs [30]. We note that without randomized column signs, the partial Fourier matrix with uniformly subsampled rows is not incoherent with wavelet bases. As recently shown in [31], the partial Fourier matrix with rows subsampled according to appropriate power law densities, after preconditioning, is incoherent with the Haar wavelet basis and does apply in Theorems 5 and 6.

Remark 3. The constant C in Theorem 6 is an absolute constant proportional to the constant obtained in Proposition 8 below. We have not tried to optimize the dependence on the values of the restricted isometry parameters in the theorems. Refinements such as those in standard CS may yield improvements in the conditions.

Remark 4. Theorems 5 and 6 require the image side-length to be a power of 2, $N = 2^n$. This is not actually a restriction, as an image of arbitrary side-length $N \in \mathbb{N}$ can be reflected horizontally and vertically to produce an at most $2N \times 2N$ image with the same total variation up to a factor of 4.

The remainder of this paper is dedicated to the proofs of Theorems 5 and 6. The proof of Theorem 5 has two parts: we first prove the bounds (22) and (23) concerning stable recovery of the discrete gradient. We then prove a strengthened Sobolev inequality for images in the null space of an RIP matrix, and stable image recovery follows. The proof of Theorem 6 is similar, but more direct, and does not use a Sobolev inequality explicitly.

3. Stable gradient recovery for discrete images.

In this section we prove statements (22) and (23) from Theorem 5, showing that total variation minimization recovers the gradient image robustly.

3.1. Proof of stable gradient recovery, bounds (22) and (23).

Since $[\mathcal{A} \ \ \mathcal{A}']$ satisfies the RIP, in light of Proposition 3, it suffices to show that the discrete gradient $\nabla(\boldsymbol{X} - \hat{\boldsymbol{X}})$, regarded as a vector, satisfies the tube and cone constraints.

Let $\boldsymbol{D} = \boldsymbol{X} - \hat{\boldsymbol{X}}$, and set $\boldsymbol{L} = (\boldsymbol{D}_x, \boldsymbol{D}_y^T)$. For convenience, let P denote the mapping of indices which maps the index of a nonzero entry in $\nabla \boldsymbol{D}$ to its corresponding index in $\boldsymbol{L}$. Observe that by definition of the gradient, $\boldsymbol{L}$ has the same norm as $\nabla \boldsymbol{D}$. That is,

$\|\boldsymbol{L}\|_2 = \|\boldsymbol{\nabla D}\|_2$ and $\|\boldsymbol{L}\|_1 = \|\boldsymbol{\nabla D}\|_1$. It thus now suffices to show that the matrix $\boldsymbol{L}$ satisfies the tube and cone constraints.

Let $\boldsymbol{A_1}, \boldsymbol{A_2}, \dots \boldsymbol{A_{m_1}}, \boldsymbol{A'_1}, \boldsymbol{A'_2}, \dots \boldsymbol{A'_{m_1}}$ be such that

$$\mathcal{A}(\boldsymbol{Z})_j = \langle \boldsymbol{A_j}, \boldsymbol{Z}\rangle, \quad \mathcal{A}'(\boldsymbol{Z})_j = \langle \boldsymbol{A'_j}, \boldsymbol{Z}\rangle.$$

Cone constraint. The cone constraint holds by minimality of $\hat{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{D}$. Indeed, by this and the fact that $\boldsymbol{X}$ is also a feasible solution, letting S denote the support of the largest s entries of $\boldsymbol{\nabla X}$, we have

$$\begin{aligned}
\|(\boldsymbol{\nabla X})_S\|_1 - \|(\boldsymbol{\nabla D})_S\|_1 - \|(\boldsymbol{\nabla X})_{S^c}\|_1 + \|(\boldsymbol{\nabla D})_{S^c}\|_1 \\
&\le \|(\boldsymbol{\nabla X})_S - (\boldsymbol{\nabla D})_S\|_1 + \|(\boldsymbol{\nabla X})_{S^c} - (\boldsymbol{\nabla D})_{S^c}\|_1 \\
&= \|\boldsymbol{\nabla}\hat{\boldsymbol{X}}\|_1 \\
&\le \|\boldsymbol{\nabla X}\|_1 \\
&= \|(\boldsymbol{\nabla X})_S\|_1 + \|(\boldsymbol{\nabla X})_{S^c}\|_1.
\end{aligned}$$

Rearranging, this yields

$$\|(\boldsymbol{\nabla D})_{S^c}\|_1 \le \|(\boldsymbol{\nabla D})_S\|_1 + 2\|\boldsymbol{\nabla X} - (\boldsymbol{\nabla X})_s\|_1.$$

Since $\boldsymbol{L}$ contains all the same nonzero entries as $\boldsymbol{\nabla D}$, this implies that $\boldsymbol{L}$ satisfies the cone constraint

$$\|\boldsymbol{L}_{P(S)^c}\|_1 \le \|(\boldsymbol{\nabla D})_{P(S)}\|_1 + 2\|\boldsymbol{\nabla X} - (\boldsymbol{\nabla X})_s\|_1.$$

By definition of P, note that $|P(S)| \le |S| = s$.

Tube constraint. First note that $\boldsymbol{D}$ satisfies a tube constraint,

$$\begin{aligned}
\|\mathcal{M}(\boldsymbol{D})\|_2^2 &\le 2\|\mathcal{M}(\boldsymbol{X}) - \boldsymbol{y}\|_2^2 + 2\|\mathcal{M}(\hat{\boldsymbol{X}}) - \boldsymbol{y}\|_2^2 \\
&\le 4\varepsilon^2.
\end{aligned}$$

Now by Lemma 4,

$$\begin{aligned}
|\langle \boldsymbol{A}_j, \boldsymbol{D}_x\rangle|^2 &= \left|\left\langle [\boldsymbol{A_j}]^0, \boldsymbol{D}\right\rangle - \langle [\boldsymbol{A_j}]_0, \boldsymbol{D}\rangle\right|^2 \\
&\le 2\left|\left\langle [\boldsymbol{A_j}]^0, \boldsymbol{D}\right\rangle\right|^2 + 2|\langle [\boldsymbol{A_j}]_0, \boldsymbol{D}\rangle|^2
\end{aligned} \tag{26}$$

and

$$\begin{aligned}
\left|\left\langle \boldsymbol{A}'_j, \boldsymbol{D}_y^T\right\rangle\right|^2 &= \left|\left\langle [\boldsymbol{A'_j}]^0, \boldsymbol{D}^T\right\rangle - \left\langle [\boldsymbol{A'_j}]_0, \boldsymbol{D}^T\right\rangle\right|^2 \\
&\le 2\left|\left\langle [\boldsymbol{A'_j}]^0, \boldsymbol{D}^T\right\rangle\right|^2 + 2\left|\left\langle [\boldsymbol{A'_j}]_0, \boldsymbol{D}^T\right\rangle\right|^2.
\end{aligned} \tag{27}$$

Thus $\boldsymbol{L}$ also satisfies a tube constraint:

$$\begin{aligned}
\|[\mathcal{A}\ \mathcal{A}'](\boldsymbol{L})\|_2^2 &= \sum_{j=1}^{m} |\langle \boldsymbol{A}_j, \boldsymbol{D}_x\rangle|^2 + \left|\left\langle \boldsymbol{A}'_j, \boldsymbol{D}_y^T\right\rangle\right|^2 \\
&\le 2\|\mathcal{M}(\boldsymbol{D})\|_2^2 \\
&\le 8\varepsilon^2.
\end{aligned} \tag{28}$$

Proposition 3 then completes the proof. ■

4. A strengthened Sobolev inequality for incoherent null spaces.

As a corollary of the classical Sobolev embedding of the space of functions of bounded variation $BV(\mathbb{R}^2)$ into $L_2(\mathbb{R}^2)$ [1], the Frobenius norm of a mean-zero image is bounded by its total variation seminorm.

Proposition 7 (Sobolev inequality for images). *Let $\boldsymbol{X} \in \mathbb{C}^{N\times N}$ be a mean-zero image. Then*

$$\|\boldsymbol{X}\|_2 \le \|\boldsymbol{X}\|_{TV}. \tag{29}$$

This inequality also holds if, instead of being mean-zero, $\boldsymbol{X} \in \mathbb{C}^{N\times N}$ contains some zero-valued pixel. In the appendix, we give a direct proof of the Sobolev inequality (29) in the case that all pixels in the first column and first row of $\boldsymbol{X} \in \mathbb{C}^{N\times N}$ are zero-valued, $X_{1,j} = X_{j,1} = 0$.

In light of the total variation error estimate (23), the Sobolev inequality allows a preliminary estimate for the image error, assuming it is mean-zero:

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2 \le \|\boldsymbol{\nabla X} - (\boldsymbol{\nabla X})_S\|_1 + \sqrt{s}\varepsilon.$$

We will be able to derive a sharper bound on the error by looking to a deep and nontrivial theorem from [19] which says that the bivariate Haar coefficient vector of a function $f \in BV(Q)$ on the unit square $Q = [0,1)^2$ is in *weak* ℓ_1, and its weak ℓ_1-norm is proportional to its bounded variation seminorm. As a corollary of that result, we can bound the magnitude of the kth largest bivariate Haar coefficient of an image by the total variation of the image.

Proposition 8. *Suppose $\boldsymbol{X} \in \mathbb{C}^{N\times N}$ is mean-zero, and let $c_{(k)}(\boldsymbol{X})$ be the bivariate Haar coefficient of $\boldsymbol{X}$ having kth largest magnitude, or the entry of the bivariate Haar transform $\mathcal{H}(\boldsymbol{X})$ having kth largest magnitude. Then for all $k \ge 1$,*

$$|c_{(k)}(\boldsymbol{X})| \le C\frac{\|\boldsymbol{X}\|_{TV}}{k}.$$

The derivation of Proposition 8 from Theorem 8.1 of [19] is provided in the appendix.

Proposition 8 bounds the decay of the Haar wavelet coefficients by the image total variation seminorm. At the same time, vectors lying in the null space of a matrix with the RIP must be sufficiently *flat*, with the ℓ_2-energy in their largest s components in magnitude bounded by the ℓ_1-norm of the remaining components (the so-called *null-space property*) [18]. As a result, the norm of the bivariate Haar transform of the error, and thus the norm of the error itself, must be sufficiently small. Specifically, *the error $\boldsymbol{X} - \hat{\boldsymbol{X}}$ satisfies a Sobolev inequality that is stronger than the standard inequality* (29) *by a factor of* $\log(N^2/s)/\sqrt{s}$.

Theorem 9 (strong Sobolev inequality). *Let $\mathcal{B} : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ be a linear map such that $\mathcal{B}\mathcal{H}^{-1} : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ has the RIP of order $2s+1$ and level $\delta < 1$, where $\mathcal{H} : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ is the bivariate Haar transform. Suppose that $\boldsymbol{D} \in \mathbb{C}^{N\times N}$ satisfies the tube constraint $\|\mathcal{B}(\boldsymbol{D})\|_2 \le \varepsilon$. Then*

$$\|\boldsymbol{D}\|_2 \lesssim \left(\frac{\|\boldsymbol{D}\|_{TV}}{\sqrt{s}}\right)\log\left(\frac{N^2}{s}\right) + \varepsilon. \tag{30}$$

Proof. Let $\boldsymbol{Y} = \mathcal{H}(\boldsymbol{D}) \in \mathbb{C}^{N\times N}$ be the bivariate Haar transform of $\boldsymbol{D}$, and let $c_{(j)} \in \mathbb{C}$ be the jth largest entry (pixel) of $\boldsymbol{Y}$ in absolute magnitude. Note that $\boldsymbol{D}$ can be decomposed orthogonally into a mean-zero image and a constant image, and that the Haar transform of a constant image is one-sparse. Thus, by assuming the RIP of order $2s+1$ as opposed to order $2s$, we can assume without loss of generality that $\boldsymbol{D}$ is itself mean-zero.

Decompose $\boldsymbol{Y} = \boldsymbol{Y}_S + \boldsymbol{Y}_{S^c}$, where $\boldsymbol{Y}_S$ is the s-sparse image consisting of the s largest-magnitude entries of $\boldsymbol{Y}$. Write $\boldsymbol{Y}_{S^c} = \boldsymbol{Y}^{(1)} + \boldsymbol{Y}^{(2)} + \cdots + \boldsymbol{Y}^{(r)}$, where $r = \lfloor \frac{N^2}{s} \rfloor$, where $\boldsymbol{Y}^{(1)}$ is the s-sparse image consisting of the s largest-magnitude entries of $\boldsymbol{Y}_{S^c}$, and so on.

By Proposition 8, we know that $|c_{(j)}| \le C\|\boldsymbol{D}\|_{TV}/j$. Then

$$\begin{aligned}\|\boldsymbol{Y}_{S^c}\|_1 &= \sum_{j=s+1}^{N^2} |c_{(j)}| \\ &\le C\|\boldsymbol{D}\|_{TV} \sum_{j=s+1}^{N^2} \frac{1}{j} \\ &\le C\|\boldsymbol{D}\|_{TV} \log\left(\frac{N^2}{s}\right), \qquad (31)\end{aligned}$$

where the second inequality follows from properties of the geometric summation. We can similarly bound the ℓ_2-norm of the residual image:

$$\begin{aligned}\|\boldsymbol{Y}_{S^c}\|_2^2 &= \sum_{j=s+1}^{N^2} |c_{(j)}|^2 \\ &\le C(\|\boldsymbol{D}\|_{TV})^2 \sum_{j=s+1}^{N^2} \frac{1}{j^2} \\ &\le C(\|\boldsymbol{D}\|_{TV})^2/s, \qquad (32)\end{aligned}$$

obtaining $\|\boldsymbol{Y}_{S^c}\|_2 \le C\|\boldsymbol{D}\|_{TV}/\sqrt{s}$.

We now use the assumed tube constraint for $\boldsymbol{D}$ and RIP for $\mathcal{B}\mathcal{H}^{-1}$,

$$\begin{aligned}\varepsilon &\ge \|\mathcal{B}(\boldsymbol{D})\|_2 \\ &\ge \|\mathcal{B}\mathcal{H}^{-1}(\boldsymbol{Y}_S + \boldsymbol{Y}^{(1)})\|_2 - \sum_{j=2}^{r} \|\mathcal{B}\mathcal{H}^{-1}(\boldsymbol{Y}^{(j)})\|_2 \\ &\ge (1-\delta)\|\boldsymbol{Y}_S + \boldsymbol{Y}^{(1)}\|_2 - (1+\delta)\sum_{j=2}^{r} \|\boldsymbol{Y}^{(j)}\|_2 \\ &\ge (1-\delta)\|\boldsymbol{Y}_S\|_2 - (1+\delta)\sum_{j=2}^{r} \|\boldsymbol{Y}^{(j)}\|_2 \\ &\ge (1-\delta)\|\boldsymbol{Y}_S\|_2 - (1+\delta)\frac{1}{\sqrt{s}}\sum_{j=1}^{r} \|\boldsymbol{Y}^{(j)}\|_1 \\ &= (1-\delta)\|\boldsymbol{Y}_S\|_2 - (1+\delta)\frac{1}{\sqrt{s}}\|\boldsymbol{Y}_{S^c}\|_1. \qquad (33)\end{aligned}$$

The last inequality applies the blockwise bound $\|\boldsymbol{Y}^{(j)}\|_2 \leq \frac{\|\boldsymbol{Y}^{(j-1)}\|_1}{\sqrt{s}}$, which holds because the magnitude of each component of $\boldsymbol{Y}^{(j-1)}$ is larger than the average magnitude of the components of $\boldsymbol{Y}^{(j)}$.

Combined with the ℓ_1-norm bound (31) on $\boldsymbol{Y}_{S^c}$, this gives

$$\|\boldsymbol{Y}_S\|_2 \lesssim \varepsilon + \log\left(\frac{N^2}{s}\right)\left(\frac{\|\boldsymbol{D}\|_{TV}}{\sqrt{s}}\right). \tag{34}$$

This bound together with the ℓ_2-tail bound (32) gives

$$\|\boldsymbol{D}\|_2 = \|\boldsymbol{Y}\|_2 \leq \|\boldsymbol{Y}_S\|_2 + \|\boldsymbol{Y}_{S^c}\|_2 \lesssim \varepsilon + \log\left(\frac{N^2}{s}\right)\left(\frac{\|\boldsymbol{D}\|_{TV}}{\sqrt{s}}\right), \tag{35}$$

where the first equality follows by orthonormality of the Haar transform. This completes the proof. ■

4.1. Proof of Theorem 5.

We now have all the ingredients to prove our main result, Theorem 5.

Proof. Since bounds (22) and (23) were already proved in section 3.1, it remains to prove the final stability bound (24). Since the measurements of $\boldsymbol{X}$ are of the form (20), the image error $\boldsymbol{D} = \boldsymbol{X} - \hat{\boldsymbol{X}}$ satisfies the tube-constraint $\|\mathcal{B}\boldsymbol{D}\|_2 \leq \varepsilon$. Thus we may apply Theorem 9 and then the total variation bound (23) on $\boldsymbol{D}$ which shows

$$\begin{aligned}
\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2 &= \|\boldsymbol{D}\|_2 \\
&\lesssim \varepsilon + \log\left(\frac{N^2}{s}\right)\left(\frac{\|\boldsymbol{D}\|_{TV}}{\sqrt{s}}\right) \\
&\lesssim \varepsilon + \log\left(\frac{N^2}{s}\right)\left(\frac{\|\nabla\boldsymbol{X} - (\nabla\boldsymbol{X})_s\|_1 + \sqrt{s}\varepsilon}{\sqrt{s}}\right) \\
&\lesssim \log\left(\frac{N^2}{s}\right)\left(\frac{\|\nabla\boldsymbol{X} - (\nabla\boldsymbol{X})_s\|_1}{\sqrt{s}} + \varepsilon\right).
\end{aligned}$$

This completes the proof of Theorem 5. ■

5. Proof of Theorem 6.

We will use the following two basic lemmas about the bivariate Haar system.

Lemma 10. *Let $N = 2^n$. For any indices (j,k) and $(j,k+1)$, there are at most $6n$ bivariate Haar wavelets which are not constant on these indices.*

Proof. The lemma follows by showing that for fixed dyadic scale p between 1 and n, there are at most six Haar wavelets with side dimension 2^{n-p} which are not constant on these two indices. Indeed, if the edge between (j,k) and $(j,k+1)$ coincides with a dyadic edge at scale p, then the three wavelets supported on each of the two adjacent dyadic squares transition from being zero to nonzero along this edge. The only other case to consider is that (j,k) coincides with a dyadic edge at dyadic scale $p+1$ but does not coincide with a dyadic edge at scale p; in this case the three wavelets supported on the dyadic square centered at $(j,k+1),(j,k)$ can change from negative to positive value. ■

Lemma 11. *The bivariate Haar wavelets satisfy*

$$\|\nabla \boldsymbol{h}_{j,k}^e\|_1 \leq 8 \qquad \forall\, j,k,e.$$

Proof. The wavelet $\boldsymbol{h}_{j,k}^e$ is supported on a dyadic square of side-length 2^{n-j}, and it has constant magnitude on its support $|h_{j,k}^e| = 2^{j-n}$. Thus at the four boundary edges of the square, there is a jump of 2^{j-n}, and at the (at most two) dyadic edges in the middle of the square where the sign changes there is a jump of $2 \cdot 2^{j-n}$. Then $\|\nabla \boldsymbol{h}_{j,k}^e\|_1 \leq 8 \cdot 2^{n-j} \cdot 2^{j-n} = 8$. ■

We are now in a position to prove Theorem 6.

Proof of Theorem 6. Let $\boldsymbol{D} = \boldsymbol{X} - \hat{\boldsymbol{X}}$ denote the residual error by (TV), which we may assume without loss is mean-zero. Let $\mathcal{H} : \mathbb{C}^{N^2} \to \mathbb{C}^{N^2}$ denote the unitary matrix that computes the bivariate Haar transform, and let $c_{(j)} = c_{(j)}(\boldsymbol{D})$ denote the jth largest-magnitude Haar coefficient of $\boldsymbol{D}$ among $\boldsymbol{c} = \mathcal{H}\boldsymbol{D}$. Let $\boldsymbol{h}_{(j)}$ denote the Haar wavelet associated to $c_{(j)}$. We assume that $\mathcal{A}\mathcal{H}^* = \mathcal{A}\mathcal{H}^{-1}$ has the RIP of order

$$\widetilde{s} = C'' s \log^3(N) \tag{36}$$

for a universal constant C'' derived below and level $\delta < 1/3$.

Cone constraint on $\boldsymbol{\nabla D}$. As shown in section 3.1, we have the cone constraint

$$\|(\boldsymbol{\nabla D})_{S^c}\|_1 \leq \|(\boldsymbol{\nabla D})_S\|_1 + 2\|\boldsymbol{\nabla X} - (\boldsymbol{\nabla X})_s\|_1. \tag{37}$$

Cone constraint on $\boldsymbol{c} = \boldsymbol{\mathcal{H}D}$. Proposition 8 allows us to pass from a cone constraint on the gradient to a cone constraint on the Haar transform of $\boldsymbol{D}$. Recall that S is the subset of s largest-magnitude entries of $\nabla\boldsymbol{D}$. By Lemma 10, the set Ω of wavelets which are nonconstant over S has cardinality at most $|\Omega| = 8s\log(N)$. Decompose $\boldsymbol{D}$ as

$$\boldsymbol{D} = \sum_j c_{(j)}\boldsymbol{h}_{(j)} = \sum_{j\in\Omega} c_{(j)}\boldsymbol{h}_{(j)} + \sum_{j\in\Omega^c} c_{(j)}\boldsymbol{h}_{(j)} =: \boldsymbol{D}_\Omega + \boldsymbol{D}_{\Omega^c}. \tag{38}$$

By linearity of the gradient, $\nabla\boldsymbol{D} = \nabla\boldsymbol{D}_\Omega + \nabla\boldsymbol{D}_{\Omega^c}$. Moreover, by construction of Ω, we have immediately that $(\nabla\boldsymbol{D}_{\Omega^c})_S = 0$, leaving the equality $(\nabla\boldsymbol{D})_S = (\nabla\boldsymbol{D}_\Omega)_S$. By Lemma 11 and the triangle inequality,

$$\begin{aligned} \|(\nabla\boldsymbol{D})_S\|_1 = \|(\nabla\boldsymbol{D}_\Omega)_S\|_1 &\leq \|\nabla\boldsymbol{D}_\Omega\|_1 \\ &\leq \sum_{j\in\Omega} |c_{(j)}|\|\nabla\boldsymbol{h}_{(j)}\|_1 \\ &\leq 8\sum_{j\in\Omega}|c_{(j)}|. \end{aligned} \tag{39}$$

Combining (39) with Proposition 8 concerning the decay of the wavelet coefficients and the cone constraint (37), and letting

$$k = 8s\log(N) = |\Omega|,$$

we arrive at a cone constraint on the wavelet coefficients:

$$\begin{aligned}\sum_{j=k+1}^{N^2} |c_{(j)}| \leq \sum_{j=s+1}^{N^2} |c_{(j)}| &\leq C \log\left(\frac{N^2}{s}\right) \|\nabla \boldsymbol{D}\|_1 \\ &= C \log\left(\frac{N^2}{s}\right)\left(\|(\nabla \boldsymbol{D})_S\|_1 + \|(\nabla \boldsymbol{D})_{S^c}\|_1\right) \\ &\leq C \log\left(\frac{N^2}{s}\right)\left(2\|(\nabla \boldsymbol{D})_S\|_1 + 2\|\nabla \boldsymbol{X} - (\nabla \boldsymbol{X})_S\|_1\right) \\ &\leq C \log\left(\frac{N^2}{s}\right)\left(16\sum_{j\in\Omega}|c_{(j)}| + 2\|\nabla \boldsymbol{X} - (\nabla \boldsymbol{X})_S\|_1\right) \\ &\leq C' \log\left(\frac{N^2}{s}\right)\left(\sum_{j=1}^{k}|c_{(j)}| + \|\nabla \boldsymbol{X} - (\nabla \boldsymbol{X})_S\|_1\right).\end{aligned}$$

Tube constraint $\|\mathcal{A}\mathcal{H}^*\boldsymbol{c}\|_2 \leq 2\varepsilon$. By assumption, $\mathcal{A}\mathcal{H}^* : \mathbb{C}^{N^2} \rightarrow \mathbb{C}^m$ has the RIP of order $\overline{s} = 8s\log^3(N) > k$. Since both $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are in the feasible region of (TV), we have for $\boldsymbol{c} = \mathcal{H}\boldsymbol{D} = \mathcal{H}\boldsymbol{X} - \mathcal{H}\hat{\boldsymbol{X}}$ and by the triangle inequality

$$\|\mathcal{A}\mathcal{H}^*\boldsymbol{c}\|_2 \leq \|\mathcal{A}\boldsymbol{X}\|_2 + \|\mathcal{A}\hat{\boldsymbol{X}}\|_2 \leq 2\varepsilon.$$

Using the derived cone and tube constraints on $\boldsymbol{c} = \mathcal{H}(\boldsymbol{X} - \hat{\boldsymbol{X}})$, along with the RIP bound on $\mathcal{A}\mathcal{H}^*$, the proof is complete by applying Proposition 3 using $\gamma = C'\log(N^2/s) \leq 2C'\log(N)$, $k = 8s\log N$, and $\sigma = \log(N^2/s)\|\nabla\boldsymbol{X} - (\nabla\boldsymbol{X})_S\|_1$. In fact, this is where we need that the RIP order is $\widetilde{s}$ in (36), to accommodate the factors γ and k in Proposition 3. ■

6. Conclusion.

Compressed sensing techniques provide reconstruction of compressible signals from a few linear measurements. A fundamental application is image compression and reconstruction. Since images are compressible with respect to wavelet bases, standard CS methods such as ℓ_1-minimization guarantee reconstruction to within a factor of the error of best s-term wavelet approximation. The story does not end here, though. Images are more compressible with respect to their discrete gradient representation, and indeed the advantages of total variation minimization over wavelet-coefficient minimization have been empirically well documented (see, e.g., [10, 11]). It had been well known that without measurement noise, images with perfectly sparse gradients are recovered exactly via total variation minimization [14]. Of course in practice, images do not have exactly sparse gradients, and measurements are corrupted with additive or quantization noise. To the best of our knowledge, our main results, Theorems 5 and 6, are the first to provably guarantee robust image recovery via total variation minimization. Analogously to the standard CS results, the number of measurements in Theorem 5 required for reconstruction is optimal, up to a single logarithmic factor in the image dimension. Theorem 5 has been extended to the multidimensional case for signals with higher dimensional structure such as movies [41]. On the other hand, the proof of Theorem 6 is specific to properties of the bivariate Haar system, and extending it to higher dimensions (as well as for $d = 1$) remains an open problem. Theorem 6 applies, for example, to partial Fourier matrices subsampled according to appropriate variable densities [31]. Finally, we believe our proof technique can be used for analysis operators beyond the total variation operator. For example, in practice one often finds that minimizing the sum of a total variation seminorm and wavelet norm gives better image reconstructions. We leave this and the study of more general analysis-type operators as future work.

Appendix. Proofs of lemmas and propositions.

A.1. Proof of Proposition 3.

Here we include a proof of Proposition 3, which is a modest generalization of results from [11].

Let $s = k\gamma^2$, and let $S \subset [N]$ be the support set of the best s-term approximation of $\boldsymbol{D}$.

Proof. By assumption, we suppose that $\boldsymbol{D}$ obeys the cone constraint

$$\|\boldsymbol{D}_{S^c}\|_1 \leq \gamma\|\boldsymbol{D}_S\|_1 + \sigma \tag{40}$$

and the tube constraint $\|\mathcal{A}(\boldsymbol{D})\|_2 \leq \varepsilon$.

We write $\boldsymbol{D}_{S^c} = \boldsymbol{D}_{S_1} + \boldsymbol{D}_{S_2} + \cdots + \boldsymbol{D}_{S_r}$, where $r = \lfloor \frac{N^2}{4s} \rfloor$. Here $\boldsymbol{D}_{S_1}$ consists of the $4s$ largest-magnitude components of $\boldsymbol{D}$ over S^c, $\boldsymbol{D}_{S_2}$ consists of the $4s$ largest-magnitude components of $\boldsymbol{D}$ over $S^c \setminus S_1$, and so on. Note that $\boldsymbol{D}_S$ and similar expressions below can have the meaning of both restricting $\boldsymbol{D}$ to the indices in S as well as being the array whose entries are set to zero outside S.

Since the magnitude of each nonzero component of $\boldsymbol{D}_{S_{j-1}}$ is larger than the average magnitude of the nonzero components of $\boldsymbol{D}_{S_j}$,

$$\|\boldsymbol{D}_{S_j}\|_2 \leq \frac{\|\boldsymbol{D}_{S_{j-1}}\|_1}{2\sqrt{s}}, \quad j = 2, 3, \ldots.$$

Combining this with the cone constraint gives

$$\sum_{j=2}^{r} \|\boldsymbol{D}_{S_j}\|_2 \leq \frac{1}{2\gamma\sqrt{k}}\|\boldsymbol{D}_{S^c}\|_1 \leq \frac{1}{2\sqrt{k}}\|\boldsymbol{D}_S\|_1 + \frac{1}{2\gamma\sqrt{k}}\sigma \leq \frac{1}{2}\|\boldsymbol{D}_S\|_2 + \frac{1}{2\gamma\sqrt{k}}\sigma. \tag{41}$$

Now combining (41) with the tube constraint and the RIP,

$$\begin{aligned}
\varepsilon &\gtrsim \|\mathcal{A}\boldsymbol{D}\|_2 \\
&\geq \|\mathcal{A}(\boldsymbol{D}_S + \boldsymbol{D}_{S_1})\|_2 - \sum_{j=2}^{r} \|\mathcal{A}(\boldsymbol{D}_{S_j})\|_2 \\
&\geq \sqrt{1-\delta}\|\boldsymbol{D}_S + \boldsymbol{D}_{S_1}\|_2 - \sqrt{1+\delta}\sum_{j=2}^{r} \|\boldsymbol{D}_{S_j}\|_2 \\
&\geq \sqrt{1-\delta}\|\boldsymbol{D}_S + \boldsymbol{D}_{S_1}\|_2 - \sqrt{1+\delta}\left(\frac{1}{2}\|\boldsymbol{D}_S\|_2 + \frac{1}{2\gamma\sqrt{k}}\sigma\right) \\
&\geq \left(\sqrt{1-\delta} - \frac{\sqrt{1+\delta}}{2}\right)\|\boldsymbol{D}_S + \boldsymbol{D}_{S_1}\|_2 - \sqrt{1+\delta}\frac{1}{2\gamma\sqrt{k}}\sigma. \qquad (42)
\end{aligned}$$

Then, since $\delta < 1/3$,

$$\|\boldsymbol{D}_S + \boldsymbol{D}_{S_1}\|_2 \leq 5\varepsilon + \frac{3\sigma}{\gamma\sqrt{k}}.$$

Finally, because $\|\sum_{j=2}^{r} \boldsymbol{D}_{S_j}\|_2 \leq \sum_{j=2}^{r} \|\boldsymbol{D}_{S_j}\|_2 \leq \frac{1}{2}\|\boldsymbol{D}_S + \boldsymbol{D}_{S_1}\|_2 + \frac{1}{2\gamma\sqrt{k}}\sigma$, we have

$$\|\boldsymbol{D}\|_2 \leq 8\varepsilon + \frac{5\sigma}{\gamma\sqrt{k}},$$

confirming (13).

To confirm (14), note that the cone constraint allows the estimate

$$\begin{aligned}\|\boldsymbol{D}\|_1 &\leq (\gamma+1)\|\boldsymbol{D}_S\|_1 + \sigma \\ &\leq 2\gamma\sqrt{s}\|\boldsymbol{D}_S\|_2 + \sigma \\ &\leq 2\gamma\sqrt{k}\left(5\varepsilon + \frac{3\sigma}{\gamma\sqrt{k}}\right) + \sigma. \qquad (43)\end{aligned}$$

■

A.2. Proof of Proposition 7. Here we give a direct proof of the discrete Sobolev inequality (29) for images $\boldsymbol{X} \in \mathbb{C}^{N\times N}$ whose first row and first column of pixels are zero-valued, $X_{1,j} = X_{j,1} = 0$.

Proof. For any $1 \leq k \leq i \leq N$ we have

$$\begin{aligned}|X_{i,j}| &= \left|X_{1,j} + \sum_{\ell=1}^{i-1}\left(X_{\ell+1,j} - X_{\ell,j}\right)\right| \\ &\leq \sum_{\ell=1}^{i-1}|X_{\ell+1,j} - X_{\ell,j}| \\ &\leq \sum_{\ell=1}^{N-1}|X_{\ell+1,j} - X_{\ell,j}|. \qquad (44)\end{aligned}$$

Similarly, by reversing the order of indices we also have

$$|X_{i,j}| \leq \sum_{\ell=1}^{N-1}|X_{i,\ell+1} - X_{i,\ell}|. \qquad (45)$$

For ease of notation let

$$f(j) = \sum_{\ell=1}^{N-1}|X_{\ell+1,j} - X_{\ell,j}|,$$

and let

$$g(i) = \sum_{\ell=1}^{N-1}|X_{i,\ell+1} - X_{i,\ell}|.$$

Combining the two bounds (44) and (45) on $X_{i,j}$ results in the bound $|X_{i,j}|^2 \leq f(j)\cdot g(i)$.

Summing this inequality over all pixels (i, j),

$$\begin{aligned}
\|\boldsymbol{X}\|^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} |X_{i,j}|^2 &\leq \left(\sum_{j=1}^{N} f(j)\right)\left(\sum_{i=1}^{N} g(i)\right) \\
&\leq \frac{1}{4}\cdot\left(\sum_{j=1}^{N} f(j) + \sum_{i=1}^{N} g(i)\right)^2 \\
&\leq \frac{1}{4}\cdot\left(\sum_{j=1}^{N}\sum_{k=1}^{N-1} |X_{k+1,j} - X_{k,j}| + \sum_{i=1}^{N}\sum_{k=1}^{N-1} |X_{i,k+1} - X_{i,k}|\right)^2 \\
&\leq \frac{1}{4}\|\boldsymbol{\nabla X}\|_1^2 \\
&= \frac{1}{4}\|\boldsymbol{X}\|_{TV}^2. \qquad \blacksquare
\end{aligned} \tag{46}$$

A.3. Derivation of Proposition 8. Recall that a function $f(u, v)$ is in the space $L_p(\Omega)$ $(1 \leq p < \infty)$ if

$$\|f\|_{L_p(\Omega)} := \left(\int_{\Omega\subset\mathbb{R}^2} |f(x)|^p dx\right)^{1/p} < \infty,$$

and the space of functions with bounded variation on the unit square is defined as follows.

Definition 12. *$BV(Q)$ is the space of functions of bounded variation on the unit square $Q := [0, 1)^2 \subset \mathbb{R}^2$. For a vector $\boldsymbol{v} \in \mathbb{R}^2$, we define the difference operator $\Delta_{\boldsymbol{v}}$ in the direction of $\boldsymbol{v}$ by*

$$\Delta_{\boldsymbol{v}}(f, \boldsymbol{x}) := f(\boldsymbol{x} + \boldsymbol{v}) - f(\boldsymbol{x}).$$

We say that a function $f \in L_1(Q)$ is in $BV(Q)$ if and only if

$$V_Q(f) := \sup_{h>0} h^{-1} \sum_{j=1}^{2} \|\Delta_{h\boldsymbol{e}_j}(f, \cdot)\|_{L_1(Q(h\boldsymbol{e}_j))} = \lim_{h\to 0} h^{-1} \sum_{j=1}^{2} \|\Delta_{h\boldsymbol{e}_j}(f, \cdot)\|_{L_1(Q(h\boldsymbol{e}_j))}$$

is finite, where $\boldsymbol{e}_j$ denotes the jth coordinate vector. Here, the last equality follows from the fact that $\|\Delta_{h\boldsymbol{e}_j}(f, \cdot)\|_{L_1(Q)}$ is subadditive. $V_Q(f)$ provides a seminorm for BV:

$$|f|_{BV(Q)} := V_Q(f).$$

Theorem 8.1 of [19] bounds the rate of decay of a function's bivariate Haar coefficients by its bounded variation seminorm.

Theorem 13 (Theorem 8.1 of [19]). *Consider a function mean-zero $f \in BV(Q)$ and its bivariate Haar coefficients arranged in decreasing order according to their absolute value, $c_{(k)}(f)$. We have*

$$c_{(k)}(f) \leq C_1 \frac{|f|_{BV}}{k},$$

where $C_1 = 36(480\sqrt{5} + 168\sqrt{3})$.

As discrete images are isometric to piecewise-constant functions of the form (9), the bivariate Haar coefficients of the image $\boldsymbol{X} \in \mathbb{C}^{N \times N}$ are equal to those of the function $f_{\boldsymbol{X}} \in L_2(Q)$ given by

$$f_{\boldsymbol{X}}(u,v) = N\boldsymbol{X}_{i,j}, \quad \frac{i-1}{N} \leq u < \frac{i}{N}, \quad \frac{j-1}{N} \leq v < \frac{j}{N}, \quad 1 \leq i,j \leq N. \tag{47}$$

To derive Proposition 8, it will suffice to verify that the *bounded variation* of $f_{\boldsymbol{X}}$ can be bounded by the *total variation* of $\boldsymbol{X}$.

Lemma 14. $|f_{\boldsymbol{X}}|_{BV} \leq \|\boldsymbol{X}\|_{TV}$.

Proof. For $h < \frac{1}{N}$,

$$\Delta_{he_1}(f_{\boldsymbol{X}},(u,v)) = \begin{cases} N(\boldsymbol{X}_{i+1,j} - \boldsymbol{X}_{i,j}), & \frac{i}{N} - h \leq u \leq \frac{i}{N}, \quad \frac{j}{N} \leq v \leq \frac{j+1}{N}, \\ 0 & \text{else} \end{cases}$$

and

$$\Delta_{he_2}(f_{\boldsymbol{X}},(u,v)) = \begin{cases} N(\boldsymbol{X}_{i,j+1} - \boldsymbol{X}_{i,j}), & \frac{i}{N} \leq u \leq \frac{i+1}{N}, \quad \frac{j}{N} - h \leq v \leq \frac{j}{N}, \\ 0 & \text{else.} \end{cases}$$

Then

$$\begin{aligned} |f_{\boldsymbol{X}}|_{BV} &= \lim_{h \to 0} \frac{1}{h} \left[\int_0^1 \int_0^1 |f_{\boldsymbol{X}}(u+h,v) - f_{\boldsymbol{X}}(u,v)|\, dudv \right. \\ &\qquad \left. + \int_0^1 \int_0^1 |f_{\boldsymbol{X}}(u,v+h) - f_{\boldsymbol{X}}(u,v)|\, dvdu \right] \\ &= \sum_{j=1}^{N-1} \sum_{i=1}^{N-1} |\boldsymbol{X}_{i+1,j} - \boldsymbol{X}_{i,j}| + \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} |\boldsymbol{X}_{i,j+1} - \boldsymbol{X}_{i,j}| \\ &\leq \|\boldsymbol{X}\|_{TV} \qquad \blacksquare \end{aligned} \tag{48}$$

Acknowledgment. We would like to thank Arie Israel, Christina Frederick, Felix Krahmer, Stan Osher, Yaniv Plan, Justin Romberg, Joel Tropp, and Mark Tygert for invaluable discussions and improvements. We also would like to thank Fabio Lanzoni and his agent, Eric Ashenberg.

REFERENCES

[1] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of bounded variation and free discontinuity problems*, The Clarendon Press, Oxford University Press, New York, 2000.

[2] R. G. Baraniuk, M. Davenport, R. A. DeVore, and M. Wakin, *A simple proof of the Restricted Isometry Property for random matrices*, Constr. Approx., 28 (2008), pp. 253–263.

[3] J. Bioucas-Dias and M. Figueiredo, *A new twist: Two-step iterative thresholding algorithm for image restoration*, IEEE Trans. Image Process., 16 (2007), pp. 2992–3004.

[4] T. Blumensath and M. E. Davies, *Iterative hard thresholding for compressed sensing*, Appl. Comput. Harmon. Anal., 27 (2009), pp. 265–274.

[5] L. Bregman, *The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. Math. Phys., 7 (1967), pp. 200–217.

[6] J.-F. Cai, B. Dong, S. Osher, and Z. Shen, *Image restoration: Total variation, wavelet frames, and beyond*, J. Amer. Math. Soc., 25 (2012), pp. 1033–1089.

[7] E. Candès, *The restricted isometry property and its implications for compressed sensing*, C. R. Math. Acad. Sci. Paris, 346 (2008), pp. 589–592.

[8] E. Candès, Y. Eldar, D. Needell, and P. Randall, *Compressed sensing with coherent and redundant dictionaries*, Appl. Comput. Hamon. Anal., 31 (2011), pp. 59–73.

[9] E. Candès and F. Guo, *New multiscale transforms, minimum total variation synthesis: Applications to edge-preserving image reconstruction*, Signal Process., 82 (2002), pp. 1519–1543.

[10] E. Candès and J. Romberg, *Signal recovery from random projections*, in Proceedings of the SPIE Conference on Computational Imaging III, Vol. 5674, SPIE, Bellingham, WA, 2005, pp. 76–86.

[11] E. Candès, J. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math., 59 (2006), pp. 1207–1223.

[12] E. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory, 51 (2005), pp. 4203–4215.

[13] E. Candès and T. Tao, *Near optimal signal recovery from random projections: Universal encoding strategies?*, IEEE Trans. Inform. Theory, 52 (2006), pp. 5406–5425.

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

[15] A. Chambolle, *An algorithm for total variation minimization and applications*, J. Math. Imaging Vision, 20 (2004), pp. 89–97.

[16] T. F. Chan, J. Shen, and H. M. Zhou, *Total variation wavelet inpainting*, J. Math. Imaging Vision, 25 (2006), pp. 107–125.

[17] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by Basis Pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33–61.

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

[19] A. Cohen, R. DeVore, P. Petrushev, and H. Xu, *Nonlinear approximation and the space* $BV(\mathbb{R}^2)$, Amer. J. Math., 121 (1999), pp. 587–628.

[20] *Compressed Sensing Resources*, http://www.dsp.ece.rice.edu/cs/.

[21] G. B. Dantzig and M. N. Thapa, *Linear Programming*, Springer, New York, 1997.

[22] R. A. DeVore, B. Jawerth, and B. J. Lucier, *Image compression through wavelet transform coding*, IEEE Trans. Inform. Theory, 38 (1992), pp. 719–746.

[23] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, *Single-pixel imaging via compressive sampling*, IEEE Signal Processing Mag., 25 (2008), pp. 83–91.

[24] ℓ_1*-MAGIC software*, http://users.ece.gatech.edu/~justin/l1magic/.

[25] A. Yu. Garnaev and E. D. Gluskin, *The widths of a Euclidean ball*, Soviet Math. Dokl., 30 (1984), pp. 200–204.

[26] T. Goldstein and S. Osher, *The split Bregman algorithm for L1-regularized problems*, SIAM J. Imaging Sci., 2 (2009), pp. 323–343.

[27] B. Kai Tobias, U. Martin, and F. Jens, *Suppression of MRI truncation artifacts using total variation constrained data extrapolation*, Int. J. Biomed. Imaging, 2008 (2008), 184123.

[28] B. Kashin, *The widths of certain finite dimensional sets and classes of smooth functions*, Izvestia, 41 (1977), pp. 334–351.

[29] S. L. Keeling, *Total variation based convex filters for medical imaging*, Appl. Math. Comput., 139 (2003), pp. 101–119.

[30] F. Krahmer and R. Ward, *New and improved Johnson–Lindenstrauss embeddings via the restricted isometry property*, SIAM J. Math. Anal., 43 (2011), pp. 1269–1281.

[31] F. Krahmer and R. Ward, *Beyond incoherence: Stable and robust sampling strategies for compressive imaging*, submitted.

[32] Y. Liu and Q. Wan, *Total Variation Minimization Based Compressive Wideband Spectrum Sensing for Cognitive Radios*, preprint, arXiv:1106.3629v1 [cs.IT], 2011.

[33] M. Lustig, D. Donoho, and J. M. Pauly, *Sparse MRI: The application of compressed sensing for rapid MRI imaging*, Magn. Reson. Med., 58 (2007), pp. 1182–1195.

[34] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, *Compressed sensing MRI*, IEEE Signal. Processing Mag., 25 (2008), pp. 72–82.

[35] S. Ma, W. Yin, Y. Zhang, and A. Chakraborty, *An efficient algorithm for compressed MR imaging using total variation and wavelets*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[36] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, *Uniform uncertainty principle for Bernoulli and subgaussian ensembles*, Constr. Approx., 28 (2008), pp. 277–289.

[37] Q. Mo and S. Li, *New bounds on the restricted isometry constant δ_{2k}*, Appl. Comput. Harmon. Anal., 31 (2011), pp. 460–468.

[38] S. Nam, M. E. Davies, M. Elad, and R. Gribonval, *The cosparse analysis model and algorithms*, Appl. Comp. Harmon. Anal., 34 (2013), pp. 30–56.

[39] B. K. Natarajan, *Sparse approximate solutions to linear systems*, SIAM J. Comput., 24 (1995), pp. 227–234.

[40] D. Needell and J. A. Tropp, *CoSaMP: Iterative signal recovery from noisy samples*, Appl. Comput. Harmon. Anal., 26 (2008), pp. 301–321.

[41] D. Needell and R. Ward, *Total variation minimization for stable multidimensional signal recovery*, submitted.

[42] B. Nett, J. Tang, S. Leng, and G. H. Chen, *Tomosynthesis via total variation minimization reconstruction and prior image constrained compressed sensing (PICCS) on a C-arm system*, Proc. Soc. Photo Opt. Instrum. Eng., 6913 (2008), 92672.

[43] S. Osher, A. Solé, and L. Vese, *Image decomposition and restoration using total variation minimization and the H^{-1} norm*, Multiscale Model. Simul., 1 (2003), pp. 349–370.

[44] V. Patel, R. Maleh, A. C. Gilbert, and R. Chellappa, *Gradient-based image recovery methods from incomplete Fourier measurements*, IEEE Trans. Image Process., 21 (2012), pp. 94–105.

[45] H. Rauhut, *Compressive sensing and structured random matrices*, in Theoretical Foundations and Numerical Methods for Sparse Recovery, M. Fornasier, ed., Radon Ser. Comput. Appl. Math. 9, Walter de Gruyter, Berlin, 2010, pp. 1–92.

[46] H. Rauhut, J. Romberg, and J. A. Tropp, *Restricted isometries for partial random circulant matrices*, Appl. Comput. Harmon. Anal., 32 (2012), pp. 242–254.

[47] H. Rauhut and R. Ward, *Sparse Legendre expansions via ℓ_1-minimization*, J. Approx. Theory, 164 (2012), pp. 517–533.

[48] M. Rudelson and R. Vershynin, *On sparse reconstruction from Fourier and Gaussian measurements*, Comm. Pure Appl. Math., 61 (2008), pp. 1025–1045.

[49] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.

[50] D. Strong and T. Chan, *Edge-preserving and scale-dependent properties of total variation regularization*, Inverse Problems, 19 (2003), pp. S165–S187.

[51] S. Vaiter, G. Peyré, C. Dossal, and J. Fadili, *Robust sparse analysis regularization*, IEEE Trans. Inform. Theory, 59 (2013), pp. 2001–2016.

[52] J. Yang, Y. Zhang, and W. Yin, *A fast alternating direction method for TVL_1-L_2 signal reconstruction from partial Fourier data*, IEEE J. Sel. Topics Signal Process., 4 (2010), pp. 288–297.