Guarantees of Total Variation Minimization for Signal Recovery

Jian-Feng Cai* Weiyu Xu†

March 11, 2015

Abstract

In this paper, we consider using total variation (TV) minimization to recover signals whose gradients have a sparse support, from a small number of measurements. We establish a proof for the performance guarantee of TV minimization in recovering *one-dimensional* signal with sparse gradient support. This answers the open question of proving the fidelity of TV minimization in such a setting. We have shown that, when the number of Gaussian measurements $M \gtrsim \sqrt{NK}\log N$, the TV minimization guarantees the exact recovery of any signal of size N with at most K nonzero gradients with high probability; when $M \lesssim \sqrt{NK}$, the TV minimization cannot find the original signal with a moderate probability. Last but not least, when M grows linearly with the signal dimension, we will also show that the recoverable sparsity K grows linearly with the signal dimension as well.

1 Introduction

Compressed sensing has recently gained a lot of attention in many applications, because it enables acquiring sparse signals from a much smaller number of samples than the ambient dimension of signal. Compressed sensing takes advantage of the fact that most signals of interest in practice are sparse: there are only a few nonzero or big elements when the signals are represented over a certain dictionary such as wavelet basis. For these types of sparse or compressible signals, compressed sensing theory [7, 13] has established that a small number of nonadaptive measurements are often sufficient to efficiently recover them under methods such as ℓ_1-minimization [6–8, 13].

To begin with, let us assume that $\bar{\boldsymbol{x}} \in \mathbb{R}^N$ is a one-dimensional (compared with 2-dimensional images and 3-dimensional videos) signal vector of N elements, and has no more than K ($K < N$) nonzero elements.

*Department of Mathematics, University of Iowa, Iowa City, IA 52242, USA. Email: `jianfeng-cai@uiowa.edu`

†Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, USA. Email: `weiyu-xu@uiowa.edu`

In compressed sensing, we sample $\bar{\boldsymbol{x}}$ using M $(M < N)$ linear projections

$$\boldsymbol{y} = A\bar{\boldsymbol{x}},$$

where A is an $M \times N$ measuring matrix and $\boldsymbol{y}$ is an $M \times 1$ measurement vector. Knowing A and the measurement $\boldsymbol{y}$, ℓ_1 minimization is often used to recover the sparse $\bar{\boldsymbol{x}}$:

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \qquad \text{subject to} \;\; \boldsymbol{y} = A\boldsymbol{x}. \tag{1}$$

It has been shown that under suitable conditions on the measuring matrix A, it is guaranteed that the original $\bar{\boldsymbol{x}}$ is the unique solution to ℓ_1-minimization (1). In fact, if A satisfies the so-called restricted isometry property (RIP), then the solution of (1) matches exactly with the original signal [4, 7, 16]. When noise is presented and/or the signal is not exactly but approximately sparse, a variant of (1) leads to a solution that does not deviate too much from the original signal [4]. Various results concerning the perfect and stable reconstruction of the original signal by solving (1) have been established via the restricted isometry condition, the null space condition and the exact recovery principle in [7, 9, 13–16, 19, 25, 45, 47].

The results above hold true only for sparse signals, and they can be extended to signals that are synthesized by a linear combination of few atoms in a (redundant) dictionary with incoherent atoms [36]. However, there are numerous practical examples in which a signal of interest does not fall into the category in the aforementioned theoretical work. One such an example is signal that has a sparse gradient (i.e., the signal is piecewise constant), which arises frequently from imaging. Images with little detail are usually modelled as piecewise constant functions. For simplicity, we assume that $\bar{\boldsymbol{x}} \in \mathbb{R}^N$ is a vector representing one-dimensional piecewise constant signal. Let $D\bar{\boldsymbol{x}}$ be its finite difference defined by $[D\bar{\boldsymbol{x}}]_i = \bar{x}_{i+1} - \bar{x}_i$ for $i = 1, 2, \ldots, N-1$. Since $\bar{\boldsymbol{x}}$ is piecewise constant, we must have that $D\bar{\boldsymbol{x}}$ is sparse. Assume that $D\bar{\boldsymbol{x}}$ has only K $(K < N)$ nonzero entries. Let $\boldsymbol{y} = A\bar{\boldsymbol{x}} \in \mathbb{R}^M$ be M linear samples of $\bar{\boldsymbol{x}}$. Then, to recover $\bar{\boldsymbol{x}}$, one usually solves

$$\min_{\boldsymbol{x}} \|D\boldsymbol{x}\|_1 \qquad \text{subject to} \;\; \boldsymbol{y} = A\boldsymbol{x}. \tag{2}$$

The regularization term $\|D\boldsymbol{x}\|_1$ is called the *total variation (TV)* of $\boldsymbol{x}$. When $\bar{\boldsymbol{x}} \in \mathbb{R}^{N^d}$ is generated from d-dimensional signals, we only need to replace D by the concatenation of directional finite differences, and $\|D\boldsymbol{x}\|_1$ is the anisotropic TV of $\boldsymbol{x}$.

TV regularization has been used extensively in the literature for decades in imaging sciences [2, 26, 38, 39] and other related fields [10,42]. The minimization problem (2) has the same form as the minimization in the analysis-based compressed sensing in [5]. However, the perfect/stable reconstruction result in [5] can not be applied to (2), as the rows of D do not form a frame (D has a nontrivial null space and a large condition number). Furthermore, the analysis in cosparse analysis model [29,30] and other ℓ_1 analysis models [17,22,23] cannot be applied directly as well. Despite the great importance of the TV minimization in applications, rigorous proofs of conditions of successfully recovering signal by using the TV minimization have only recently been established [31,32]. To establish such conditions, [31,32] first transformed d-dimensional ($d \geq 2$) signals with sparse gradients into signals compressible over the Haar orthogonal wavelet basis. Then a modified restricted isometry condition, which takes into account the Haar orthogonal wavelet transformation, was established for the matrix A such that (2) offers a stable recovery of $\boldsymbol{x}$. However, it is noted in [31, 32] that establishing conditions for successfully recovering one-dimensional (namely d=1) signal vector remains an open problem. This is partially due to the fact that small TV of a one-dimensional signal does not necessarily imply fast decay of its Haar wavelet coefficients. In fact, TV minimization for 1-dimensional signal plays a very important role in signal processing, for example, in soil moisture monitoring [44]. It should be noted that some applications change an image into a one-dimensional signal and adapt the finite difference operator [30, 39, 41].

In this paper, we establish the proof for performance guarantees of TV minimization in recovering *one-dimensional* signal with sparse gradient support. This partially answers the open problem of proving the fidelity of total variation minimization in such a setting [31]. We will prove the following results.

1. When $M \gtrsim \sqrt{NK} \log N$, the TV minimization (2) finds the original signal with overwhelming probability.

2. When $M \lesssim \sqrt{NK}$, the TV minimization (2) fails to find some signals with sparse gradient with a moderate probability.

3. When M grows linearly with N, the recoverable threshold of K can grow linearly with N as well. Moreover, in this case, the recovery is robust to noisy data and non-exactly but approximately sparse signals.

Compared with [31, 32], our results do not use the restricted isometry condition, but directly work on the null space condition of the measuring matrix A. To establish the null space condition of interest, we use "Escape through the Mesh" theorem [11,21,37,40] to estimate the Gaussian width [21,37] of a cone specified by the null space condition. We further extend our results to TV minimization for higher dimensional signals. For $d \geq 2$, we have obtained performance bounds for TV minimization comparable to results in [31, 32]. In [12], an average-case phase transition was calculated for an approximate message passing (AMP) algorithm using the a total variation denoiser, through evaluating the asymptotic minimax Mean Square Error (MSE) for a separate denoising problem using total variation regularizer. Experimental results have shown excellent agreements between the phase transition curves calculated for the AMP algorithm in [12], and the empirical performance of convex programming based total variation minimization approach. However, the phase transition bounds in [12] rely on two key assumptions: the density evolution for the AMP algorithm is correct; and the convex programming based total variation minimization indeed has the same phase transition as the AMP algorithm. These two assumptions remain to be proved for total variation minimization. The recent work [33] by Oymak et al. interestingly shows that the minimax MSE derived in [12] is related to the expectation of the square of the distance between a Gaussian vector and the λ-scaled subdifferential of $\|Dx\|_1$ at x, for some optimized $\lambda > 0$. However, it is not clear whether that expectation is equal to the Gaussian width for TV minimization studied in this paper, which is instead equal to the expectation of the square of the distance between a Gaussian vector and the conic hull of the subdifferential of $\|Dx\|_1$. In addition, compared with [12], our results discuss worst-case performance guarantees which are uniformly true for all the possible supports for the signal gradient. The average-case phase transition results from [12] are mainly for the case when M/N and K/N approach a nonzero constant, and it is not straightforward to extend them to worst-case performance bounds for K not growing linearly with the problem dimension.

After our paper was submitted, there have been some recent works on sample complexity bounds for TV minimization, including [35] and [24]. [35] discusses the performance of total variation under non-uniformly chosen Fourier measurements, and has discovered that the performance of total variation under such measurements not only depends on the chosen Fourier measurements, but also on the signal structure, namely how far apart the nonzero signal gradients are from each other. This observation of TV minimization performance depending on the signal structure beyond signal gradient sparsity, is consistent with our worst-case result from a particular gradient support pattern, and also consistent with our worst-case and average-case simulation results, albeit we are discussing random Gaussian measurements. [24] also studies the performance of TV minimization using the tool of Gaussian width, which is also used in our paper. Our work shows that for any $\alpha > 0$ such that $M = \alpha N$, the recoverable sparsity K of Dx can grow linearly with the problem dimension N. In comparison, [24] gives performance bounds on K when $M/N = \alpha$ is bigger than a nonzero constant. More precisely, for $K = 1$, the result in [24] needs $M \geq (1 - 1/\pi)N$, while our result requires only

$M \gtrsim \sqrt{N} \log N$ and is optimal up to a logarithm factor.

The rest of this paper is organized as follows. In Section 2, we present main theorems of this paper, and then these theorems are proved in Sections 3 and 4 respectively. In Section 5, numerical experiments are performed to demonstrate our results in the main theorems. In Section 6, we extend our results to TV minimization for multidimensional signals. Section 7 concludes our paper and discusses future directions.

2 Main Results

In this section, we establish the main result of this paper on the performance guarantee of TV minimization in recovering one-dimensional signal with sparse gradient support.

Let $D \in \mathbb{R}^{(N-1)\times N}$ be a discrete gradient operator defined as

$$D = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}.$$

Note that here a Neumann boundary condition is used in the discrete gradient operator. One can impose other boundary conditions such as zero and periodic boundary conditions, and our results in the paper can be extended to those cases without too much difficulty.

We will first assume that $\bar{\boldsymbol{x}}$ has a sparse gradient, i.e.,

$$K = \|D\bar{\boldsymbol{x}}\|_0 < N, \tag{3}$$

where $\|\cdot\|_0$ is the number of nonzero elements of a vector. We would like to recover $\bar{\boldsymbol{x}}$ from its linear samples $\boldsymbol{y} = A\bar{\boldsymbol{x}} \in \mathbb{R}^M$, where $A \in \mathbb{R}^{M\times N}$ is a random Gaussian matrix whose entries are independently drawn from the standard normal distribution. If we recover x by finding the solution to "$\arg\min \|Dx\|_0$ subject to $y = Ax$", then one can show that we will need at least $(2K+1)$ measurements to guarantee recovering any signal with K-sparse gradient. However, minimizing $\|Dx\|_0$ is computationally challenging. Instead, to recover $\bar{\boldsymbol{x}}$ from $\boldsymbol{y}$, we often solve the convex optimization problem (2).

Our main results of this paper are Theorems 2.1 and 2.2 below. In Theorem 2.1, we will prove: when the number of measurements $M \gtrsim \sqrt{NK}\log N$, the TV minimization (2) guarantees the exact recovery of any signal of size N with at most K nonzero gradients with high probability; when $M \lesssim \sqrt{NK}$, the TV minimization (2) cannot find the original signal with a moderate probability. Throughout the paper, we will use $C_0, C_1, \ldots$ for constants that are independent of N, M, K. To save notations, the same C_i in different theorems and lemmas may refer to different constants.

Theorem 2.1. *Let $A \in \mathbb{R}^{M\times N}$ be a random Gaussian matrix whose entries follow independently the standard normal distribution. Let $\hat{\boldsymbol{x}}$ be a solution of (2) with data $\boldsymbol{y} = A\bar{\boldsymbol{x}}$.*

(a). There exist positive constants $C_1, C_2, C_3, C_4 > 0$ such that, with probability at least $1 - C_1 e^{-C_2\sqrt{M}}$,

$$\hat{\boldsymbol{x}} = \bar{\boldsymbol{x}}, \quad \forall\ \bar{\boldsymbol{x}} \text{ satisfying } \|D\bar{\boldsymbol{x}}\|_0 \le K,$$

provided

$$M \ge C_3\sqrt{NK}(\log N + C_4).$$

(b). For any $0 < \eta < 1$, there exist positive constants $\widetilde{C}_0, \widetilde{C}_1$ and a universal positive constant $\widetilde{C}_2$ such that the following statement holds. Let $K \ge \widetilde{C}_0$ and $(K+1) < N/4$. There exists infinitely many $\bar{\boldsymbol{x}} \in \mathbb{R}^N$ with $\|D\bar{\boldsymbol{x}}\|_0 = K$ such that, with probability at least $1 - \eta$,

$$\hat{\boldsymbol{x}} \ne \bar{\boldsymbol{x}}$$

provided

$$M \le \widetilde{C}_1\sqrt{NK} - \widetilde{C}_2.$$

Though the above result works for any K and N, the bound $M \gtrsim \sqrt{NK}\log N$ that guarantees the perfect recovery can be refined to $M \gtrsim K$ for large M that grows linearly with N. Moreover, the perfect recovery result for noiseless data and exactly sparse signal can be extended to noisy data and non-exactly but approximately sparse signals. More precisely, we assume $\boldsymbol{y} = A\bar{\boldsymbol{x}} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a noise and satisfies $\|\boldsymbol{\epsilon}\|_2 \le \epsilon$; besides, suppose that $D\bar{\boldsymbol{x}}$ is not necessarily exactly but approximately sparse, in particular, $\min_{|\mathcal{K}|\le K} \|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1$ is small (note that $\min_{|\mathcal{K}|\le K} \|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1$ is the error of the best K-term approximation of the gradient). In this case, we let the recovered signal be a solution $\hat{\boldsymbol{x}}$ of the following optimization problem

$$\min_{\boldsymbol{x}} \|D\boldsymbol{x}\|_1 \quad \text{subject to } \|A\boldsymbol{x} - \boldsymbol{y}\|_2 \le \epsilon. \tag{4}$$

Our second main theorem is as follows.

Theorem 2.2. *Let $A \in \mathbb{R}^{M\times N}$ be a random Gaussian matrix whose entries follow independently the standard normal distribution. Let $\hat{\boldsymbol{x}}$ be a solution of* (4) *with the data $\boldsymbol{y}$ satisfying $\|A\bar{\boldsymbol{x}} - \boldsymbol{y}\|_2 \leq \epsilon$. Then, for any constant $0 < \alpha < 1$, there exist positive constants $\delta, C_0, C_1, C_2, C_3$ such that the following statement holds true. Let $M = \alpha N$ and $K = \delta N$. Then, with probability at least $1 - C_0 e^{-C_1 N}$,*

$$\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}\|_2 \leq C_2 \frac{\min_{|\mathcal{K}|\leq K} \|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1}{\sqrt{N}} + C_3 \frac{\epsilon}{\sqrt{N}}, \tag{5}$$

The term $\min_{|\mathcal{K}|\leq K} \|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1$ in (5) is simply the error of the best K-sparse approximation to $D\bar{\boldsymbol{x}}$. In other words, if we choose $\mathcal{K}_0$ to be the set of indices of the largest K components of $D\bar{\boldsymbol{x}}$ in absolute value, then it is obvious that $\|(D\bar{\boldsymbol{x}}) - (D\bar{\boldsymbol{x}})_{\mathcal{K}_0}\|_1 = \min_{|\mathcal{K}|\leq K} \|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1$. Therefore, our result in Theorem 2.2 shows that, even if $D\bar{\boldsymbol{x}}$ is not exactly sparse, the error of the solution to (4) is only a multiple of the best K sparse approximation error.

3 Proof of Theorem 2.1

In this section, we prove Theorem 2.1 on the performance guarantee of TV minimization in recovering one-dimensional signal with sparse gradient support. Our proof uses "Escape through the Mesh" theorem [21, 37, 40] for the bound of M for the successful recovery, and the bound of M for the unsuccessful recovery is proved by a theorem from [1]. In both proofs, Gaussian width plays a fundamental role.

3.1 Bound for successful recovery

In this section, we give the bound of M for the successful recovery, i.e., we provide a proof of Part (a) of Theorem 2.1. Our proof is based on the null space property (c.f. Lemma 3.1 below) and the "escape through the mesh" theorem [21] (c.f. Theorem 3.2 below).

Lemma 3.1. *Let $A \in \mathbb{R}^{M\times N}$. Let $\hat{\boldsymbol{x}}$ be a solution of* (2) *with data $\boldsymbol{y} = A\bar{\boldsymbol{x}}$. Then,*

$$\hat{\boldsymbol{x}} = \bar{\boldsymbol{x}}, \quad \forall\ \bar{\boldsymbol{x}}\ s.t.\ \|D\bar{\boldsymbol{x}}\|_0 \leq K$$

if and only if the following condition holds: for every nonzero vector $\boldsymbol{z}$ in the null space of A (namely $A\boldsymbol{z} = 0$, $\boldsymbol{z} \neq \boldsymbol{0}$),

$$\|(D\boldsymbol{z})_{\mathcal{K}}\|_1 < \|(D\boldsymbol{z})_{\mathcal{K}^c}\|_1, \quad \forall\ \mathcal{K} \subset \{1, 2, \ldots, N-1\},\ s.t.\ |\mathcal{K}| \leq K. \tag{6}$$

Proof. We omit the detailed proof of this lemma, since it is very similar to the proof of null space conditions for ℓ_1 minimization; see, for example, [40, 47]. $\square$

Theorem 3.2 (Escape through the mesh [21])**.** *Let $\mathcal{S}$ be a subset of the unit Euclidean sphere in $\mathbb{R}^N$. Let $\mathcal{Y}$ be a random $(N-M)$-dimensional subspace of $\mathbb{R}^N$, distributed uniformly in the Grassmanian with respect to the Haar measure. Assume that $w(\mathcal{S}) < (\sqrt{M} - \frac{1}{2\sqrt{M}})$, where*

$$w(\mathcal{S}) = E\left(\sup_{\boldsymbol{x}\in\mathcal{S}} \langle \boldsymbol{x}, \boldsymbol{g}\rangle\right), \tag{7}$$

with the entries of $\boldsymbol{g} \in \mathbb{R}^N$ following independently the standard normal distribution and E being the expectation, is the Gaussian width of the set $\mathcal{S}$. Then

$$P(\mathcal{Y} \bigcap \mathcal{S} = \emptyset) > 1 - 3.5\exp\left(-\frac{(\sqrt{M} - 1/(2\sqrt{M})) - w(\mathcal{S})}{18}\right).$$

In order to use Theorem 3.2 to prove Part (a) of Theorem 2.1, we use $\mathcal{S}$ be the intersection of the unit Euclidean sphere and the set that violates (6), i.e.,

$$\mathcal{S} = \{\boldsymbol{x} \ : \ \|\boldsymbol{x}\|_2 = 1, \quad \text{and} \quad \exists \mathcal{K} \subset \{1, 2, \dots, N\} \text{ s.t. } |\mathcal{K}| \le K, \ \|(D\boldsymbol{x})_{\mathcal{K}}\|_1 \ge \|(D\boldsymbol{x})_{\mathcal{K}^c}\|_1\}. \tag{8}$$

In the following, we estimate the Gaussian width $w(\mathcal{S})$ of $\mathcal{S}$. We will prove that an upper bound of $w(\mathcal{S})$ is $O\left((NK)^{\frac{1}{4}} \cdot \log^{\frac{1}{2}} N\right)$ and a lower bound is $O\left((NK)^{\frac{1}{4}}\right)$. Therefore, our estimation is quite tight as the the lower bound and upper bound are in the same order up to a log factor. The results are summarized into the following theorem.

Theorem 3.3. *The Gaussian width $w(\mathcal{S})$ with $\mathcal{S}$ defined in* (8) *satisfies*

$$\sqrt{\frac{2}{\pi}}\left(\frac{1-\frac{K+1}{N}}{2+\sqrt{\frac{K}{N}}}(NK)^{\frac{1}{4}} - 2(NK)^{-\frac{1}{4}}\right) \le w(\mathcal{S}) \le 2^{\frac{5}{4}}(2\sqrt{5}+\sqrt{10})(NK)^{\frac{1}{4}} \cdot \sqrt{\log(2N)}.$$

Proof. We estimate the upper bound first. For any $\boldsymbol{x} \in \mathcal{S}$, we have that

$$\|(D\boldsymbol{x})_{\mathcal{K}^c}\|_1 \le \|(D\boldsymbol{x})_{\mathcal{K}}\|_1 \le \sqrt{K}\|(D\boldsymbol{x})_{\mathcal{K}}\|_2 \le \sqrt{K}\|D\boldsymbol{x}\|_2 \le 2\sqrt{K}\|\boldsymbol{x}\|_2 = 2\sqrt{K},$$

which implies $\|D\boldsymbol{x}\|_1 \le 4\sqrt{K}$. Therefore

$$\mathcal{S} \subset \widetilde{\mathcal{S}} := \{\boldsymbol{x} \ : \ \|\boldsymbol{x}\|_2 \le 1, \ \ \|D\boldsymbol{x}\|_1 \le 4\sqrt{K}\}.$$

and further $w(\mathcal{S}) \le w\big(\widetilde{\mathcal{S}}\big)$. We thus instead estimate an upper bound of $w\big(\widetilde{\mathcal{S}}\big)$.

We temporarily assume that $N = 2^L$. For any $\boldsymbol{x} \in \widetilde{\mathcal{S}}$, we decompose $\boldsymbol{x}$ according to a Haar wavelet transform as

$$\boldsymbol{x} = \hat{\boldsymbol{z}}^{(1)} + \ldots + \hat{\boldsymbol{z}}^{(L)} + \hat{\boldsymbol{y}}^{(L)}, \tag{9}$$

where

$$\hat{\boldsymbol{z}}^{(\ell)} = \boldsymbol{z}^{(\ell)} \otimes [\underbrace{1 \ \ldots \ 1}_{2^{\ell-1}} \ \underbrace{-1 \ \ldots \ -1}_{2^{\ell-1}}], \quad \boldsymbol{z}^{(\ell)} = [z_1^{(\ell)} \ z_2^{(\ell)} \ \ldots \ z_{N/2^\ell}^{(\ell)}]$$

and

$$\hat{\boldsymbol{y}}^{(\ell)} = \boldsymbol{y}^{(\ell)} \otimes [\underbrace{1 \ 1 \ \ldots \ 1}_{2^{\ell}}], \quad \boldsymbol{y}^{(\ell)} = [y_1^{(\ell)} \ \ldots \ y_{N/2^\ell}^{(\ell)}].$$

Here $\otimes$ is the Kronecker product, i.e., $\boldsymbol{a} \otimes \boldsymbol{b} := [a_1\boldsymbol{b} \ a_2\boldsymbol{b} \ \ldots \ a_n\boldsymbol{b}]$. The decomposition (9) is done recursively as follows. We first define $\boldsymbol{y}^{(0)} = \hat{\boldsymbol{y}}^{(0)} = \boldsymbol{x}$. Then, at level ℓ, we decompose $\hat{\boldsymbol{y}}^{(\ell)}$ as

$$\hat{\boldsymbol{y}}^{(\ell)} = \hat{\boldsymbol{y}}^{(\ell+1)} + \hat{\boldsymbol{z}}^{(\ell+1)},$$

where

$$y_i^{(\ell+1)} = \frac{y_{2i-1}^{(\ell)} + y_{2i}^{(\ell)}}{2}, \quad \text{and} \quad z_i^{(\ell+1)} = \frac{y_{2i-1}^{(\ell)} - y_{2i}^{(\ell)}}{2}.$$

The decomposition (9) possesses the following properties.

(a) Components in decomposition (9) are orthogonal to each others. Consequently,

$$\|\boldsymbol{x}\|_2^2 = \|\hat{\boldsymbol{z}}^{(1)}\|_2^2 + \|\hat{\boldsymbol{z}}^{(2)}\|_2^2 + \ldots + \|\hat{\boldsymbol{z}}^{(L)}\|_2^2 + \|\hat{\boldsymbol{y}}^{(L)}\|_2^2 = \sum_{\ell=1}^{L} \left(2^\ell \|\boldsymbol{z}^{(\ell)}\|_2^2\right) + 2^L\|\boldsymbol{y}\|_2^2.$$

Since $\boldsymbol{x} \in \widetilde{\mathcal{S}}$ implies $\|\boldsymbol{x}\|_2^2 \le 1$, we have

$$\sum_{\ell=1}^{L} \left(2^\ell \|\boldsymbol{z}^{(\ell)}\|_2^2\right) + 2^L\|\boldsymbol{y}^{(L)}\|_2^2 \le 1. \tag{10}$$

(b) By direct calculation, we have

$$\begin{aligned}
\|D\hat{\boldsymbol{y}}^{(\ell)}\|_1 = \sum_{i=1}^{N/2^\ell-1} |y_{i+1}^{(\ell)} - y_i^{(\ell)}| &= \sum_{i=1}^{N/2^\ell-1} \left| \frac{y_{2i+1}^{(\ell-1)} + y_{2i+2}^{(\ell-1)}}{2} - \frac{y_{2i-1}^{(\ell-1)} + y_{2i}^{(\ell-1)}}{2} \right| \\
&= \sum_{i=1}^{N/2^\ell-1} \left| \frac{y_{2i+2}^{(\ell-1)} - y_{2i+1}^{(\ell-1)}}{2} + \left(y_{2i+1}^{(\ell-1)} - y_{2i}^{(\ell-1)}\right) + \frac{y_{2i}^{(\ell-1)} - y_{2i-1}^{(\ell-1)}}{2} \right| \\
&\le \sum_{i=1}^{N/2^\ell-1} \left(\left| \frac{y_{2i+2}^{(\ell-1)} - y_{2i+1}^{(\ell-1)}}{2} \right| + \left| y_{2i+1}^{(\ell-1)} - y_{2i}^{(\ell-1)} \right| + \left| \frac{y_{2i}^{(\ell-1)} - y_{2i-1}^{(\ell-1)}}{2} \right| \right) \\
&\le \sum_{i=1}^{N/2^{\ell-1}-1} |y_{i+1}^{(\ell-1)} - y_i^{(\ell-1)}| = \|D\hat{\boldsymbol{y}}^{(\ell-1)}\|_1,
\end{aligned}$$

which implies

$$\|\boldsymbol{z}^{(\ell)}\|_1 \le \|D\boldsymbol{y}^{(\ell-1)}\|_1/2 = \|D\hat{\boldsymbol{y}}^{(\ell-1)}\|_1/2 \le 2\sqrt{K}. \tag{11}$$

Now we are ready to estimate an upper bound of $w(\widetilde{\mathcal{S}})$. Let $\boldsymbol{g}$ be a vector whose entries are i.i.d. Gaussian random variables with mean 0 and variance 1. Since (10) implies $\|\boldsymbol{z}^{(\ell)}\|_2 \le \sqrt{2^{-\ell}}$, we have, by Cauchy-Schwartz inequality, $\|\boldsymbol{z}^{(\ell)}\|_1 \le \sqrt{N/2^\ell}\|\boldsymbol{z}^{(\ell)}\|_2 \le 2^{-\ell}\sqrt{N}$. This together with (11) implies that $\|\boldsymbol{z}^{(\ell)}\|_1 \le \min\{2^{-\ell}\sqrt{N}, 2\sqrt{K}\}$. Then,

$$\langle \hat{\boldsymbol{z}}^{(\ell)}, \boldsymbol{g}\rangle = \langle \boldsymbol{z}^{(\ell)}, \boldsymbol{g}^{(\ell)}\rangle \le \|\boldsymbol{z}^{(\ell)}\|_1 \|\boldsymbol{g}^{(\ell)}\|_\infty, \tag{12}$$

where $\boldsymbol{g}^{(\ell)} = [g_1^{(\ell)}\ g_2^{(\ell)}\ \ldots\ g_{N/2^\ell}^{(\ell)}]$ with $g_i^{(\ell)} = \sum_{j=1}^{2^{\ell-1}} (g_{j+(i-1)2^\ell} - g_{j+2^{\ell-1}+(i-1)2^\ell})$. Since the components in $\boldsymbol{g}^{(\ell)}$ are i.i.d. random Gaussian variables with mean 0 and variance 2^ℓ, it follows from [18, Proposition 8.1] that

$$E\left(\|\boldsymbol{g}^{(\ell)}\|_\infty\right) \le \sqrt{2^\ell}\sqrt{2\log(2N/2^\ell)} \tag{13}$$

Combining (12) and (13) leads to

$$\begin{aligned}
E\left(\sup_{\boldsymbol{x}\in\widetilde{\mathcal{S}}} \langle \hat{\boldsymbol{z}}^\ell, \boldsymbol{g}\rangle\right) &\le E\left(\sup_{\boldsymbol{x}\in\widetilde{\mathcal{S}}} \|\hat{\boldsymbol{z}}^\ell\|_1 \|\boldsymbol{g}^{(\ell)}\|_\infty\right) \le \min\{2^{-\ell}\sqrt{N}, 2\sqrt{K}\} \cdot E\left(\|\boldsymbol{g}^{(\ell)}\|_\infty\right) \\
&\le \min\{\sqrt{2^{-\ell}N}, 2\sqrt{2^\ell K}\} \cdot \sqrt{2\log(2N/2^\ell)}.
\end{aligned} \tag{14}$$

This together with (9) implies that

$$w(\widetilde{\mathcal{S}}) = E\left(\sup_{\boldsymbol{x}\in\widetilde{\mathcal{S}}} \langle \boldsymbol{x}, \boldsymbol{g}\rangle\right) \le \sqrt{2/\pi} + \sum_{\ell=1}^{L} \min\{\sqrt{2^{-\ell}N}, 2\sqrt{2^\ell K}\} \cdot \sqrt{2\log(2N/2^\ell)}.$$

Here $\sqrt{2/\pi}$ is an upper bound of the expectation of $|\langle \hat{\boldsymbol{y}}^{(L)}, \boldsymbol{g}\rangle|$ with $\hat{\boldsymbol{y}}^{(L)} = c\mathbf{1}$ and $|c| \le 1/\sqrt{N}$. Now we estimate the constant $\sum_{\ell=1}^{L} \min\{\sqrt{2^{-\ell}N}, 2\sqrt{2^{\ell}K}\} \cdot \sqrt{2\log(2N/2^{\ell})}$. Let L_0 be the maximum integer that satisfies $\sqrt{2^{-L_0}N} \geq 2\sqrt{2^{L_0}K}$, namely, $2^{L_0} \leq \sqrt{N/K}/2 \leq 2^{L_0+1}$. It is obvious that $\min\{\sqrt{2^{-\ell}N}, 2\sqrt{2^{\ell}K}\} = 2\sqrt{2^{\ell}K}$ if $\ell \leq L_0$ and $\min\{\sqrt{2^{-\ell}N}, 2\sqrt{2^{\ell}K}\} = \sqrt{2^{-\ell}N}$ otherwise. Therefore, if $N > 1$ and $K > 1$, then

$$
\begin{aligned}
&\sum_{\ell=1}^{L}\left(\min\{\sqrt{2^{-\ell}N}, 2\sqrt{2^{\ell}K}\} \cdot \sqrt{2\log(2N/2^{\ell})}\right) \\
\leq&\sqrt{2\log N}\left(2\sqrt{K}\textstyle\sum_{\ell=1}^{L_0} 2^{\frac{\ell}{2}} + \sqrt{N}\sum_{\ell=L_0+1}^{L} 2^{-\frac{\ell}{2}}\right) \\
=&\sqrt{2\log N}\left(2\sqrt{2}\sqrt{K}\frac{2^{L_0/2}-1}{\sqrt{2}-1} + \sqrt{N}2^{-\frac{L_0+1}{2}}\frac{1-2^{-\frac{L-L_0}{2}}}{1-2^{-\frac{1}{2}}}\right) \\
\leq&\sqrt{2\log N}\left(2\left(2+\sqrt{2}\right)\sqrt{K}2^{L_0/2} + (2+\sqrt{2})\sqrt{N2^{-(L_0+1)}}\right) - \sqrt{2/\pi} \\
\leq&\sqrt{2\log N}\left(2(2+\sqrt{2})\sqrt{K}\sqrt{\sqrt{N/K}/2} + (2+\sqrt{2})\sqrt{N/(\sqrt{N/K}/2)}\right) \\
&\qquad - \sqrt{2/\pi} \\
=&\left(8+4\sqrt{2}\right)(NK)^{1/4}\sqrt{\log N} - \sqrt{2/\pi}.
\end{aligned}
\tag{15}
$$

Finally, we get

$$
w(\mathcal{S}) \leq w(\widetilde{\mathcal{S}}) \leq \left(8+4\sqrt{2}\right)(NK)^{\frac{1}{4}}\sqrt{\log N}.
$$

When N is not in the form of 2^L, we can extend $\boldsymbol{x} \in \widetilde{S}$ to $\tilde{\boldsymbol{x}}$ of size $\tilde{N} = 2^L \leq 2N$ by padding zeros. Then, $\|\tilde{\boldsymbol{x}}\|_2 \leq 1$ and $\|D\tilde{\boldsymbol{x}}\|_1 \leq 4\sqrt{K}+1 \leq 4\sqrt{\tilde{K}}$ with $\tilde{K} = 25K/16$. Furthermore, $\langle \boldsymbol{x}, \boldsymbol{g}\rangle = \langle \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{g}}\rangle$, where $\tilde{\boldsymbol{g}}$ is a Gaussian vector of size $\tilde{N}$. Altogether, we have

$$
\begin{aligned}
w(\mathcal{S}) \leq w(\widetilde{\mathcal{S}}) &\leq (8+4\sqrt{2})(\tilde{N}\tilde{K})^{\frac{1}{4}} \cdot \sqrt{\log \tilde{N}} \\
&\leq 2^{\frac{5}{4}}(2\sqrt{5}+\sqrt{10})(NK)^{\frac{1}{4}} \cdot \sqrt{\log(2N)}
\end{aligned}
$$

Let us estimate the lower bound of $w(\mathcal{S})$. Let $\boldsymbol{g} \in \mathbb{R}^N$ be fixed. Let $\widetilde{L}$ be a positive number that is to be determined later. We partition $\boldsymbol{g}$ as

$$
\boldsymbol{g} = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_H, \boldsymbol{g}_{H+1}],
$$

where $H = \lfloor (N - \max\{K+1, \widetilde{L}\})/\widetilde{L} \rfloor$ and $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_H \in \mathbb{R}^{\widetilde{L}}$ and $\boldsymbol{g}_{H+1}$ is the remaining entries of $\boldsymbol{g}$. Let $\boldsymbol{a}$ be a vector that has the same size as $\boldsymbol{g}_{H+1}$. Since the length of $\boldsymbol{g}_{H+1}$ is larger than $K+1$, we can define

$$
\boldsymbol{a} = [0, \ldots, 0, (-1)^0, (-1)^1, (-1)^2, \ldots, (-1)^{K-1}] \tag{16}
$$

Therefore, the support $\mathcal{K}_0$ of $D\boldsymbol{a}$ has a cardinality K. Define

$$\boldsymbol{z} = [\nu s_1 \mathbf{1} \ \dots \ \nu s_H \mathbf{1} \ \mu \boldsymbol{a}], \tag{17}$$

where $s_i = \text{sgn}(\langle \boldsymbol{g}_i, \mathbf{1} \rangle)$ for $i = 1, \dots, H$, and μ and ν are positive numbers to be determined later. Therefore, we have $\|\boldsymbol{z}\|_2 = (\nu^2 H \widetilde{L} + \mu^2 K)^{1/2}$. In order $\|\boldsymbol{z}\|_2 = 1$, we need

$$\nu^2 H \widetilde{L} + \mu^2 K = 1. \tag{18}$$

Moreover, from the construction, $\|(D\boldsymbol{z})_{\mathcal{K}_0}\|_1 = (2K-1)\mu$, and $\|(D\boldsymbol{z})_{\mathcal{K}_0^c}\|_1 \leq (2H-1)\nu \leq (2N/\widetilde{L})\nu$. In order that $\boldsymbol{z} \in \mathcal{S}$, we should have

$$(2K-1)\mu \geq 2\nu N/\widetilde{L}. \tag{19}$$

We pick

$$\mu = 1/\sqrt{2K} \cdot \sqrt{\frac{2N}{N + H\widetilde{L}}}, \quad \nu = 1/\sqrt{2N} \cdot \sqrt{\frac{2N}{N + H\widetilde{L}}}, \quad L = \lceil 2\sqrt{N/K} \rceil. \tag{20}$$

Then, both (18) and (19) are satisfied, and hence $\boldsymbol{z} \in \mathcal{S}$. Therefore,

$$\begin{aligned} w(\mathcal{S}) = E\left(\sup_{\boldsymbol{x} \in \mathcal{S}} \langle \boldsymbol{x}, \boldsymbol{g} \rangle\right) &\geq E\left(\nu \textstyle\sum_{i=1}^{H} |\langle \boldsymbol{g}_i, \mathbf{1} \rangle| + \mu \langle \boldsymbol{g}_{H+1}, \boldsymbol{a} \rangle\right) = \nu H \cdot E(|\langle \boldsymbol{g}_1, \mathbf{1} \rangle|) + \mu E(\langle \boldsymbol{g}_{H+1}, \boldsymbol{a} \rangle) \\ = \nu H \cdot E(|\langle \boldsymbol{g}_1, \mathbf{1} \rangle|) &\geq \sqrt{\frac{2N}{N + H\widetilde{L}}} \cdot H/\sqrt{2N} \cdot \sqrt{2\sqrt{N/K}} \sqrt{2/\pi} \geq H \cdot \sqrt{2/\pi} (NK)^{-\frac{1}{4}}. \end{aligned} \tag{21}$$

It remains to find a lower bound of H. If $\widetilde{L} \geq K+1$, then $H \geq (N - \widetilde{L})/\widetilde{L} - 1 = N/\widetilde{L} - 2 \geq N/(2\sqrt{N/K} + 1) - 2 \geq \sqrt{NK}/(2 + \sqrt{K/N}) - 2$, otherwise, $H \geq (N-K-1)/\widetilde{L} - 1 \geq (N-K-1)/(2\sqrt{N/K}+1) - 1 \geq (1 - (K+1)/N)/(2 + \sqrt{K/N}) \cdot \sqrt{NK} - 1$. Therefore,

$$H \geq \frac{1 - \frac{K+1}{N}}{2 + \sqrt{K/N}} \sqrt{NK} - 2. \tag{22}$$

$\square$

Now we are ready for the proof of Part (a) of Theorem 2.1.

Proof of Part (a) of Theorem 2.1. Let $\mathcal{Y}$ be the null space of A. Then $\mathcal{Y}$ satisfies the random model in Theorem 3.2 [6, 37, 40]. Besides, $\mathcal{S}$ in (8) is a subset of the unit Euclidean ball. Now, if we choose $M \geq 2C_3^2 \sqrt{NK} \log(2N)$ with $C_3 = 2^{\frac{5}{4}}(2\sqrt{5} + \sqrt{10})$, then, by Theorem 3.3, $w(\mathcal{S}) \leq \sqrt{M/2} < \sqrt{M} - \frac{1}{2\sqrt{M}}$. This together with Theorem 3.2 implies that the probability $\mathcal{Y}\bigcap\mathcal{S}=\emptyset$ is larger than

$$1-3.5e^{-\frac{\sqrt{M}-\frac{1}{2\sqrt{M}}-w(\mathcal{S})}{18}}=1-3.5e^{\frac{1}{36\sqrt{M}}}e^{-\frac{\sqrt{M}-w(\mathcal{S})}{18}}\geq 1-3.5e^{\frac{1}{36}}e^{-\frac{(1-1/\sqrt{2})\sqrt{M}}{18}}\equiv 1-C_1e^{-C_2\sqrt{M}}.$$

According to Lemma 3.1, $\mathcal{Y}\bigcap\mathcal{S}=\emptyset$ leads to $\hat{\boldsymbol{x}}=\bar{\boldsymbol{x}}$ for all $\bar{\boldsymbol{x}}$ satisfying $\|D\bar{\boldsymbol{x}}\|_0\leq K$, which concludes the proof. □

3.2 Bound for unsuccessful recovery

In this section, we give the bound of M for the failure of (2) in recovery of $\bar{\boldsymbol{x}}$ with $\|D\bar{\boldsymbol{x}}\|_0\leq K$, i.e., we prove Part (b) of Theorem 2.1. Our argument uses a result in [1] (c.f. Theorem 3.4 below) and the lower bound of the Gaussian width of the descent cone for (2).

To describe the success of convex optimization for linear inverse problems, [1] introduced the descent cone (applied to (2)) as follows

$$\mathcal{D}(\boldsymbol{x})=\cup_{\tau>0}\{\boldsymbol{z}\mid \|D(\boldsymbol{x}+\tau\boldsymbol{z})\|_1\leq\|D(\boldsymbol{x})\|_1\}.$$

To measure the success rate, one may use the Gaussian width $w(\mathcal{D}(\boldsymbol{x})\cap\mathbb{S}^{N-1})$, where $\mathbb{S}^{N-1}$ is the unit Euclidean sphere in $\mathbb{R}^N$.

Theorem 3.4 (A corollary of Proposition 10.1 and Theorem II in [1])**.** *Fix a tolerance $\eta\in(0,1)$. Let $\bar{\boldsymbol{x}}\in\mathbb{R}^N$ be a fixed vector. Suppose $A\in\mathbb{R}^{M\times N}$ has independent standard normal entries and $\boldsymbol{y}=A\bar{\boldsymbol{x}}$. Set the quantity $a_\eta=4\sqrt{\log(4/\eta)}$. Then, as long as $M\leq w^2(\mathcal{D}(\bar{\boldsymbol{x}})\cap\mathbb{S}^{N-1})-a_\eta\sqrt{N}$, $\bar{\boldsymbol{x}}$ is not a solution of* (2) *with probability at least $1-\eta$.*

Now let us prove Part (b) of Theorem 2.1.

Proof of Part (b) of Theorem 2.1. For a fixed $\bar{\boldsymbol{x}}\in\mathbb{R}^N$ with gradient support $\mathcal{K}_0:=\operatorname{supp}(D\bar{\boldsymbol{x}})$ satisfying $|\mathcal{K}_0|\leq K$, it is easy to check that the descent cone is

$$\mathcal{D}(\bar{\boldsymbol{x}})=\{\boldsymbol{z}\mid\exists\,\tau>0,\ \text{s.t. }\tau\|(D\boldsymbol{z})_{\mathcal{K}_0}^c\|_1+\|(D(\bar{\boldsymbol{x}}+\tau\boldsymbol{z}))_{\mathcal{K}_0}\|_1\leq\|(D\bar{\boldsymbol{x}})_{\mathcal{K}_0}\|_1\}.$$

We shall estimate a lower bound of the Gaussian width $w(\mathcal{D}(\bar{\boldsymbol{x}})\cap\mathbb{S}^{N-1})$, by using a similar argument from the proof of the lower bound in Theorem 3.3.

Choose $\bar{\boldsymbol{x}}$ such that its gradient support $\mathcal{K}_0 = \{N-K, N-K+1, \ldots, N-1\}$ and $\text{sgn}((D\bar{\boldsymbol{x}})_{\mathcal{K}_0}) = [-1, 1, -1, 1, \ldots]^T$. In other words, components of $\bar{\boldsymbol{x}}$ keep constant until $N-K$ and then they decrease and increase alternatively. Define $\boldsymbol{z}$ by (16), (17), and (20). It can be seen that $\text{sgn}((D\bar{\boldsymbol{x}})_{\mathcal{K}_0}) = -\text{sgn}((D\boldsymbol{z})_{\mathcal{K}_0})$, and, consequently, for a sufficiently small $\tau > 0$, we have $\|(D(\bar{\boldsymbol{x}} + \tau\boldsymbol{z}))_{\mathcal{K}_0}\|_1 = \|D(\bar{\boldsymbol{x}})_{\mathcal{K}_0}\|_1 - \tau\|(D\boldsymbol{z})_{\mathcal{K}_0}\|_1$. This together with (18) and (19) implies that $\boldsymbol{z} \in \mathcal{D}(\bar{\boldsymbol{x}}) \cap \mathbb{S}^{N-1}$. An estimation similar to (21) and (22) leads to

$$w^2(\mathcal{D}(\bar{\boldsymbol{x}}) \cap \mathbb{S}^{N-1}) \geq \frac{2}{\pi}\left(\frac{3}{10}(NK)^{\frac{1}{4}} - 2(NK)^{-\frac{1}{4}}\right)^2 \geq \frac{9\sqrt{K}}{50\pi}\sqrt{N} - \frac{12}{5\pi},$$

provided $K+1 \leq N/4$. Therefore, if K is large enough such that $a_\eta < \frac{9\sqrt{K}}{50\pi}$, then by Theorem 3.4, as long as $M \leq \left(\frac{9}{50\pi} - \frac{a_\eta}{\sqrt{K}}\right)\sqrt{KN} - \frac{12}{5\pi}$, (2) fails to find $\bar{\boldsymbol{x}}$ with probability larger than $1-\eta$. □

4 Proof of Theorem 2.2

This section is to prove Theorem 2.2. We first present some necessary lemmas, and then the main proof is brought in Section 4.3.

4.1 Balanced Condition

In this subsection, we prove the balanced condition for $\|D\boldsymbol{x}\|_1$ stated as in below.

Lemma 4.1 (Balanced Condition)**.** *Let $A \in \mathbb{R}^{M\times N}$ be a random Gaussian matrix whose entries are drawn from independent standard normal distributions. Then, for any constant $0 < \alpha < 1$ and $0 < C < 1$, there exist positive constants δ, C_0, and C_1 such that the following statement holds true, with probability at least $1 - C_0 e^{-C_1 N}$: For all subsets $\mathcal{K} \subseteq \{1, 2, ..., N-1\}$ with cardinality $|\mathcal{K}| \leq \delta N$, and for every nonzero vector $\boldsymbol{x}$ in the null space of A (namely $A\boldsymbol{x} = 0$, $\boldsymbol{x} \neq \mathbf{0}$) with $M = \alpha N$,*

$$\|(D\boldsymbol{x})_{\mathcal{K}}\|_1 < C\|(D\boldsymbol{x})_{\mathcal{K}^c}\|_1, \tag{23}$$

where $\mathcal{K}^c = \{1, 2, ..., N-1\} \setminus \mathcal{K}$.

To prove this lemma, we need several lemmas.

Lemma 4.2. *Let $H \in \mathbb{R}^{N\times(N-M)}$ be a random matrix whose entries follow independent standard Gaussian distributions. Then for any $C_0 > 0$, $\gamma > 0$ and $\alpha > 0$, there exists $\delta > 0$ and $C_1, C_2 > 0$, such that, for*

$M = \alpha N$,

$$P\left(\forall \|\boldsymbol{z}\|_2 = 1 \ \textit{and}\ |\mathcal{K}| \le \delta N, \quad \|(DH\boldsymbol{z})_{\mathcal{K}}\|_1 \le \frac{C_0\gamma}{2} N\right) \ge 1 - C_1 e^{-C_2 N}. \tag{24}$$

Proof. For a fixed $|\mathcal{K}| \le \delta N$, there exists a set $\mathcal{K}'$ satisfying $|\mathcal{K}'| \le 2\delta N$ such that, for any $\boldsymbol{z}$ satisfying $\|\boldsymbol{z}\|_2 = 1$,

$$\|(DH\boldsymbol{z})_{\mathcal{K}}\|_1 \le 2\|(H\boldsymbol{z})_{\mathcal{K}'}\|_1 = 2\|H_{\mathcal{K}'}\boldsymbol{z}\|_1 \le 2\sqrt{2\delta N}\|H_{\mathcal{K}'}\boldsymbol{z}\|_2 \le 2\sqrt{2\delta N}\|H_{\mathcal{K}'}\|_2, \tag{25}$$

where $H_{\mathcal{K}'}$ are the rows of H indexed by $\mathcal{K}'$. From [43, Corollary 5.35], we know that, for any $\theta > 0$,

$$P\left(\|H_{\mathcal{K}'}\|_2 \le (1+\theta)(\sqrt{2\delta} + \sqrt{1-\alpha})\sqrt{N}\right) \ge 1 - 2\cdot \exp(-\theta^2(2\delta + 1 - \alpha + 2\sqrt{2\delta(1-\alpha)})N/2).$$

Let $\boldsymbol{x} = H\boldsymbol{z}$. Then, together with (25), we get

$$P\left(\|(D\boldsymbol{x})_{\mathcal{K}}\|_1 \le 2(1+\theta)(2\delta + \sqrt{2\delta(1-\alpha)})N\right) \ge 1 - 2\cdot \exp(-\theta^2(2\delta + 1 - \alpha + 2\sqrt{2\delta(1-\alpha)})N/2).$$

Therefore, the uniform probability

$$\begin{aligned} P\big(\forall\ \|\boldsymbol{z}\|_2 = 1 \text{ and } |\mathcal{K}| \le \delta N, \quad \|(D\boldsymbol{x})_{\mathcal{K}}\|_1 \le 2(1+\theta)(2\delta + \sqrt{2\delta(1-\alpha)})N\big) \ge \\ 1 - 2\cdot \binom{N}{\delta N} \cdot \exp(-\theta^2(2\delta + 1 - \alpha + 2\sqrt{2\delta(1-\alpha)})N/2) \end{aligned}$$

By Stirling's formula,

$$\begin{aligned} \binom{N}{\delta N} &= \frac{N!}{((1-\delta)N)!(\delta N)!} \le \frac{\frac{e}{\sqrt{2\pi}}\sqrt{2\pi N}\left(\frac{N}{e}\right)^N}{\sqrt{2\pi(1-\delta)N}\left(\frac{(1-\delta)N}{e}\right)^{(1-\delta)N}\sqrt{2\pi\delta N}\left(\frac{\delta N}{e}\right)^{\delta N}} \\ &\le \frac{e}{\pi\sqrt{N}}\frac{1}{\sqrt{\delta(1-\delta)}}\left(\frac{1}{(\delta)^\delta(1-\delta)^{1-\delta}}\right)^N \end{aligned}$$

Since the function $\delta\log(1/\delta) + (1-\delta)\log(1/(1-\delta)) \le \log 2$, we choose $\theta = \sqrt{\frac{2\log 2}{1-\alpha}}$ and therefore

$$\delta\log(1/\delta) + (1-\delta)\log(1/(1-\delta)) \le \log 2 = \theta^2(1-\alpha)/2 < \theta^2(2\delta + 1 - \alpha + 2\sqrt{2\delta(1-\alpha)})/2.$$

As $2(1+\theta)(2\delta + \sqrt{2\delta(1-\alpha)}) \to 0$ when $\delta \to 0$, we can choose a small $\delta > 0$ such that $2(1+\theta)(2\delta + \sqrt{2\delta(1-\alpha)}) \le \frac{C_0\gamma}{2}$, and we obtain (24). □

Lemma 4.3. *Let $H \in \mathbb{R}^{N\times(N-M)}$ be a random matrix whose entries follow independent standard Gaussian* *distributions* $\mathcal{N}(0,1)$. *Then for any* $\alpha > 0$, *there exists* $C_0, C_1, C_2 > 0$, *such that, for* $M = \alpha N$,

$$P\left(\forall\ \|\boldsymbol{z}\|_2 = 1, \quad \|DH\boldsymbol{z}\|_1 \leq C_0 N\right) \geq 1 - C_1 e^{-C_2 N}. \tag{26}$$

Proof. By [43, Corollary 5.35], for any $\theta > 0$, we have

$$P(\|H\|_2 \leq (1+\theta)(1+\sqrt{1-\alpha})\sqrt{N}) \geq 1 - 2\cdot\exp(-\theta^2(2-\alpha+2\sqrt{1-\alpha})N/2),$$

Moreover,

$$\|DH\boldsymbol{z}\|_1 \leq 2\|H\boldsymbol{z}\|_1 \leq 2\sqrt{N}\|H\boldsymbol{z}\|_2 \leq 2\sqrt{N}\|H\|_2$$

Therefore,

$$P\big(\forall\ \|\boldsymbol{z}\|_2 = 1, \quad \|DH\boldsymbol{z}\|_1 \leq 2(1+\theta)(1+\sqrt{1-\alpha})N\big) \geq 1 - 2\cdot\exp(-\theta^2(2-\alpha+2\sqrt{1-\alpha})N/2).$$

□

Lemma 4.4. *Let* $H \in \mathbb{R}^{N\times(N-M)}$ *be a random matrix whose entries follow independent standard Gaussian distributions* $\mathcal{N}(0,1)$. *Then, for any* $0 < \alpha < 1$, *there exist constants* $\gamma > 0$, $C_0 > 0$, *and* $C_1 > 0$, *such that, for* $M = \alpha N$,

$$P\left(\forall\ \|\boldsymbol{z}\|_2 = 1, \quad \|DH\boldsymbol{z}\|_1 \geq \gamma N\right) \geq 1 - C_0 e^{-C_1 N}. \tag{27}$$

Proof. For any fixed $\boldsymbol{z} \in \mathbb{R}^{N-M}$ satisfying $\|\boldsymbol{z}\|_2 = 1$, $\boldsymbol{x} = H\boldsymbol{z}$ is a random vector whose entries follows independent standard normal distribution. Then $\boldsymbol{y} := DH\boldsymbol{z}$ follows a multivariate Gaussian $\mathcal{N}(\boldsymbol{0}, \sqrt{DD^T})$, and its probability density function is

$$f(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^{N-1}\cdot\det(DD^T)}}\exp(-\boldsymbol{y}^T(DD^T)^{-1}\boldsymbol{y}/2) = \frac{1}{\sqrt{(2\pi)^{N-1}N}}\exp(-\boldsymbol{y}^T(DD^T)^{-1}\boldsymbol{y}/2),$$

where we have used the fact that $\det(DD^T) = N$. Since DD^T is symmetric positive definite, we have $\exp(-\boldsymbol{y}^T(DD^T)^{-1}\boldsymbol{y}/2) \leq 1$. Thus, $f(\boldsymbol{y}) \leq \frac{1}{\sqrt{(2\pi)^{N-1}N}}$. Let θ be a positive number determined later. We have, by Markov inequality,

$$\begin{aligned}P\left(\|D\boldsymbol{x}\|\leq\gamma N\right)=&P\left(\sum_{i=1}^{N-1}|y_i|\leq\gamma N\right)=P\left(-\theta\sum_{i=1}^{N-1}|y_i|\geq-\theta\gamma N\right)\\=&P\left(e^{-\theta\sum\limits_{i=1}^{N-1}|y_i|}\geq e^{-\theta\gamma N}\right)\leq e^{\theta\gamma N}\cdot E(e^{-\theta\sum_{i=1}^{N-1}|y_i|}).\end{aligned}$$

The expectation can be estimated from below

$$\begin{aligned}E(e^{-\theta\sum_{i=1}^{N-1}|y_i|})&=\int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty}e^{-\theta\sum_{i=1}^{N-1}|y_i|}f(\boldsymbol{y})dy_1\dots dy_{N-1}\\&\leq\frac{1}{\sqrt{(2\pi)^{N-1}N}}\int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty}e^{-\theta\sum_{i=1}^{N-1}|y_i|}dy_1\dots dy_{N-1}\\&=\frac{1}{\sqrt{(2\pi)^{N-1}N}}\left(\int_{-\infty}^{+\infty}e^{-\theta|y_1|}dy_1\right)^{N-1}\\&=\frac{1}{\sqrt{(2\pi)^{N-1}N}}\left(\int_{-\infty}^{0}e^{\theta y_1}dy_1+\int_{0}^{+\infty}e^{-\theta y_1}dy_1\right)^{N-1}=\frac{1}{\sqrt{(2\pi)^{N-1}N}}\left(\frac{2}{\theta}\right)^{N-1}\end{aligned}$$

from which we derive

$$P\left(\|D\boldsymbol{x}\|_1\leq\gamma N\right)\leq e^{\theta\gamma N}\frac{1}{\sqrt{(2\pi)^{N-1}N}}\left(\frac{2}{\theta}\right)^{N-1}.$$

Choosing $\theta=1/\gamma$ yields

$$P\left(\|D\boldsymbol{x}\|_1\leq\gamma N\right)\leq\frac{e}{\sqrt{N}}\cdot\left(\frac{2\gamma e}{\sqrt{2\pi}}\right)^{N-1}. \tag{28}$$

Now we let $\boldsymbol{z}$ vary on the sphere $\{\boldsymbol{z}|\ \ \|\boldsymbol{z}\|_2=1\}$. We cover the sphere with ϵ-net, where $\epsilon=C_2\gamma$, $C_2>0$ are constants we will choose later. ϵ-net is a finite set $V=\{\boldsymbol{v}_1,...,\boldsymbol{v}_L\}$ on $\{\boldsymbol{z}|\ \|\boldsymbol{z}\|_2=1\}$ such that every point $\boldsymbol{z}$ from $\{\boldsymbol{z}|\ \|\boldsymbol{z}\|_2=1\}$, there is a $\boldsymbol{v}_l\in V$ such that $\|\boldsymbol{z}-\boldsymbol{v}_l\|_2\leq\epsilon$. It is known that the size of the ϵ-net can be taken no bigger than $(1+\frac{2}{\epsilon})^{N-M}$ [27]. For any $\boldsymbol{z}$ such that $\|\boldsymbol{z}\|_2=1$, there exists a point $\boldsymbol{v}_0$ (we change the subscript numbering for V to index the order) in V such that $\|\boldsymbol{z}-\boldsymbol{v}_0\|_2\triangleq\epsilon_1\leq\epsilon$. Let $\boldsymbol{z}_1$ denote $\boldsymbol{z}-\boldsymbol{v}_0$, then $\|\boldsymbol{z}_1-\epsilon_1\boldsymbol{v}_1\|_2\triangleq\epsilon_2\leq\epsilon_1\epsilon\leq\epsilon^2$ for some $\boldsymbol{v}_1$ in V. Repeating this process, we have $\boldsymbol{z}=\sum_{j\geq0}\epsilon_j\boldsymbol{v}_j$, where $\epsilon_0=1$, $\epsilon_j\leq\epsilon^j$ and $\boldsymbol{v}_j\in V$. Then

$$\begin{aligned}\|D(H\boldsymbol{z})\|_1=\|D\sum_{j\geq0}(\epsilon_jH\boldsymbol{v}_j)\|_1&\geq\|(D(H\boldsymbol{v}_0)\|_1-\sum_{j\geq1}\epsilon_j\|D(H\boldsymbol{v}_j)\|_1\\&\geq\|(D(H\boldsymbol{v}_0)\|_1-\sum_{j\geq1}\epsilon^j\|D(H\boldsymbol{v}_j)\|_1\geq\gamma N-\frac{\epsilon}{1-\epsilon}\times C_3N\geq\left(\gamma-\frac{C_2C_3\gamma}{1-C_2C_3\gamma}\right)N\end{aligned} \tag{29}$$

where the first inequality follows from the triangle inequality, and the last second inequality follows from

Lemma 4.3. Furthermore, the probability that (29) fails is at most $C_4 e^{-C_5 N}$. If we choose C_2 such that $C_2 C_3 = \frac{1}{3}$, then, when $\gamma < 1$, we have

$$P\left(\forall \ \|\boldsymbol{z}\|_2 = 1, \ \ \|DH\boldsymbol{z}\|_1 \geq \frac{\gamma}{2} N\right) \geq 1 - C_4 e^{-C_5 N} - P,$$

where $P = \left(1 + \frac{2}{\epsilon}\right)^{N-M} \cdot \frac{e}{\sqrt{N}} \cdot \left(\frac{2\gamma e}{\sqrt{2\pi}}\right)^{N-1}$. Let $1 > \mu > 1 - \alpha$ be a constant. We then have

$$\begin{aligned}
P &= \left(1 + \frac{2}{C_2 \gamma}\right)^{(1-\alpha)N} \frac{\frac{\sqrt{2\pi}}{2\gamma}}{\sqrt{N}} \cdot \left(\frac{2\gamma e}{\sqrt{2\pi}}\right)^{N} = \frac{\frac{\sqrt{2\pi}}{2\gamma}}{\sqrt{N}} \left(\frac{2e}{\sqrt{2\pi}}\right)^{N} \gamma^{(1-\mu)N} \left(\gamma^{\frac{\mu}{1-\alpha}} + \frac{2}{C_2}\gamma^{\frac{\mu}{1-\alpha}-1}\right)^{(1-\alpha)N} \\
&\leq \frac{\frac{\sqrt{2\pi}}{2\gamma}}{\sqrt{N}} \left(\frac{2e}{\sqrt{2\pi}}\right)^{N} \gamma^{(1-\mu)N} \left(1 + \frac{2}{C_2}\right)^{(1-\alpha)N} = \frac{\frac{\sqrt{2\pi}}{2\gamma}}{\sqrt{N}} \left[\left(\frac{2e}{\sqrt{2\pi}}\right)^{\frac{1}{1-\mu}} \gamma \left(1 + \frac{2}{C_2}\right)^{\frac{1-\alpha}{1-\mu}}\right]^{(1-\mu)N} \\
&\equiv \frac{\frac{\sqrt{2\pi}}{2\gamma}}{\sqrt{N}} \left(C_6 \gamma\right)^{(1-\mu)N},
\end{aligned}$$

Therefore, by choosing γ such that $C_7 \equiv C_6 \gamma < 1$ and defining $C_8 \equiv \frac{\sqrt{2\pi}}{2\gamma}$, we obtain

$$P\left(\forall \ \|\boldsymbol{z}\|_2 = 1, \ \ \|DH\boldsymbol{z}\|_1 \geq \frac{\gamma}{2} N\right) \geq 1 - C_4 e^{-C_5 N} - \frac{C_8}{\sqrt{N}} C_7^{(1-\mu)N}.$$

By choosing proper positive constants C_0 and C_1, we have $C_4 e^{-C_5 N} + \frac{C_8}{\sqrt{N}} C_7^{(1-\mu)N} \leq C_0 e^{-C_1 N}$, which concludes the proof. □

We are now ready to prove Lemma 4.1.

Proof of Lemma 4.1. It is well known that the distribution of vectors in the null space of A is the same as $\boldsymbol{x} = H\boldsymbol{z}$, where $H \in \mathbb{R}^{N \times (N-M)}$ is a random Gaussian matrix [6, 45]. Notice that the $C\|(D\boldsymbol{x})_{\mathcal{K}^c}\|_1 > \|(D\boldsymbol{x})_{\mathcal{K}}\|_1$ is invariant of the length of $\boldsymbol{x}$. Therefore, we can fix the length of $\boldsymbol{x}$ (consequently $\boldsymbol{z}$). Without loss of generality, we consider $\|\boldsymbol{z}\|_2 = 1$. By Lemma 4.4, we can find a γ such that, with overwhelming probability,

$$\|D\boldsymbol{x}\|_1 \geq \gamma N. \tag{30}$$

Fixing this γ, by Lemma 4.2, we can find $\delta > 0$ such that, with overwhelming probability,

$$\|(D\boldsymbol{x})_{\mathcal{K}}\|_1 \leq \frac{C\gamma}{2} N. \tag{31}$$

Altogether, we have $C\|D\boldsymbol{x}\|_1 \geq 2\|(D\boldsymbol{x})_{\mathcal{K}}\|_1 > (1+C)\|(D\boldsymbol{x})_{\mathcal{K}}\|_1$, which implies $C\|(D\boldsymbol{x})_{\mathcal{K}^c}\|_1 > \|(D\boldsymbol{x})_{\mathcal{K}}\|_1$.

The probability that (30) or (31) fails is at most $C_0 e^{-C_1 N}$. □

4.2 Almost Euclidean Property

Lemma 4.5 (Almost Euclidean Property)**.** *Suppose that the measurement matrix A is an $M \times N$ matrix having i.i.d. standard zero mean Gaussian elements. For any constant $0 < \alpha < 1$, there exists constants $\beta > 0$, $C_0 > 0$, and $C_1 > 1$, such that the following statement holds true, with probability at least $1 - C_0 e^{-C_1 N}$: For every nonzero vector $\boldsymbol{x}$ in the null space of A (namely $A\boldsymbol{x} = 0$, $\boldsymbol{x} \neq \mathbf{0}$) with $M = \alpha N$,*

$$\|D\boldsymbol{x}\|_1 \geq \beta\sqrt{N}\|\boldsymbol{x}\|_2. \tag{32}$$

Proof. The distribution of vectors in the null space of A is the same as $\boldsymbol{x} = H\boldsymbol{z}$, where $H \in \mathbb{R}^{N \times (M-N)}$ is a random Gaussian matrix. Because (32) is invariant of scales, we assume $\|\boldsymbol{z}\|_2 = 1$ without loss of generality. By Lemma 4.4, there exists a constant $\gamma > 0$, such that, with overwhelming probability,

$$\|D\boldsymbol{x}\|_1 \geq \gamma N. \tag{33}$$

Since $\boldsymbol{x}$ follows an independent Gaussian distribution, then, with probability at least $1 - 2e^{-\theta^2 N}$,

$$\|\boldsymbol{x}\|_2 \leq (1+\theta)\sqrt{N}. \tag{34}$$

When both (33) and (34) happens, we have (32) with $\beta = \gamma/(1+\theta)$. The probability that (33) or (34) fails is at most $2e^{-\theta^2 N} + C_2 e^{-C_3 N}$, which can be bounded by $C_0 e^{-C_1 N}$ for some positive constants C_0 and C_1. □

4.3 Proof of Theorem 2.2

Proof of Theorem 2.2. Let $0 < C < 1$ and $0 < \alpha < 1$ be given and δ be the constant δ found in Lemma 4.1. Let $\mathcal{K} \subseteq \{1, 2, ..., N-1\}$ be a minimizer of $\min_{|\mathcal{K}| \leq \delta N} \|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1$. Let $\boldsymbol{w} = \bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}$, and we decompose it orthogonally as $\boldsymbol{w} = \boldsymbol{w}_1 + \boldsymbol{w}_2$, where $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are in the null space of A and the range of A^T respectively. Then, we have

$$\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}\|_2 \leq \|\boldsymbol{w}_1\|_2 + \|\boldsymbol{w}_2\|_2. \tag{35}$$

Since $\boldsymbol{w}_2$ is in the range of A^T,

$$\|\boldsymbol{w}_2\|_2 \leq \frac{1}{\sigma_{min}(A^T)}\|A\boldsymbol{w}_2\|_2 = \frac{1}{\sigma_{min}(A^T)}\|A\boldsymbol{w}\|_2 \leq \frac{1}{\sigma_{min}(A^T)}(\|A\bar{\boldsymbol{x}}-\boldsymbol{y}\|_2+\|A\hat{\boldsymbol{x}}-\boldsymbol{y}\|_2) \leq \frac{2}{\sigma_{min}(A^T)}\epsilon. \quad (36)$$

Since $\boldsymbol{w}_1$ is in the kernel of A, by Lemma 4.5, we have

$$\|\boldsymbol{w}_1\|_2 \leq \frac{1}{\beta\sqrt{N}}\|D\boldsymbol{w}_1\|_1 \quad (37)$$

with overwhelming probability. Let us estimate $\|D\boldsymbol{w}_1\|_1$. The minimality of $\|D\hat{\boldsymbol{x}}\|_1$ implies

$$\begin{aligned}\|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1 + \|(D\bar{\boldsymbol{x}})_{\mathcal{K}}\|_1 = \|D\bar{\boldsymbol{x}}\|_1 \geq \|D\hat{\boldsymbol{x}}\|_1 = \|D\bar{\boldsymbol{x}} + D\boldsymbol{w}\|_1 = \|(D\bar{\boldsymbol{x}} + D\boldsymbol{w})_{\mathcal{K}}\|_1 + \|(D\bar{\boldsymbol{x}} + D\boldsymbol{w})_{\mathcal{K}^c}\|_1 \\ \geq \|(D\bar{\boldsymbol{x}})_{\mathcal{K}}\|_1 - \|(D\boldsymbol{w})_{\mathcal{K}}\|_1 + \|(D\boldsymbol{w})_{\mathcal{K}^c}\|_1 - \|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1,\end{aligned}$$

which leads to

$$2\|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1 + \|(D\boldsymbol{w})_{\mathcal{K}}\|_1 \geq \|(D\boldsymbol{w})_{\mathcal{K}^c}\|_1.$$

Therefore,

$$\begin{aligned}2\|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1+&\|(D\boldsymbol{w}_1)_{\mathcal{K}}\|_1 + \|(D\boldsymbol{w}_2)_{\mathcal{K}}\|_1 \\ &\geq 2\|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1 + \|(D\boldsymbol{w})_{\mathcal{K}}\|_1 \geq \|(D\boldsymbol{w})_{\mathcal{K}^c}\|_1 \geq \|(D\boldsymbol{w}_1)_{\mathcal{K}^c}\|_1 - \|(D\boldsymbol{w}_2)_{\mathcal{K}^c}\|_1,\end{aligned}$$

and thus

$$\begin{aligned}\|(D\boldsymbol{w}_1)_{\mathcal{K}^c}\|_1 &\leq 2\|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1 + \|D\boldsymbol{w}_2\|_1 + \|(D\boldsymbol{w}_1)_{\mathcal{K}}\|_1 \leq 2\|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1 + 2\|\boldsymbol{w}_2\|_1 + \|(D\boldsymbol{w}_1)_{\mathcal{K}}\|_1 \\ &\leq 2\|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1 + \frac{4}{\sigma_{\min}(A^T)}\sqrt{N}\epsilon + \|(D\boldsymbol{w}_1)_{\mathcal{K}}\|_1.\end{aligned} \quad (38)$$

Moreover, by Lemma 4.1,

$$\|(D\boldsymbol{w}_1)_{\mathcal{K}}\|_1 \leq C\|(D\boldsymbol{w}_1)_{\mathcal{K}^c}\|_1$$

with overwhelming probability. This together with (38) implies

$$\|(D\boldsymbol{w}_1)_{\mathcal{K}}\|_1 \leq 2C\|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1 + \frac{4C}{\sigma_{\min}(A^T)}\sqrt{N}\epsilon + C\|(D\boldsymbol{w}_1)_{\mathcal{K}}\|_1$$

and further

$$\|(D\boldsymbol{w}_1)_{\mathcal{K}}\|_1 \leq \frac{2C}{1-C}\|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1 + \frac{4C}{(1-C)\sigma_{\min}(A^T)}\sqrt{N}\epsilon.$$

Substituting it into (38) again yields

$$\begin{aligned}\|(D\boldsymbol{w}_1)_{\mathcal{K}^c}\|_1 &\leq (2+\frac{2C}{1-C})\|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1 + \left(\frac{4}{\sigma_{\min}(A^T)} + \frac{4C}{(1-C)\sigma_{\min}(A^T)}\right)\sqrt{N}\epsilon \\ &= \frac{2}{1-C}\|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1 + \frac{4}{(1-C)\sigma_{\min}(A^T)}\sqrt{N}\epsilon.\end{aligned}$$

We obtain

$$\|D\boldsymbol{w}_1\|_1 = \|(D\boldsymbol{w}_1)_{\mathcal{K}}\|_1 + \|(D\boldsymbol{w}_1)_{\mathcal{K}^c}\|_1 \leq \frac{2(1+C)}{(1-C)}\|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1 + \frac{4(1+C)}{(1-C)\sigma_{\min}(A^T)}\sqrt{N}\epsilon. \tag{39}$$

Finally, combine (35), (36), (37), and (39) and get

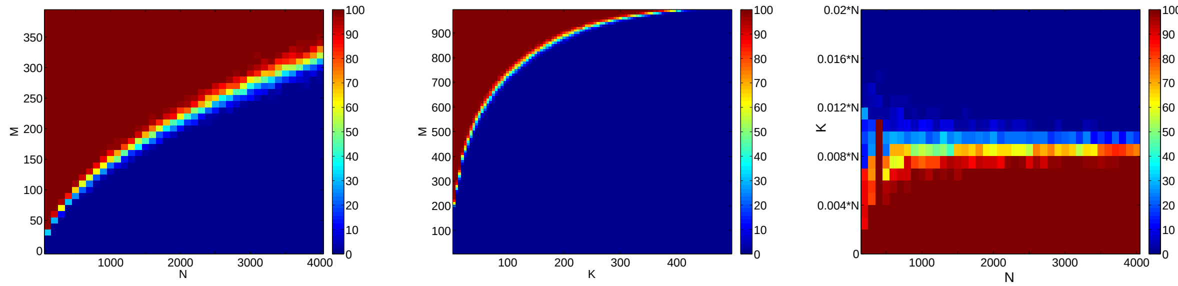
$$\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}\|_2 \leq \left(\frac{2(1+C)}{\beta(1-C)}\right)\frac{\|(D\bar{\boldsymbol{x}})_{\mathcal{K}^c}\|_1}{\sqrt{N}} + \left(2 + \frac{4(1+C)}{\beta(1-C)}\right)\frac{\epsilon}{\sigma_{\min}(A^T)}.$$

To conclude the proof, we use the well-known fact that the minimum non-zero singular values of an $M \times N$ Gaussian random matrix is the order of $\sqrt{N} - \sqrt{M} = (1-\sqrt{\alpha})\sqrt{N}$. By [43, Corollary 5.35], the probability that $\sigma_{\min}(A^T) < \frac{1-\sqrt{\alpha}}{2}\sqrt{N}$ is at most $2e^{-(1-\sqrt{\alpha})^2 N/8}$. The probabilities that other analysis above fails are all in the form of $C_0 e^{-C_1 N}$, whose sum is again in the form of $C_0 e^{-C_1 N}$ with some other positive constants C_0 and C_1. □

5 Numerical experiments

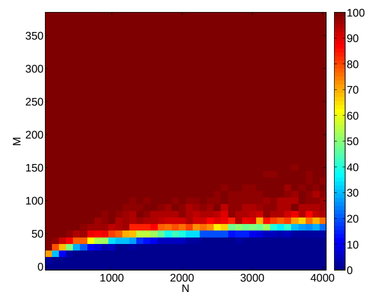
In this section, we demonstrate empirical bounds under different settings of M, N, and K, through numerical experiments.

We generate the true signal $\bar{\boldsymbol{x}} \in \mathbb{R}^N$ with K-sparse gradient as follows. The entries of $\bar{\boldsymbol{x}}$ keep to be 1 from $\bar{x}_1$ through $\bar{x}_{N-K}$, and then the values of $\bar{x}_{N-K+1}$ through $\bar{x}_N$ alternate between -1 and 1. The entries of $A \in \mathbb{R}^{M\times N}$ are drawn from i.i.d. standard normal distribution. The minimization (2) are solved by the split Bregman algorithm [3, 20]. For one set of parameters (M, N, K), we test the TV compressed sensing with 100 realizations of A, and plot the rate of successful recovery. The results are shown in Figure 1. Figure 1(a) demonstrates the successful recovery rate when we fix $K = 5$ and variate M and N. It shows that the recovery threshold of M is of the order $\sqrt{N}$, which is evidenced by the fact that the threshold bound doubles when N quadruples (e.g. the threshold bound is about 150 and 300 respectively when N is 1000 and 4000 respectively). This confirms our result in Theorem 2.1 up to a logarithm factor. In Figure 1(b),



(a) $K = 5$ is fixed. (b) $N = 1000$ is fixed. (c) $M = 0.2 * N$ is fixed.

Figure 1: Simulation results when $\bar{\boldsymbol{x}}$ is fixed and particularly chosen.



(a) $K = 5$ is fixed.

Figure 2: Simulation results when the support of $D\bar{\boldsymbol{x}}$ is distributed uniformly at random.

we fix $N = 1000$ and test different M and K. We see that the phase transition happens at the narrow band centered at $O(\sqrt{NK})$, which also supports our result in Theorem 2.1 up to a log factor. In Figure 1(c), we fix $M = 0.2N$ and let K and N change. It can be seen clearly that the maximum K that guarantees the exact recovery is proportional to N, which verifies the result of noiseless case in Theorem 2.2.

The experiments above are for a specific signal, which reflects the performance of TV-CS in the worst case. Next we demonstrate the average performance of TV-CS. We generate the true signal $\bar{\boldsymbol{x}} \in \mathbb{R}^N$ by a ± 1-valued signal, whose support of gradient $D\bar{\boldsymbol{x}}$ is chosen randomly according to the uniform distribution. All other settings are the same as those in the previous paragraph. Fig. 2(a) depicts the success rate with different pairs of (M, N) while $K = 5$ is fixed. We see that the values of M for phase transition increases very slowly with respect to N, compared to Fig. 1(a). By a further numerical examination, we find that a logarithmic function fits M for phase transition against N very well. This indicates that the recovery threshold of M for average signals is likely to follow typical bound $O(K \log(N/K))$ in compressed sensing.

6 Extensions to Multidimensional Signals

In this section, we extend our results from one-dimensional signal to d-dimensional ($d \geq 2$, for example $d = 2$ for image and $d = 3$ for videos) signal vectors, and get results that are comparable to those in [31, 32].

In particular, let $\bar{\boldsymbol{X}} \in \mathbb{R}^{N^d}$ be a multi-indexed vector that is from a d-dimensional signal. Let $A \in \mathbb{R}^{M \times N^d}$ be a measurement matrix whose elements are i.i.d. Gaussian random variables, and $\boldsymbol{Y} = A\bar{\boldsymbol{X}}$ be its corresponding measurements of $\bar{\boldsymbol{X}}$. Define $D\bar{\boldsymbol{X}}$ be the discrete gradient of $\bar{\boldsymbol{X}}$. Assume that $D\bar{\boldsymbol{X}}$ contains at most K nonzero entries. In order to recover $\bar{\boldsymbol{X}}$, similar to (2), we solve the following minimization

$$\min_{\boldsymbol{X}} \ \|D\boldsymbol{X}\|_1, \qquad \text{subject to} \quad A\boldsymbol{X} = \boldsymbol{Y}. \tag{40}$$

In the remaining of this section, we prove that a solution $\hat{\boldsymbol{X}}$ of (40) is exactly the original $\bar{\boldsymbol{X}}$ with high probability, as long as

$$M \geq \begin{cases} C_1 K \log^3 N, & \text{if } d = 2, \\ C_2 K \log N, & \text{if } d > 2, \end{cases}$$

where $C_1 > 0$ and $C_2 > 0$ are two constants depending on d. Note that $\|D\boldsymbol{X}\|_1$ in (40) is the anisotropic TV. Our proof can be generalized to isotropic TV without too much difficulty, since the anisotropic and isometric TV's are equivalent and their ratio is in $[1, \sqrt{2}]$.

We will use the same architecture as in the proof of Part (a) of Theorem 2.1. Similar to Lemma 3.1, a sufficient condition for the original $\bar{\boldsymbol{X}}$ being the unique solution of (40) is the following null space condition

$$\|(D\boldsymbol{X})_{\mathcal{K}}\|_1 < \|(D\boldsymbol{X})_{\mathcal{K}^c}\|_1 \quad \forall \mathcal{K} \in [N]^d \times [d] \text{ s.t. } |\mathcal{K}| \leq K. \tag{41}$$

Here we have used $[N] = \{1, \ldots, N\}$. Different from one-dimensional case, this null space condition (41) is only a sufficient condition for higher dimensional signals. Then, using Theorem 3.2, (41) holds true with overwhelming probability if the Gaussian width satisfies $w(\mathcal{S}^{(d)}) < \sqrt{M} - \frac{1}{2\sqrt{M}}$, where

$$\mathcal{S}^{(d)} = \{\boldsymbol{X} \in \mathbb{R}^{N^d} \ : \ \|\boldsymbol{X}\|_2 = 1, \quad \text{and} \quad \|(D\boldsymbol{X})_{\mathcal{K}}\|_1 \geq \|(D\boldsymbol{X})_{\mathcal{K}^c}\|_1 \ \exists \mathcal{K} \subset [N]^d \times [d] \text{ s.t. } |\mathcal{K}| \leq K\}.$$

Given any vector $\boldsymbol{X} \in \mathcal{S}^{(d)}$, we have

$$\begin{aligned}\|D\boldsymbol{X}\|_1 &= \|(D\boldsymbol{X})_{\mathcal{K}^c}\|_1 + \|(D\boldsymbol{X})_{\mathcal{K}}\|_1 \leq 2\|(D\boldsymbol{X})_{\mathcal{K}}\|_1 \leq 2\sqrt{K}\|(D\boldsymbol{X})_{\mathcal{K}}\|_2 \\ &\leq 2\sqrt{K}\|D\boldsymbol{X}\|_2 \leq 4\sqrt{d}\sqrt{K}\|\boldsymbol{X}\|_2 = 4\sqrt{d}\sqrt{K}.\end{aligned}$$

We have used the fact that $\|D\boldsymbol{X}\|_2 \leq 2\sqrt{d}\|\boldsymbol{X}\|_2$. Therefore,

$$\mathcal{S}^{(d)} \subset \widetilde{\mathcal{S}}^{(d)} := \{\boldsymbol{X} \in \mathbb{R}^{N^d} \ : \ \|\boldsymbol{X}\|_2 \leq 1, \ \ \|D\boldsymbol{X}\|_1 \leq 4\sqrt{d}\sqrt{K}\}.$$

In the following, we estimate the Gaussian width of $\widetilde{\mathcal{S}}^{(d)}$. Similar to one-dimensional signal, we consider only the case where $N = 2^L$, and the Gaussian width for other N is the same order. For any $\boldsymbol{X} \in \widetilde{\mathcal{S}}^{(d)}$, we decompose $\boldsymbol{X}$ according to a Haar transform for d-tuple indexed vector as

$$\boldsymbol{X} = \textstyle\sum_{\ell=1}^{L} \sum_{\boldsymbol{i} \in \{0,1\}^d \setminus \boldsymbol{0}} \hat{\boldsymbol{Z}}^{(\ell,\boldsymbol{i})} + \hat{\boldsymbol{Y}}^{(L)}, \tag{42}$$

where $\hat{\boldsymbol{Z}}^{(\ell,\boldsymbol{i})} = \boldsymbol{Z}^{(\ell,\boldsymbol{i})} \otimes \boldsymbol{H}^{(\boldsymbol{i})} \otimes \boldsymbol{1}_{2^{\ell-1}}$ with $\boldsymbol{Z}^{(\ell,\boldsymbol{i})} \in \mathbb{R}^{(N/2^\ell)^d}$, $\hat{\boldsymbol{Y}}^{(L)} = \boldsymbol{Y}^{(L)} \otimes \boldsymbol{1}_N$ with $\boldsymbol{Y}^{(L)} \in \mathbb{R}$, and $\boldsymbol{i} \in \{0,1\}^d$ is a multi-index. Furthermore, here $\boldsymbol{1}_n \in \mathbb{R}^{n^d}$ is the d-tuple indexed vector whose entries are all 1, and $\otimes$ is the Kronecker product, i.e., $\boldsymbol{A} \otimes \boldsymbol{B}$ is the block d-tuple indexed vector whose $(j_1, j_2, \ldots, j_d)$ block is $A_{j_1 j_2 \ldots j_d}\boldsymbol{B}$. Moreover, $\boldsymbol{H}^{(\boldsymbol{i})} \in \mathbb{R}^{2^d}$ is the (scaled) Haar filter defined by $H^{(\boldsymbol{i})}_{j_1 j_2 \ldots j_d} = \prod_{k=1}^{d} h^{(i_k)}_{j_k}$ with $\boldsymbol{h}^{(0)} = [h^{(0)}_0 \ h^{(0)}_1]^T = [1 \ 1]^T$ and $\boldsymbol{h}^{(1)} = [h^{(1)}_0 \ h^{(1)}_1]^T = [1 \ -1]^T$. In particular, we have $\boldsymbol{H}^{(0)} = \boldsymbol{1}_2$.

The decomposition (42) is done recursively as follows. We first define $\boldsymbol{Y}^{(0)} = \hat{\boldsymbol{Y}}^{(0)} = \boldsymbol{X}$. At level ℓ, we decompose $\boldsymbol{Y}^{(\ell)}$ as

$$\hat{\boldsymbol{Y}}^{(\ell)} = \textstyle\sum_{\boldsymbol{i} \in \{0,1\}^d \setminus \boldsymbol{0}} \hat{\boldsymbol{Z}}^{(\ell+1,\boldsymbol{i})} + \hat{\boldsymbol{Y}}^{(\ell+1)},$$

where

$$Y^{(\ell+1)}_{\boldsymbol{k}} = \frac{\sum_{\boldsymbol{j} \in \{0,1\}^d} H^{(\boldsymbol{0})}_{\boldsymbol{j}} Y^{(\ell)}_{2\boldsymbol{k}-\boldsymbol{j}}}{2^d} = \frac{\sum_{\boldsymbol{j} \in \{0,1\}^d} Y^{(\ell)}_{2\boldsymbol{k}-\boldsymbol{j}}}{2^d}, \qquad Z^{(\ell+1,\boldsymbol{i})}_{\boldsymbol{k}} = \frac{\sum_{\boldsymbol{j} \in \{0,1\}^d} H^{(\boldsymbol{i})}_{\boldsymbol{j}} Y^{(\ell)}_{2\boldsymbol{k}-\boldsymbol{j}}}{2^d}.$$

The decomposition (42) has the following properties.

- Obviously, components in decomposition (42) are orthogonal to each others. Consequently,

$$\begin{aligned}\|\boldsymbol{X}\|_2^2 &= \textstyle\sum_{\ell=1}^{L} \sum_{\boldsymbol{i} \in \{0,1\}^d \setminus \boldsymbol{0}} \|\hat{\boldsymbol{Z}}^{(\ell,\boldsymbol{i})}\|_2^2 + \|\hat{\boldsymbol{Y}}^{(L)}\|_2^2 \\ &= \textstyle\sum_{\ell=1}^{L} \left(2^{d\ell} \sum_{\boldsymbol{i} \in \{0,1\}^d \setminus \boldsymbol{0}} \|\boldsymbol{Z}^{(\ell,\boldsymbol{i})}\|_2^2\right) + 2^{dL}\|\boldsymbol{Y}^{(L)}\|_2^2.\end{aligned}$$

Since $\boldsymbol{X} \in \tilde{\mathcal{S}}^{(d)}$ implies $\|\boldsymbol{X}\|_2^2 \leq 1$, we have

$$\sum_{\ell=1}^{L} \left(2^{d\ell} \sum_{\boldsymbol{i}\in\{0,1\}^d\setminus\boldsymbol{0}} \|\boldsymbol{Z}^{(\ell,\boldsymbol{i})}\|_2^2\right) + 2^{dL}\|\boldsymbol{Y}^{(L)}\|_2^2 \leq 1 \tag{43}$$

- It can be shown that

$$\|D\hat{\boldsymbol{Y}}^{(\ell)}\|_1 \leq \|D\hat{\boldsymbol{Y}}^{(\ell-1)}\|_1 \tag{44}$$

and, consequently,

$$\|D\hat{\boldsymbol{Y}}^{(\ell)}\|_1 \leq \|D\boldsymbol{X}\|_1 \leq 4\sqrt{d}\sqrt{K}. \tag{45}$$

Indeed, let D_i be the difference matrix along the i-th dimension. Then, similar to the one-dimensional case, one can show that

$$\begin{aligned}\|D_i\hat{\boldsymbol{Y}}^{(\ell)}\|_1 &= 2^{\ell(d-1)}\|D_i\boldsymbol{Y}^{(\ell)}\|_1 \leq 2^{\ell(d-1)} \cdot \frac{2}{2^d}\|D_i\boldsymbol{Y}^{(\ell-1)}\|_1 \\ &= 2^{\ell(d-1)} \cdot \frac{2}{2^d}\frac{1}{2^{(\ell-1)(d-1)}}\|D_i\hat{\boldsymbol{Y}}^{(\ell-1)}\|_1 = \|D_i\hat{\boldsymbol{Y}}^{(\ell-1)}\|_1.\end{aligned}$$

Summing over i yields (44).

- Furthermore, for any vector $\boldsymbol{G}$, we have

$$\sum_{\boldsymbol{i}\in\{0,1\}^d\setminus\boldsymbol{0}} \langle \boldsymbol{G}, \hat{\boldsymbol{Z}}^{(\ell,\boldsymbol{i})}\rangle \leq \sqrt{d}2^{d+1}\frac{\sqrt{K}}{2^{(\ell-1)(d-1)}}\|\tilde{\boldsymbol{G}}^{(\ell-1)}\|_\infty, \tag{46}$$

where $\tilde{\boldsymbol{G}}^{(\ell-1)} \in \mathbb{R}^{(N/2^{\ell-1})^d}$ is a d-tuple indexed signal whose $\boldsymbol{i}$-th entry is the sum of the entries of $\boldsymbol{G}$ on the $\boldsymbol{i}$-th block of size $2^{\ell-1} \times 2^{\ell-1}$. Eq. (46) is shown as follows. We have

$$\begin{aligned}\sum_{\boldsymbol{i}\in\{0,1\}^d\setminus\boldsymbol{0}} \langle \boldsymbol{G}, \hat{\boldsymbol{Z}}^{(\ell,\boldsymbol{i})}\rangle &= \sum_{\boldsymbol{i}\in\{0,1\}^d\setminus\boldsymbol{0}} \langle \tilde{\boldsymbol{G}}^{(\ell-1)}, \boldsymbol{Z}^{(\ell,\boldsymbol{i})} \otimes \boldsymbol{H}^{(\boldsymbol{i})}\rangle \leq \|\tilde{\boldsymbol{G}}^{(\ell-1)}\|_\infty \cdot \left\|\sum_{\boldsymbol{i}\in\{0,1\}^d\setminus\boldsymbol{0}} \boldsymbol{Z}^{(\ell,\boldsymbol{i})} \otimes \boldsymbol{H}^{(\boldsymbol{i})}\right\|_1 \\ &\leq \|\tilde{\boldsymbol{G}}^{(\ell-1)}\|_\infty \cdot \sum_{\boldsymbol{j}\in\{0,1\}^d} \left\|\sum_{\boldsymbol{i}\in\{0,1\}^d\setminus\boldsymbol{0}} H_{\boldsymbol{j}}^{(\boldsymbol{i})}\boldsymbol{Z}^{(\ell,\boldsymbol{i})}\right\|_1 \\ &\leq \|\tilde{\boldsymbol{G}}^{(\ell-1)}\|_\infty \cdot 2^d \sum_{\boldsymbol{i}\in\{0,1\}^d\setminus\boldsymbol{0}} \left\|\boldsymbol{Z}^{(\ell,\boldsymbol{i})}\right\|_1,\end{aligned} \tag{47}$$

where we have used the fact that $|H_{\boldsymbol{j}}^{(\boldsymbol{i})}| = 1$ for all indices $\boldsymbol{i}$ and $\boldsymbol{j}$. Define $\boldsymbol{H}^{(\boldsymbol{i}\setminus k)} \in \mathbb{R}^{2^{d-1}}$ and $H_{j_1\cdots j_{k-1}j_{k+1}\cdots j_d}^{(\boldsymbol{i}\setminus k)} = \prod_{p=1,p\neq k}^{d} h_{j_p}^{(i_p)}$ with $\boldsymbol{h}^{(0)} = [1\ 1]^T$, $\boldsymbol{h}^{(1)} = [1\ -1]^T$. For any fixed $\tilde{\boldsymbol{i}} = (\tilde{i}_1, \ldots, \tilde{i}_d) \in \{0,1\}^d \setminus \boldsymbol{0}$, there exists a k_0 such that $\tilde{i}_{k_0} = 1$. Since $\langle \boldsymbol{H}^{(\boldsymbol{i}\setminus k_0)}, \boldsymbol{H}^{(\boldsymbol{j}\setminus k_0)}\rangle = 2^{d-1}\prod_{k\neq k_0}\delta_{i_k j_k}$ with δ_{ij}

being 1 if $i = j$ and 0 otherwise, we have

$$\sum_{\boldsymbol{i}\in\{0,1\}^d,\ i_{k_0}=1} \sum_{\boldsymbol{j}\in\{0,1\}^d} H_{\boldsymbol{j}}^{\boldsymbol{i}\backslash k_0} H_{\boldsymbol{j}}^{\tilde{\boldsymbol{i}}\backslash k_0} = 2^{d-1},$$

and thus

$$\boldsymbol{Z}^{(\ell,\tilde{\boldsymbol{i}})} = \frac{1}{2^{d-1}} \sum_{\boldsymbol{j}\in\{0,1\}^{d-1}} \left(H_{\boldsymbol{j}}^{(\tilde{\boldsymbol{i}}\backslash k_0)} \sum_{\boldsymbol{i}\in\{0,1\}^d,\ i_{k_0}=1} H_{\boldsymbol{j}}^{(\boldsymbol{i}\backslash k_0)} \boldsymbol{Z}^{(\ell,\boldsymbol{i})} \right),$$

which leads to

$$\begin{aligned} \|\boldsymbol{Z}^{(\ell,\tilde{\boldsymbol{i}})}\|_1 &\leq \frac{1}{2^{d-1}} \sum_{\boldsymbol{j}\in\{0,1\}^{d-1}} \left\| H_{\boldsymbol{j}}^{(\tilde{\boldsymbol{i}}\backslash k_0)} \sum_{\boldsymbol{i}\in\{0,1\}^d,\ i_{k_0}=1} H_{\boldsymbol{j}}^{(\boldsymbol{i}\backslash k_0)} \boldsymbol{Z}^{(\ell,\boldsymbol{i})} \right\|_1 \\ &= \frac{1}{2^{d-1}} \sum_{\boldsymbol{j}\in\{0,1\}^{d-1}} \left\| \sum_{\boldsymbol{i}\in\{0,1\}^d,\ i_{k_0}=1} H_{\boldsymbol{j}}^{(\boldsymbol{i}\backslash k_0)} \boldsymbol{Z}^{(\ell,\boldsymbol{i})} \right\|_1 . \end{aligned} \tag{48}$$

In the following, we estimate the last term. Let D_k be the difference along the k-th direction. We have that $D_k \boldsymbol{H}^{(\boldsymbol{i})} = 2\boldsymbol{H}^{(\boldsymbol{i}\backslash k)}$ if $i_k = 1$ and $D_k \boldsymbol{H}^{(\boldsymbol{i})} = \boldsymbol{0}$ if $i_k = 0$. Note that, for any d-tuple indexed vector $\boldsymbol{a}$, we have

$$\begin{aligned} \left\| D \left(\sum_{\boldsymbol{i}\in\{0,1\}^d} a_{\boldsymbol{i}} \boldsymbol{H}^{(\boldsymbol{i})} \otimes \boldsymbol{1}_{2^{\ell-1}} \right) \right\|_1 &= 2^{(d-1)(\ell-1)} \sum_{k=1}^d \left\| \sum_{\boldsymbol{i}\in\{0,1\}^d} a_{\boldsymbol{i}} D_k \boldsymbol{H}^{(\boldsymbol{i})} \right\|_1 \\ &= 2 \cdot 2^{(d-1)(\ell-1)} \sum_{k=1}^d \left\| \sum_{\boldsymbol{i}\in\{0,1\}^d,\ i_k=1} a_{\boldsymbol{i}} \boldsymbol{H}^{(\boldsymbol{i}\backslash k)} \right\|_1 \end{aligned}$$

This together with (45) and the definition $\hat{\boldsymbol{Z}}^{(\ell,\boldsymbol{i})} = \boldsymbol{Z}^{(\ell,\boldsymbol{i})} \otimes \boldsymbol{H}^{(\boldsymbol{i})} \otimes \boldsymbol{1}_{2^{\ell-1}}$ implies (defining $\boldsymbol{Z}^{(\ell,\boldsymbol{0})} = \boldsymbol{Y}^{(\ell)}$ and $\hat{\boldsymbol{Z}}^{(\ell,\boldsymbol{0})} = \hat{\boldsymbol{Y}}^{(\ell)}$)

$$\begin{aligned} 4\sqrt{d}\sqrt{K} \geq \|D\hat{\boldsymbol{Y}}^{(\ell-1)}\|_1 &= \left\| D \left(\hat{\boldsymbol{Y}}^{(\ell)} + \sum_{\boldsymbol{i}\in\{0,1\}^d\backslash\boldsymbol{0}} \hat{\boldsymbol{Z}}^{(\ell,\boldsymbol{i})} \right) \right\|_1 \\ &\geq \sum_{\boldsymbol{p}\in[N/2^\ell]^d} \left\| D \left(\sum_{\boldsymbol{i}\in\{0,1\}^d} Z_{\boldsymbol{p}}^{(\ell,\boldsymbol{i})} \boldsymbol{H}^{(\boldsymbol{i})} \otimes \boldsymbol{1}_{2^{\ell-1}} \right) \right\|_1 \\ &= 2 \cdot 2^{(d-1)(\ell-1)} \sum_{k=1}^d \sum_{\boldsymbol{p}\in[N/2^\ell]^d} \left\| \sum_{\boldsymbol{i}\in\{0,1\}^d,\ i_k=1} Z_{\boldsymbol{p}}^{(\ell,\boldsymbol{i})} \boldsymbol{H}^{(\boldsymbol{i}\backslash k)} \right\|_1 \\ &= 2 \cdot 2^{(d-1)(\ell-1)} \sum_{k=1}^d \sum_{\boldsymbol{j}\in\{0,1\}^{d-1}} \left\| \sum_{\boldsymbol{i}\in\{0,1\}^d,\ i_k=1} H_{\boldsymbol{j}}^{(\boldsymbol{i}\backslash k)} \boldsymbol{Z}^{(\ell,\boldsymbol{i})} \right\|_1 \end{aligned}$$

and therefore

$$\sum_{k=1}^d \sum_{\boldsymbol{j}\in\{0,1\}^{d-1}} \left\| \sum_{\boldsymbol{i}\in\{0,1\}^d,\ i_k=1} H_{\boldsymbol{j}}^{(\boldsymbol{i}\backslash k)} \boldsymbol{Z}^{(\ell,\boldsymbol{i})} \right\|_1 \leq \frac{2\sqrt{d}}{2^{(d-1)(\ell-1)}} \sqrt{K}. \tag{49}$$

Let $n(\boldsymbol{i})$ be the number of 1's in $\boldsymbol{i}$. By combining (48) and (49), we obtain

$$\begin{aligned}\sum_{\tilde{\boldsymbol{i}}\in\{0,1\}^d\setminus\boldsymbol{0}} \left\|\boldsymbol{Z}^{(\ell,\tilde{\boldsymbol{i}})}\right\|_1 &= \sum_{p=1}^{d}\sum_{\tilde{\boldsymbol{i}}:n(\tilde{\boldsymbol{i}})=p} \left\|\boldsymbol{Z}^{(\ell,\tilde{\boldsymbol{i}})}\right\|_1 = \sum_{p=1}^{d}\sum_{k=1}^{d}\sum_{\tilde{\boldsymbol{i}}:\tilde{i}_k=1,n(\tilde{\boldsymbol{i}})=p} \left\|\boldsymbol{Z}^{(\ell,\tilde{\boldsymbol{i}})}\right\|_1 \\ &\leq \frac{1}{2^{d-1}}\sum_{p=1}^{d}\sum_{k=1}^{d}\sum_{\tilde{\boldsymbol{i}}:\tilde{i}_k=1,n(\tilde{\boldsymbol{i}})=p}\sum_{\boldsymbol{j}\in\{0,1\}^{d-1}}\left\|\sum_{\boldsymbol{i}\in\{0,1\}^d,\ i_k=1} H_{\boldsymbol{j}}^{(\boldsymbol{i}\setminus k)}\boldsymbol{Z}^{(\ell,\boldsymbol{i})}\right\|_1 \qquad (50) \\ &\leq \frac{1}{2^{d-1}}\sum_{p=1}^{d}\binom{d-1}{p-1}\frac{2\sqrt{d}}{2^{(d-1)(\ell-1)}}\sqrt{K} = \frac{2\sqrt{d}}{2^{(d-1)(\ell-1)}}\sqrt{K},\end{aligned}$$

which combined with (47) yields (46).

Now we are ready to estimate the Gaussian width of $\tilde{\mathcal{S}}^{(d)}$. Let $\boldsymbol{G}$ be a vector whose entries are i.i.d. Gaussian random variables with mean 0 and variance 1. The same argument in one dimensional cases leads to

$$E(\|\tilde{\boldsymbol{G}}^{(\ell)}\|_\infty) \leq \sqrt{2}\sqrt{2^{d\ell}\log\left(2N^d/2^{d\ell}\right)}$$

which implies

$$\begin{aligned}E\left(\sup\sum_{\boldsymbol{i}\in\{0,1\}^d\setminus\boldsymbol{0}}\langle\boldsymbol{G},\hat{\boldsymbol{Z}}^{(\ell,\boldsymbol{i})}\rangle\right) &\leq \sqrt{d}2^{d+1}\frac{\sqrt{K}}{2^{(\ell-1)(d-1)}}E(\|\tilde{\boldsymbol{G}}^{(\ell-1)}\|_\infty) \\ &\leq \sqrt{2}\sqrt{d}2^{d+1}\frac{\sqrt{K}}{2^{(\ell-1)(d-1)}}\sqrt{2^{d(\ell-1)}\log\left(2N^d/2^{d(\ell-1)}\right)} \\ &\leq \sqrt{2}\sqrt{d}2^{d+1}\frac{\sqrt{K}}{2^{(\ell-1)(d-1)}}\sqrt{2^{d(\ell-1)}\log\left(2N^d\right)} \\ &= \sqrt{2}\sqrt{d}2^{d+1}\sqrt{K}2^{(\ell-1)(1-\frac{d}{2})}\sqrt{\log\left(2N^d\right)}.\end{aligned}$$

Moreover,

$$E\left(\sup\langle\boldsymbol{G},\hat{\boldsymbol{Y}}^{(L)}\rangle\right) = E\left(\sup Y^{(L)}\tilde{\boldsymbol{G}}^{(L)}\right) \leq \sup|Y^{(L)}|E\left(|\tilde{\boldsymbol{G}}^{(L)}|\right) \leq \sqrt{\frac{1}{2^{dL}}}\sqrt{2^{dL}}\sqrt{2/\pi} = \sqrt{2/\pi}.$$

Therefore,

$$E\left(\sup_{\boldsymbol{X}\in\tilde{\mathcal{S}}^{(d)}}\langle\boldsymbol{G},\boldsymbol{X}\rangle\right)=\sum_{\ell=1}^{L-1}E\left(\sup\sum_{\boldsymbol{i}\in\{0,1\}^d\setminus\boldsymbol{0}}\langle\boldsymbol{G},\hat{\boldsymbol{Z}}^{(\ell,\boldsymbol{i})}\rangle\right)+E\left(\sup\langle\boldsymbol{G},\hat{\boldsymbol{Y}}^{(L)}\rangle\right)$$
$$\leq\sqrt{2}\sqrt{d}2^{d+1}\sqrt{K}\sqrt{\log\left(2N^d\right)}\sum_{\ell=1}^{L-1}2^{(\ell-1)(1-\frac{d}{2})}+\sqrt{2/\pi}$$
$$\leq\begin{cases}\sqrt{2}\sqrt{d}2^{d+1}\sqrt{K}\sqrt{\log\left(2N^d\right)}\log_2 N+\sqrt{2/\pi} & \text{if } d=2,\\ \sqrt{2}\sqrt{d}2^{d+1}\sqrt{K}\sqrt{\log\left(2N^d\right)}\frac{1}{1-2^{1-\frac{d}{2}}}+\sqrt{2/\pi} & \text{if } d>2\end{cases}$$

We require the Gaussian width is about $\sqrt{M}$, where M is the number of measurement. So, we have

$$M\sim\begin{cases}K\log^3 N & \text{if } d=2\\ K\log N & \text{if } d>2.\end{cases}$$

7 Conclusion

In this paper, we prove the performance guarantee of total variation (TV) minimization in recovering sparse-gradient *one-dimensional* signal. The almost Euclidean property of subspaces [25, 46, 47] is used to extend our results to proving the stability of TV minimization for signals with approximately sparse gradients or under noisy measurements. Our results can also be extended to TV minimization for multidimensional signals. Stability of TV minimization has also been established for one-dimensional signal vectors with large M and K.

Our current results work only for the Gaussian ensemble of measurement matrices. One future direction is to extend our results to general deterministic and random measurement matrices, such as partial Fourier matrices, and random Bernoulli matrices. Another direction we would like to pursue is to establish the stability of TV minimization when K is small by using Gaussian width [34]. Finally, we are also interested in extending the result to general CS analysis model. Current results (e.g. [28]) usually assume that the analysis operator has a small condition number. Though the finite difference matrix D has a bad condition number, our analysis in this paper still can get linear growth of the TV recovery threshold when the number of measurements is linear to the signal size. We expect that our analysis can be applied to general CS analysis models to get recovery thresholds that are independent of the condition number of the analysis operator.

References

[1] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, *Living on the edge: Phase transitions in convex programs with random data*, Information and Inference, (2014).

[2] 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.

[3] J.-F. Cai, S. Osher, and Z. Shen, *Split Bregman methods and frame based image restoration*, Multiscale Modeling & Simulation, 8 (2009), pp. 337–369.

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

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

[6] E. J. Candès and P. A. Randall, *Highly robust error correction by convex programming*, IEEE Trans. Inform. Theory, 54 (2008), pp. 2829–2840.

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

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

[9] E. J. Candes and T. Tao, *Near-optimal signal recovery from random projections: Universal encoding strategies?*, Information Theory, IEEE Transactions on, 52 (2006), pp. 5406–5425.

[10] A. Chambolle and J. Darbon, *On total variation minimization and surface evolution using parametric maximum flows*, International Journal of Computer Vision, 84 (2009), pp. 288–307.

[11] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky, *The convex geometry of linear inverse problems*, Foundations of Computational Mathematics, 12 (2012), pp. 805–849.

[12] D. Donoho, I. Johnstone, and A. Montanari, *Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising*, Information Theory, IEEE Transactions on, 59 (2013), pp. 3396–3433.

[13] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.

[14] ———, *High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension*, Discrete Comput. Geom., 35 (2006), pp. 617–652.

[15] D. L. Donoho, A. Maleki, and A. Montanari, *The noise-sensitivity phase transition in compressed sensing*, IEEE Trans. Inform. Theory, 57 (2011), pp. 6920–6941.

[16] D. L. Donoho and J. Tanner, *Neighborliness of randomly projected simplices in high dimensions*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 9452–9457 (electronic).

[17] J. Fadili, G. Peyré, S. Vaiter, C.-A. Deledalle, and J. Salmon, *Stable recovery with analysis decomposable priors*, in SPARS, 2013.

[18] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, Springer, 2013.

[19] A. Y. Garnaev and E. D. Gluskin, *The widths of a Euclidean ball*, Dokl. Akad. Nauk SSSR, 277 (1984), pp. 1048–1052.

[20] T. Goldstein and S. Osher, *The split Bregman method for L1 regularized problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 323–343.

[21] Y. Gordon, *On Milman's inequality and random subspaces which escape through a mesh in* $\mathbf{R}^n$, in Geometric aspects of functional analysis (1986/87), vol. 1317 of Lecture Notes in Math., Springer, Berlin, 1988, pp. 84–106.

[22] M. Grasmair, *Linear convergence rates for tikhonov regularization with positively homogeneous functionals*, Inverse Problems, 27 (2011), p. 075014.

[23] M. Haltmeier, *Stable signal reconstruction via* ℓ^1*-minimization in redundant, non-tight frames*, IEEE Transactions on Signal Processing, 61 (2013), pp. 420–426.

[24] M. Kabanava, H. Rauhut, and H. Zhang, *Robust analysis* ℓ_1*-recovery from gaussian measurements and total variation minimization*, arXiv preprint arxiv:1407:7402, (2014).

[25] B. S. Kašin, *The widths of certain finite-dimensional sets and classes of smooth functions*, Izv. Akad. Nauk SSSR Ser. Mat., 41 (1977), pp. 334–351, 478.

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

[27] M. Ledoux, *The Concentration of Measure Phenomenon*, Mathematical Surveys and Monographs, American Mathematical Society, 2005.

[28] J. Liu, L. Yuan, and J. Ye, *Guaranteed sparse recovery under linear transformation*, arXiv preprint arXiv:1305.0047, (2013).

[29] S. Nam, M. E. Davies, M. Elad, and R. Gribonval, *Cosparse analysis modeling-uniqueness and algorithms*, in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE, 2011, pp. 5804–5807.

[30] S. Nam, M. E. Davies, M. Elad, and R. Gribonval, *The cosparse analysis model and algorithms*, Applied and Computational Harmonic Analysis, 34 (2013), pp. 30–56.

[31] D. Needell and R. Ward, *Near-optimal compressed sensing guarantees for total variation minimization*, Image Processing, IEEE Transactions on, 22 (2013), pp. 3941–3949.

[32] D. Needell and R. Ward, *Stable image reconstruction using total variation minimization*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 1035–1058.

[33] S. Oymak and B. Hassibi, *Sharp mse bounds for proximal denoising*, arXiv preprint arXiv:1305.2714, (2013).

[34] S. Oymak, C. Thrampoulidis, and B. Hassibi, *The squared-error of generalized lasso: A precise analysis*, arXiv preprint arXiv:1311.0830, (2013).

[35] C. Poon, *On the role of total variation in compressed sensing - structure dependence*, arXiv preprint arxiv:1407.5337, (2014).

[36] H. Rauhut, K. Schnass, and P. Vandergheynst, *Compressed sensing and redundant dictionaries*, IEEE Transactions on Information Theory, 54 (2008), pp. 2210–2219.

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

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

[39] E. Sidky and X. Pan, *Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization*, Physics in Medicine and Biology, 53 (2008), pp. 4777–4807.

[40] M. Stojnic, *Various thresholds for ℓ_1-optimization in f*, arXiv preprint arXiv:0907.3666, (2009).

[41] S. Vaiter, G. Peyre, C. Dossal, and J. Fadili, *Robust sparse analysis regularization*, IEEE Transactions on Information Theory, 59 (2013), pp. 2001–2016.

[42] P. Van den Berg and R. Kleinman, *A total variation enhanced modified gradient algorithm for profile reconstruction*, Inverse Problems, 11 (1999), p. L5.

[43] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, in Compressed sensing: theory and applications, Cambridge University Press, New York, 2012, pp. 210–268.

[44] X. Wu and M. Liu, *In-situ soil moisture sensing: Measurement scheduling and estimation using compressive sensing*, in Proceedings of the 11th International Conference on Information Processing in Sensor Networks, IPSN '12, New York, NY, USA, 2012, ACM, pp. 1–12.

[45] W. Xu and B. Hassibi, *Precise stability phase transitions for ℓ_1 minimization: a unified geometric framework*, IEEE Trans. Inform. Theory, 57 (2011), pp. 6894–6919.

[46] W. Xu, M. Wang, J.-F. Cai, and A. Tang, *Sparse recovery from nonlinear measurements with applications in bad data detection for power networks*, IEEE Transactions on Signal Processing, 61 (2013), pp. 6175–6187.

[47] Y. Zhang, *A simple proof for recoverability of ℓ_1-minimization: go over or under*, Rice CAAM Technical Report *http://www.caam.rice.edu/∼yzhang/reports/tr0509.pdf*, (2005).