On the Error in Phase Transition Computations for Compressed Sensing

Sajad Daei, Farzan Haddadi, Arash Amini, Martin Lotz

Abstract—**Evaluating the statistical dimension is a common tool to determine the asymptotic phase transition in compressed sensing problems with Gaussian ensemble. Unfortunately, the exact evaluation of the statistical dimension is very difficult and it has become standard to replace it with an upper-bound. To ensure that this technique is suitable, [1] has introduced an upper-bound on the gap between the statistical dimension and its approximation. In this work, we first show that the error bound in [1] in some low-dimensional models such as total variation and ℓ_1 analysis minimization becomes poorly large. Next, we develop a new error bound which significantly improves the estimation gap compared to [1]. In particular, unlike the bound in [1] that fails in some settings with overcomplete dictionaries, our bound exhibits a decaying behavior in such cases.**

Index Terms—**statistical dimension, error estimate, low-complexity models.**

I. INTRODUCTION

UNDERSTANDING the behavior of random compressed sensing problems in transition from absolute failure to success (known as phase transition) has been the subject of research in recent years [1]–[8]. Most of these works concentrate on simple sparse models and do not allude to the challenges in other low-dimensional structures such as low-rank matrices, block-sparse vectors, gradient-sparse vectors and cosparse (also known as analysis sparse [9]) vectors. For simplicity, we associate such structures with their common recovery techniques and rename the structures accordingly. For instance, total variation (TV), ℓ_1 analysis and $\ell_{1,2}$ minimization refer to both the recovery techniques and the underlying low-dimensional structures. In this work, we revisit linear inverse problems with the aim of recovering a vector $\boldsymbol{x} \in \mathbb{R}^n$ from a few random linear measurements $\boldsymbol{y} = \boldsymbol{Ax} \in \mathbb{R}^m$. This is summarized as solving the following convex program:

$$\mathsf{P}_f: \quad \min_{\boldsymbol{z}\in\mathbb{R}^n} f(\boldsymbol{z})$$
$$\text{s.t.} \quad \boldsymbol{y} = \boldsymbol{Az}, \tag{1}$$

where, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is the measurement matrix whose entries are i.i.d. random variables with normal distribution and f is a convex penalty function that promotes the low-dimensional structure. A major subject of recent research is the number of Gaussian measurements (the number of rows in $\boldsymbol{A}$) one needs to recover a structured vector $\boldsymbol{x}$ from $\boldsymbol{y} \in \mathbb{R}^m$. In [3], a bound is obtained using polytope angle calculations with asymptotic sharpness in case of $f = \|\cdot\|_1$. A link between the number of

required measurements and the error in the denoising problem is investigated in [8]. For the particular case of $f = \|\cdot\|_{\mathrm{TV}}$ in P_f, we need to consider the denoising problem

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{z}\in\mathbb{R}^n} \tau\sigma\|\boldsymbol{z}\|_{\mathrm{TV}} + \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{z}\|_2^2, \tag{2}$$

where $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{w}$ with $\boldsymbol{w}$ being an additive noise drawn from $\mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. If we define the worst-case normalized mean squared error (NMSE) as

$$\mathrm{NMSE} := \lim_{\sigma\to 0} \inf_{\tau\geqslant 0} \frac{\mathbb{E}\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2^2}{\sigma^2}, \tag{3}$$

then, the result in [8] implies that the value of NMSE is a sharp estimate of the required number of measurements (for P_{TV}) in the asymptotic case. Also, in [7], the authors showed that the mentioned NMSE is the same as the number of required measurements that TV approximate message passing (TV-AMP) algorithm needs. [2] introduced a general framework for obtaining the number of Gaussian measurements in different low-dimensional structures using Gordon min-max inequality [10] and the concept of atomic norms. Specifically, it was shown that $\omega^2(\mathcal{D}(f,\boldsymbol{x})\cap\mathbb{B}^n)+1$ measurements are sufficient. Here, $\mathcal{D}(f,\boldsymbol{x})$ is the descent cone of f at $\boldsymbol{x} \in \mathbb{R}^n$ and $\omega^2(\mathcal{D}(f,\boldsymbol{x})\cap\mathbb{B}^n)$ is the squared Gaussian width, which intuitively measures the size of this cone. In [1], it has been shown that the statistical dimension of this cone, which is defined below and differs from the squared Gaussian width above by at most 1, specifies the phase transition of the [random] convex program P_f from absolute failure to absolute success:

$$\delta(\mathcal{D}(f,\boldsymbol{x})) := \mathbb{E}\,\mathrm{dist}^2(\boldsymbol{g}, \mathrm{cone}(\partial f(\boldsymbol{x}))). \tag{4}$$

$\delta(\mathcal{D}(f,\boldsymbol{x}))$ is the average distance of a standard Gaussian i.i.d. vector $\boldsymbol{g} \in \mathbb{R}^n$ from non-negative scalings of the sub-differential at point $\boldsymbol{x} \in \mathbb{R}^n$. So far, we know that a phase transition exists in P_f and its boundary is interpreted via the statistical dimension. A natural question is how we can find an expression for the phase transition curve. The upper-bound for $\delta(\mathcal{D}(f,\boldsymbol{x}))$, first used in the context of ℓ_1 minimization by Stojnic ([6]), is given by:

$$\delta(\mathcal{D}(f,\boldsymbol{x})) \leqslant \inf_{t\geqslant 0} \mathbb{E}\,\mathrm{dist}^2(\boldsymbol{g}, t\partial f(\boldsymbol{x})) := U_\delta. \tag{5}$$

However, it is still unknown whether U_δ is sharp for different low-dimensional structures. For ease of notation, we define the error E_δ by:

$$E_\delta := U_\delta - \delta(\mathcal{D}(f,\boldsymbol{x})). \tag{6}$$

Here, U_δ represents a sufficient number of measurements that P_f needs for successful recovery. In [1], implicit formulas are derived for the upper-bound (5) in case of ℓ_1 and nuclear norm. Recently, an explicit upper-bound for U_δ in case of ℓ_1 analysis and TV minimization is presented in [11]. The proposed bound depends on a notion called "generalized analysis sparsity" and is numerically observed to be tight for many analysis operators. In [1], a general upper-bound expression for E_δ (known as the error estimate) for various structure-inducing functions (e.g. ℓ_1) is proposed (see Theorem 1). For the cases of ℓ_1 and nuclear norm minimization, it is further shown that the normalized error estimate $\frac{\widehat{E}_p}{n}$ (and thus $\frac{E_\delta}{n}$) vanishes in high dimensions. This result confirms that U_δ is a good surrogate for $\delta(\mathcal{D}(f, \boldsymbol{x}))$. However, the asymptotic behavior of $\frac{E_\delta}{n}$ in cases of $\|\cdot\|_{1,2}$, $\|\boldsymbol{\Omega}\cdot\|_1$ and $\|\cdot\|_{\mathrm{TV}} := \|\boldsymbol{\Omega_d}\cdot\|_1$ where

$$\boldsymbol{\Omega}_d = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \cdots & \cdots & 1 & -1 \end{bmatrix} \in \mathbb{R}^{n-1 \times n}, \quad (7)$$

is not studied. Thus, one could simply think of the following problems:

1) Does U_δ provide a fair estimate of the statistical dimension?
2) How to quantify the gap between the exact phase transition curve and the one obtained via U_δ?
3) Can one extend the previous error bounds obtained for ℓ_1 minimization in [1] to other low-dimensional structures such as block sparsity, TV and ℓ_1 analysis?

In this work, we try to find answers to these questions. Specifically, we want to study how well U_δ describes $\delta(\mathcal{D}(f, \boldsymbol{x}))$ in low-dimensional structures represented by $\|\cdot\|_1$, $\|\cdot\|_{1,2}$, $\|\cdot\|_*$, $\|\boldsymbol{\Omega}\cdot\|_1$ and $\|\cdot\|_{\mathrm{TV}} := \|\boldsymbol{\Omega_d}\cdot\|_1$. A generic $\boldsymbol{\Omega} \in \mathbb{R}^{p \times n}$ can be tall, i.e. $p > n$, or fat, i.e. $p \leqslant n$; due to the similarity of the arguments used in this paper, we include square matrices in the category of fat matrices. Tall $\boldsymbol{\Omega}$ matrices cover various redundant[1] analysis operators that are common in practice; in particular, redundant wavelet frames [12], and redundant random frames (widely used as a benchmark template in [9], [11], [13], [14]). Also fat matrices include examples such as the one-dimensional finite difference operator $\boldsymbol{\Omega}_d$ and non-redundant random analysis operators (used in [11, Section 3.3]). We call a signal $\boldsymbol{x} \in \mathbb{R}^n$ an analysis-sparse vector (also called as cosparse vector [9]) with respect to the analysis operator $\boldsymbol{\Omega} \in \mathbb{R}^{p \times n}$ if after applying $\boldsymbol{\Omega}$ the resulting vector becomes sparse; i.e., $\boldsymbol{\Omega x}$ is sparse. We denote the support set of $\boldsymbol{\Omega x}$ with $\mathcal{S}$; similarly, $\overline{\mathcal{S}}$ stands for the zero set of $\boldsymbol{\Omega x}$. The number of zeros in $\boldsymbol{\Omega x}$, i.e., $|\overline{\mathcal{S}}|$, is called the cosparsity of $\boldsymbol{x}$ with respect to $\boldsymbol{\Omega}$ [9], [11], [14]. We should highlight that in most of the existing literature regarding the ℓ_1-analysis problem it is assumed that $\boldsymbol{\Omega} \in \mathbb{R}^{p \times n}$ has rows in general position[2].

---

[1] The term redundant refers to an analysis operator with more number of rows than columns.

[2] Every subset of n rows of $\boldsymbol{\Omega}$ are linearly independent.

U_δ	$\delta(\mathcal{D}(\|\cdot\|_1, \boldsymbol{x}))$	$\frac{E_\delta}{n}$	$\frac{\widehat{E}_p}{n}$ in (21)	s
9.458	9.0905	0.0003	0.064	1
16.828	16.8	0.0003	0.045	2
23.4544	23.04	.0005	0.036	3
61.244	60.84	0.0004	0.02	10
104.1814	104.04	0.0001	0.015	20

Table I: In this table, we examine the error estimate [1] for function $f = \|\cdot\|_1$ and signal $\boldsymbol{x}$ with dimension $n = 1000$. The upper-bound U_δ is computed using the package SNOW in [1]. The true sample complexity is denoted by $\delta(\mathcal{D}(\|\cdot\|_1, \boldsymbol{x}))$ and computed by squaring $\omega(\mathcal{D}(\|\cdot\|_1, \boldsymbol{x}) \cap \mathbb{B}^n)$ (see Appendix E for more details). For the considered class of $\boldsymbol{x}$ signals, we observe that there is a negligible gap between $\frac{\widehat{E}_p}{n}$ in [1] and the true normalized error $\frac{E_\delta}{n}$.

A. Motivation

Tables I and II present the results of a computer experiment designed to evaluate the error of U_δ in estimating the statistical dimension. In two experiments shown in Tables I and II, we test the error bound for ℓ_1 and TV minimization. The value of $\delta(\mathcal{D}(f, \boldsymbol{x}))$ which is approximately equal to $\omega^2(\mathcal{D}(f, \boldsymbol{x}) \cap \mathbb{B}^n)$, is computed using the procedure proposed in [11, Section B.2] (see Appendix E). In the first experiment, for each sparsity level, we construct a sparse vector $\boldsymbol{x} \in \mathbb{R}^{1000}$ with random non-zero values (distributed as $\mathcal{N}(\boldsymbol{0}, \sqrt{1000})$) at uniformly random locations. In the second experiment, we set $\boldsymbol{\Omega} = \boldsymbol{\Omega}_d$, and generate a gradient sparse vector $\boldsymbol{x} \in \mathbb{R}^{1000}$ as the sum of two components: a small but non-zero component in the space $\mathrm{null}(\boldsymbol{\Omega}_{\overline{S}})$, and a large component in the space $\mathrm{null}(\boldsymbol{\Omega})$. The notation $\boldsymbol{\Omega}_{\overline{S}}$ refers to the matrix $\boldsymbol{\Omega}$ restricted to the rows indexed by $\overline{\mathcal{S}}$ (see Section IV-A for more details). For Tables I and II, the upper-bound (5) is obtained by [1, Equation D.6] and numerical optimization[3], respectively. As shown in Tables I and II, there exists a gap between the true error E_δ and the state of the art theoretical error estimate $\widehat{E}_p$ in (21). While the normalized gap, i.e., $\frac{|E_\delta - \widehat{E}_p|}{n}$ is negligible in the ℓ_1 case (Table I), it is considerable in case of TV minimization (Table II). Now, a natural question that arises is: can we find a better bound that reduces the gap?

B. Contributions

In this work, we rigorously analyze the error of estimating the phase transition. The significance of this error is to have a good understanding about the required number of measurements that P_f needs to recover a structured vector from undersampled measurements. Our analysis is general and holds for a variety of low-dimensional structures including sparse, block-sparse, analysis sparse and gradient-sparse vectors, as well as low-rank matrices. In brief, the contributions of this work can be listed as follows.

1) *Identifying a failure regime for [1]*: For $f = \|\boldsymbol{\Omega} \cdot \|_1$ in P_f, the error estimate of [1] shown in (21), can become remarkably large for some specific signals and analysis operators. For fat analysis operators, a typical signal $\boldsymbol{x}$ with such property is constructed as:

$$\boldsymbol{x} = \boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{S}})} \boldsymbol{w} + \boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega})} \boldsymbol{c} \in \mathbb{R}^n, \quad (8)$$

---

[3] See also [15, Section 4] which proposes a numerical method to calculate U_δ in case of TV minimization.

U_δ	$\delta(\mathcal{D}(\|\cdot\|_{\mathrm{TV}}, \boldsymbol{x}))$	$\frac{E_\delta}{n}$	$\frac{\widehat{E}_p}{n}$ in (21)	Average amplitude of $\boldsymbol{x}$	$\|\boldsymbol{\Omega}\boldsymbol{x}\|_\infty$	s
58	56.25	0.002	2063.3	1000	0.1	10
61.1	60.84	0.0013	5.1143	30	4	10
80.03	79.21	0.0008	0.4	3	0.9	10
982	979.69	0.002	10.4	316	0.4	850

Table II: In this table, we examine the TV error bound for gradient sparse signals with dimension $n = 1000$. The upper-bound U_δ is obtained by Monte Carlo simulations and numerical optimization. The true sample complexity $\delta(\mathcal{D}(\|\cdot\|_{\mathrm{TV}}, \boldsymbol{x})) \approx \omega^2(\mathcal{D}(\|\cdot\|_{\mathrm{TV}}, \boldsymbol{x}) \cap \mathbb{B}^n)$ is obtained using the approach presented in Appendix E. The large values of $\frac{\widehat{E}_p}{n}$ observed in this table are not necessarily caused by approximating $\delta(\mathcal{D}(f, \boldsymbol{x}))$ with U_δ.

where $\boldsymbol{w}, \boldsymbol{c} \in \mathbb{R}^n$ are arbitrary vectors. For tall analysis operators, we can find pairs of $\boldsymbol{\Omega}$ and $\boldsymbol{x}$ for which the error estimate [1] explodes. We precisely investigate this in Section IV.

2) *Obtaining an error bound for $\delta(\mathcal{D}(f, \boldsymbol{x}))$ with rather general $f(\cdot)$:* $\delta(\mathcal{D}(f, \boldsymbol{x}))$ precisely determines the boundary of failure and success of P_f. However, exact computation of $\delta(\mathcal{D}(f, \boldsymbol{x}))$ is very difficult. It is common to approximate $\delta(\mathcal{D}(f, \boldsymbol{x}))$ with U_δ. By providing an error bound, we formally show that this approximation is good. More precisely, we show that

$$\frac{E_\delta}{n} \leqslant h_2(\beta, \omega), \qquad (9)$$

where β depends on $\partial f(\boldsymbol{x})$, and $h_2(\beta, \omega)$ is a function of β and $\omega(\mathcal{D}(f, \boldsymbol{x}) \cap \mathbb{B}^n)$ that is succinctly shown by ω. Under certain conditions, we show that $h_2(\beta, \omega)$ vanishes as n grows sufficiently large. To a great extent, the setting considered for f (see (30)) is nonrestrictive. In particular, it includes the important special cases of $\|\boldsymbol{\Omega} \cdot \|_1$ for tall and fat analysis operators.

In contrast to the error estimate of [1] that directly depends on $\boldsymbol{x} \in \mathbb{R}^n$, our bound is determined by $\partial f(\boldsymbol{x})$. Besides, our error bound holds even for rank-deficient fat $\boldsymbol{\Omega} \in \mathbb{R}^{p \times n}$ matrices. We should emphasize that our error estimate bound is not sharp in all cases of analysis operators and does not necessarily fill the gap between E_δ and $\widehat{E}_p$ in (21). In fact, there are various settings in which the bound in [1], our bound, or both are effective.

C. Notation

Throughout the paper, scalars are denoted by lowercase letters, vectors by lowercase boldface letters, and matrices by uppercase boldface letters. The ith element of the vector $\boldsymbol{x}$ is given either by $x(i)$ or x_i. The notation $(\cdot)^\dagger$ stands for the pseudo-inverse operator. We reserve calligraphic uppercase letters for sets (e.g. $\mathcal{S}$) and denote the cardinality of a set $\mathcal{S}$ by $|\mathcal{S}|$. The complement of a set $\mathcal{S}$ in $\{1, ..., n\}$ (briefly represented as $[n]$) is denoted by $\overline{\mathcal{S}}$. Similarly, the complement of an event $\mathcal{E}$ is shown by $\overline{\mathcal{E}}$. For a matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ and a subset $\mathcal{S} \subseteq [n]$, the notation $\boldsymbol{X}_\mathcal{S}$ refers to the sub-matrix of $\boldsymbol{X}$ by including the rows indexed by $\mathcal{S}$. Similarly, for $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x}_\mathcal{S}$ stands for the vector in $\mathbb{R}^n$ that coincides with $\boldsymbol{x}$ at entries indexed by $\mathcal{S}$ and zero elsewhere. Also, we use the notation $\widetilde{\boldsymbol{x}}_\mathcal{S}$ to represent a sub-vector of $\boldsymbol{x}$ in $\mathbb{R}^{|\mathcal{S}|}$, that is formed by discarding the zero entries not indexed in $\mathcal{S}$. The null-space of linear operators is denoted by $\mathrm{null}(\cdot)$. For a matrix $\boldsymbol{\Omega}$, the operator norm is defined as $\|\boldsymbol{\Omega}\|_{p \to q} = \sup\limits_{\|\boldsymbol{x}\|_p \leqslant 1} \|\boldsymbol{\Omega}\boldsymbol{x}\|_q$. Also, $\kappa(\boldsymbol{\Omega}) := \frac{\sigma_{\max}(\boldsymbol{\Omega})}{\sigma_{\min}(\boldsymbol{\Omega})}$ denotes the condition number of $\boldsymbol{\Omega}$. The polar $\mathcal{K}^\circ$ of a cone $\mathcal{K} \subset \mathbb{R}^n$ is the set of vectors forming non-acute angles with every vector in $\mathcal{K}$, i.e.

$$\mathcal{K}^\circ = \{\boldsymbol{v} \in \mathbb{R}^n : \langle \boldsymbol{v}, \boldsymbol{z} \rangle \leqslant 0 \; \forall \boldsymbol{z} \in \mathcal{K}\}. \qquad (10)$$

$\mathbb{B}^n$ and $\mathbb{S}^{n-1}$ stand for the unit ball $\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 \leqslant 1\}$ and unit sphere $\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 = 1\}$, respectively. $\boldsymbol{P}_\mathcal{C}$ is the matrix associated with the orthogonal projection onto the subspace $\mathcal{C}$, that maps a vector in $\mathbb{R}^n$ onto the subspace $\mathcal{C} \subset \mathbb{R}^n$.

D. Outline

The paper is organized as follows. The required concepts from convex geometry are reviewed in Section II. Section III discusses two approaches in obtaining the error estimate. Section V is dedicated to present our main contributions. In Section IV, we investigate the estimate in [1] and introduce some examples for which the error estimate does not work. In Section VI, numerical experiments are presented which confirm our theory. Finally, the paper is concluded in Section VII.

II. CONVEX GEOMETRY

In this section, a review of basic concepts of convex geometry is provided.

A. Descent Cones

The descent cone $\mathcal{D}(f, \boldsymbol{x})$ at a point $\boldsymbol{x} \in \mathbb{R}^n$ consists of the set of directions that do not increase f and is given by:

$$\mathcal{D}(f, \boldsymbol{x}) = \bigcup_{t \geqslant 0} \{\boldsymbol{z} \in \mathbb{R}^n : f(\boldsymbol{x} + t\boldsymbol{z}) \leqslant f(\boldsymbol{x})\}. \qquad (11)$$

The descent cone reveals the local behavior of f near $\boldsymbol{x}$ and is a convex set. There is also a relation between decent cone and subdifferential [16, Chapter 23] given by:

$$\mathcal{D}^\circ(f, \boldsymbol{x}) = \mathrm{cone}(\partial f(\boldsymbol{x})) := \bigcup_{t \geqslant 0} t\partial f(\boldsymbol{x}). \qquad (12)$$

B. Statistical Dimension

Definition 1. Statistical Dimension [1]: Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a closed convex cone. The statistical dimension of $\mathcal{C}$ is defined as:

$$\delta(\mathcal{C}) := \mathbb{E}\|\mathcal{P}_\mathcal{C}(\boldsymbol{g})\|_2^2 = \mathbb{E} \, \mathrm{dist}^2(\boldsymbol{g}, \mathcal{C}^\circ), \qquad (13)$$

where $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$ is the projection of $\boldsymbol{x} \in \mathbb{R}^n$ onto the set $\mathcal{C}$ defined as: $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) = \underset{\boldsymbol{z} \in \mathcal{C}}{\arg\min} \|\boldsymbol{z} - \boldsymbol{x}\|_2$.

The statistical dimension extends the concept of linear subspaces to convex cones. Intuitively, it measures the size of a cone. Furthermore, $\delta(\mathcal{D}(f, \boldsymbol{x}))$ determines the precise location of transition from failure to success in P_f.

C. Gaussian width

Definition 2. The Gaussian width of a set C is defined as:

$$\omega(C) := \mathbb{E} \sup_{\boldsymbol{y} \in C} \langle \boldsymbol{y}, \boldsymbol{g} \rangle. \tag{14}$$

The relation between statistical dimension and Gaussian width is summarized in the following [2, Proposition 3.6], [1, Proposition 10.2].

$$\omega(C \cap \mathbb{S}^{n-1}) \leqslant \omega(C \cap \mathbb{B}^n) = \mathbb{E}\|\mathcal{P}_C(\boldsymbol{g})\|_2 = \mathbb{E} \operatorname{dist}(\boldsymbol{g}, C^\circ), \tag{15}$$

$$\omega^2(C \cap \mathbb{B}^n) = (\mathbb{E}\|\mathcal{P}_C(\boldsymbol{g})\|_2)^2 \leqslant \delta(C). \tag{16}$$

It is shown in [1, Proposition 10.2]. that the quantities $\omega^2(C \cap \mathbb{B}^n)$ and $\delta(C)$ differ numerically by at most 1.

D. Optimality Condition

In the following, we characterize when P_f succeeds in the noise-free case.

Proposition 1. *[2, Proposition 2.1] Optimality condition: Let f be a proper convex function. The vector $\boldsymbol{x} \in \mathbb{R}^n$ is the unique optimal point of P_f if and only if $\mathcal{D}(f, \boldsymbol{x}) \cap \operatorname{null}(\boldsymbol{A}) = \{\boldsymbol{0}\}$.*

The next theorem determines the number of measurements needed for successful recovery of P_f for any proper convex function f.

Theorem 1. *[1, Theorem 2]: Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ be a proper convex function and $\boldsymbol{x} \in \mathbb{R}^n$ a fixed vector. Suppose that m independent Gaussian linear measurements of $\boldsymbol{x}$ are observed via $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \in \mathbb{R}^m$. If*

$$m \geqslant \delta(\mathcal{D}(f, \boldsymbol{x})) + \sqrt{8\log(\frac{4}{\eta})n}, \tag{17}$$

for a given probability of failure (tolerance) $\eta \in [0, 1]$, then, we have

$$\mathbb{P}(\mathcal{D}(f, \boldsymbol{x}) \cap \operatorname{null}(\boldsymbol{A}) = \{\boldsymbol{0}\}) \geqslant 1 - \eta. \tag{18}$$

Besides, if

$$m \leqslant \delta(\mathcal{D}(f, \boldsymbol{x})) - \sqrt{8\log(\frac{4}{\eta})n}, \tag{19}$$

then,

$$\mathbb{P}(\mathcal{D}(f, \boldsymbol{x}) \cap \operatorname{null}(\boldsymbol{A}) = \{\boldsymbol{0}\}) \leqslant \eta. \tag{20}$$

III. RELATED WORKS IN ERROR ESTIMATION

For bounding the distance between $\delta(\mathcal{D}(f, \boldsymbol{x}))$ and U_δ, two different approaches are proposed in [1], [17]. In the following, we briefly describe these methods.

Result 1. *[1, Theorem 4.3] Let f be a norm. Then, for any $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$:*

$$0 \leqslant E_\delta \leqslant \frac{\overbrace{2 \sup_{s \in \partial f(\boldsymbol{x})} \|s\|_2}^{\text{Num}_{\text{E}}}}{f\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}\right)} := \widehat{E}_p. \tag{21}$$

Result 2. *[17, Proposition 1] Suppose that for $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$, $\partial f(\boldsymbol{x})$ satisfies a weak decomposability assumption:*

$$\exists \boldsymbol{z}_0 \in \partial f(\boldsymbol{x}) \text{ s.t. } \langle \boldsymbol{z} - \boldsymbol{z}_0, \boldsymbol{z}_0 \rangle = 0, \quad \forall \boldsymbol{z} \in \partial f(\boldsymbol{x}). \tag{22}$$

Then[4],

$$\inf_{t \geqslant 0} \mathbb{E} \operatorname{dist}(\boldsymbol{g}, t\partial f(\boldsymbol{x})) \leqslant \omega(\mathcal{D}(f, \boldsymbol{x}) \cap \mathbb{B}^n) + 6. \tag{23}$$

A. Explanations

Result 2 presents an error estimate for the Gaussian width of the descent cone (restricted to the unit ball) that is used to upper-bound the number of Gaussian measurements in various low-dimensional structures [2, Section 3.1]. For functions $f = \|\cdot\|_1, \|\cdot\|_{1,2}$ and $\|\cdot\|_*$, the constraint (22) is satisfied (see Section V-A). For $f = \|\boldsymbol{\Omega} \cdot \|_1$, however, this constraint is not generally guaranteed (see Appendix F).

Unlike Result 2, the error estimate (21) depends on ∂f at the ground-truth vector $\boldsymbol{x}$, and the vector $\boldsymbol{x}$ itself. Although [1, Theorem 4.3] restricts f to be a norm, the provided proof remains valid for semi-norms such as TV. The error bound in (21), is effective for many structure-inducing functions including $\ell_1, \ell_{1,2}$, and nuclear norm. Particularly, $\frac{\widehat{E}_p}{n}$ asymptotically vanishes in these cases. However, the normalized error estimate $\frac{\widehat{E}_p}{n}$ is large in some cases of ℓ_1 analysis and TV minimization and does not reflect the actual error $\frac{E_\delta}{n}$; in Section IV-A, we study some examples. A naive interpretation of this fact is that U_δ is a poor approximation of $\delta(\mathcal{D}(f, \boldsymbol{x}))$ in those cases. Fortunately, as we show in Section V, this argument is invalid, which in turn suggests that (21) is a loose bound in those cases.

IV. THE STUDY OF EXISTING RESULTS

A. Result 1 for various low-complexity models

Before we describe our contributions in Section V, we first evaluate the error estimate (21) when f is any of $\ell_1, \ell_{1,2}$, nuclear norm, or ℓ_1-analysis for different analysis operators. An important observation is that the error estimate (21) is increasing with $\frac{\|\boldsymbol{x}\|_2}{f(\boldsymbol{x})}$; thus, whenever this term becomes large, we might obtain a loose upper-bound. To better clarify this point, we study the case of ℓ_1-analysis in three categories of fat analysis operators $\boldsymbol{\Omega}$, rank-deficient tall analysis operators $\boldsymbol{\Omega}$, and full-rank tall analysis operators $\boldsymbol{\Omega}$.

---

[4]This result is wrongly interpreted as $\inf_{t \geqslant 0} \sqrt{\mathbb{E} \operatorname{dist}^2(\boldsymbol{g}, t\partial f(\boldsymbol{x}))} \leqslant \omega(\mathcal{D}(f, \boldsymbol{x}) \cap \mathbb{B}^n) + 6$ in [17].

- **Sparse vectors**
Since $\delta\big(\mathcal{D}(\|\cdot\|_1, \boldsymbol{x})\big) = \delta\big(\mathcal{D}(\|\cdot\|_1, \text{sgn}(\boldsymbol{x}))\big)$, the bound $\widehat{E}_p$ in (21) can be written in terms of $\text{sgn}(\boldsymbol{x})$:

$$\widehat{E}_p = \frac{2\sup_{\|\boldsymbol{z}_{\overline{\mathcal{S}}}\|_\infty \leqslant 1} \|\text{sgn}(\boldsymbol{x}) + \boldsymbol{z}_{\overline{\mathcal{S}}}\|_2}{\frac{\|\text{sgn}(\boldsymbol{x})\|_1}{\|\text{sgn}(\boldsymbol{x})\|_2}}$$

$$\geqslant \frac{2\|\text{sgn}(\boldsymbol{x}) + \mathbf{1}_{\overline{\mathcal{S}}}\|_2}{\sqrt{s}} = \frac{2\sqrt{n}}{\sqrt{s}}, \qquad (24)$$

where the inequality is for choosing a point $\mathbf{1}_{\overline{\mathcal{S}}}$ in the feasible set $\|\boldsymbol{z}_{\overline{\mathcal{S}}}\|_\infty \leqslant 1$. The above expression may lead to large errors in low sparsity regimes ($s \ll n$), however, $\lim_{n\to\infty} \frac{\widehat{E}_p}{n} = 0$. This shows that U_δ is asymptotically a fair approximation of $\delta\big(\mathcal{D}(\|\cdot\|_1, \boldsymbol{x})\big)$.

- **Block-sparse Vectors**
With the same approach as in the previous case, the actual error estimate is lower-bounded by

$$\widehat{E}_p \geqslant \frac{2\sqrt{q}}{\sqrt{s}} \qquad (25)$$

where s and q stand for the number of non-zero blocks and the total number of blocks respectively. Again for small s, the error can become large while $\lim_{q\to\infty} \frac{\widehat{E}_p}{q} = 0$; thus, U_δ is asymptotically a fair approximation of $\delta\big(\mathcal{D}(\|\cdot\|_{1,2}, \boldsymbol{x})\big)$.

- **Low-rank Matrices**
Let $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ be the rank r ground-truth matrix ($n_1 \geqslant n_2$) with the SVD decomposition $\boldsymbol{X} = \boldsymbol{U}_{n_1 \times n_1} \boldsymbol{\Sigma}_{n_1 \times n_2} \boldsymbol{V}^H_{n_2 \times n_2}$ (alternatively, we have the reduced SVD decomposition as $\boldsymbol{X} = \boldsymbol{U}_{n_1 \times r} \boldsymbol{\Sigma}_{r \times r} \boldsymbol{V}^H_{n_2 \times r}$). Since $\delta\big(\mathcal{D}(\|\cdot\|_*, \boldsymbol{X})\big) = \delta\big(\mathcal{D}(\|\cdot\|_*, \boldsymbol{U}_{n_1 \times r} \boldsymbol{V}^H_{n_2 \times r})\big)$, we can replace $\boldsymbol{X}$ with $\boldsymbol{U}_{n_1 \times r} \boldsymbol{V}^H_{n_2 \times r}$ in (21):

$$\widehat{E}_p = \frac{2\sup_{\|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\|_{2\to 2} \leqslant 1} \|\boldsymbol{U}_{n_1 \times r} \boldsymbol{V}^H_{n_2 \times r} + \mathcal{P}_{T^\perp}(\boldsymbol{Z})\|_F}{\frac{\|\boldsymbol{U}_{n_1 \times r} \boldsymbol{V}^H_{n_2 \times r}\|_*}{\|\boldsymbol{U}_{n_1 \times r} \boldsymbol{V}^H_{n_2 \times r}\|_F}}, \qquad (26)$$

where

$$\mathcal{P}_{T^\perp}(\boldsymbol{Z}) = \Big(\boldsymbol{I} - \boldsymbol{P}_{\text{span}(\boldsymbol{U}_{n_1 \times r})}\Big) \boldsymbol{Z} \Big(\boldsymbol{I} - \boldsymbol{P}_{\text{span}(\boldsymbol{V}_{n_2 \times r})}\Big).$$

Now, by setting

$$\boldsymbol{Z} = \boldsymbol{U}_{n_1 \times n_2} \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n_2 - r} \end{bmatrix} \boldsymbol{V}^H_{n_2 \times n_2}$$

in (26), we obtain a lower-bound on $\widehat{E}_p$ as

$$\widehat{E}_p \geqslant \frac{2\|\boldsymbol{U}_{n_1 \times n_2} \boldsymbol{V}^H_{n_2 \times n_2}\|_F}{\frac{\|\boldsymbol{U}_{n_1 \times r} \boldsymbol{V}^H_{n_2 \times r}\|_*}{\|\boldsymbol{U}_{n_1 \times r} \boldsymbol{V}^H_{n_2 \times r}\|_F}} = \frac{2\sqrt{n_2}}{\sqrt{r}}. \qquad (27)$$

Similar to the previous cases, when $r \ll n_2$, the bound becomes large, while, $\lim_{n_1, n_2 \to \infty} \frac{\widehat{E}_p}{n_1 n_2} = 0$. This shows that U_δ is asymptotically a fair approximation of $\delta\big(\mathcal{D}(\|\cdot\|_*, \boldsymbol{x})\big)$.

- **Cosparse vectors (fat analysis operators)**
For fat $\boldsymbol{\Omega} \in \mathbb{R}^{p \times n}$, the $\text{null}(\boldsymbol{\Omega})$ is non-trivial, and we can choose $\boldsymbol{c}$ such that $\boldsymbol{P}_{\text{null}(\boldsymbol{\Omega})} \boldsymbol{c} \neq \boldsymbol{0}$. Now, if $\boldsymbol{x} =$

$\boldsymbol{P}_{\text{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{w} + \boldsymbol{P}_{\text{null}(\boldsymbol{\Omega})} \boldsymbol{c}$, where $\boldsymbol{w}$ is an arbitrary vector, the denominator of the bound in (21) can be written as

$$\frac{\|\boldsymbol{\Omega}\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_2} = \frac{\|\boldsymbol{\Omega}_{\mathcal{S}} \boldsymbol{P}_{\text{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{w}\|_1}{\|\boldsymbol{P}_{\text{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{w} + \boldsymbol{P}_{\text{null}(\boldsymbol{\Omega})} \boldsymbol{c}\|_2}. \qquad (28)$$

By increasing the norm of $\boldsymbol{c}$ using a scalar multiplier, the above fraction decreases. In other words, we can make the denominator of the error bound (21) arbitrarily small (alternatively enlarge the error bound (21)). One of the well-known examples in this category is the finite difference operator $\boldsymbol{\Omega}_d$, where $\text{null}(\boldsymbol{\Omega}_d)$ consists of constant vectors. For this example, the denominator can be reduced by setting $\boldsymbol{c} = \alpha \mathbf{1}_{n \times 1} \in \text{null}(\boldsymbol{\Omega}_d)$ and $\alpha \gg 1$.

- **Cosparse vectors (rank-deficient tall analysis operators)**
Similar to the previous case, $\text{null}(\boldsymbol{\Omega})$ is non-trivial. Thus, the same approach can be devised to make the bound in (21) arbitrarily large.

- **Cosparse vectors (full-rank tall analysis operators)**
When $\boldsymbol{\Omega}$ is a full-rank and tall matrix, we cannot generally find $\boldsymbol{x}$ that results in a small value of $\frac{\|\boldsymbol{\Omega}\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_2}$. Here, we show the existence of pairs $(\boldsymbol{\Omega}, \boldsymbol{x})$ for which the aforementioned ratio becomes arbitrarily small.
Let $|\mathcal{S}| > n$, $|\overline{\mathcal{S}}| < n$, $\sigma > 0$ and define

$$\boldsymbol{A}_{|\mathcal{S}| \times n} = \boldsymbol{U}_{|\mathcal{S}| \times n} \text{diag}\{\underbrace{1, 1, \ldots, 1}_{(n-1)\text{ times}}, \sigma\} \boldsymbol{V}^T_{n \times n},$$

where $\boldsymbol{U}_{|\mathcal{S}| \times n}$ and $\boldsymbol{V}_{n \times n}$ are arbitrary incomplete and complete unitary matrices, respectively. It is evident that $\boldsymbol{A}_{|\mathcal{S}| \times n}$ is a full-rank tall matrix. Let $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ stand for the ith and jth columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. We know that $\boldsymbol{A}\boldsymbol{v}_n = \sigma\boldsymbol{u}_n$. Now, define $\boldsymbol{B}_{|\overline{\mathcal{S}}| \times n} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{|\overline{\mathcal{S}}|}]^T$ (since $|\overline{\mathcal{S}}| < n$, this is possible). Obviously, $\boldsymbol{B}\boldsymbol{v}_n = \boldsymbol{0}$. Finally, we define

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{bmatrix}, \quad \boldsymbol{x} = \boldsymbol{v}_n, \quad \mathcal{S} = \{1, \ldots, |\mathcal{S}|\},$$
$$\overline{\mathcal{S}} = \{|\mathcal{S}| + 1, \ldots, |\mathcal{S}| + |\overline{\mathcal{S}}|\}.$$

Indeed, the design is such that $\boldsymbol{\Omega}_{\mathcal{S}} = \boldsymbol{A}$ and $\boldsymbol{\Omega}_{\overline{\mathcal{S}}} = \boldsymbol{B}$. As $\boldsymbol{A}$ has full column rank, $\boldsymbol{\Omega}$ is also a full-rank matrix. It is straightforward to check that $\boldsymbol{x} \in \text{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})$, and

$$\frac{\|\boldsymbol{\Omega}\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_2} = \|\boldsymbol{\Omega}\boldsymbol{x}\|_1 = \left\|\begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{bmatrix} \boldsymbol{v}_n\right\|_1 = \|\boldsymbol{A}\boldsymbol{v}_n\|_1 = \sigma\|\boldsymbol{u}_n\|_1.$$

Now, we can reduce σ while keeping the rest untouched. In this way, we construct non-trivial pairs of $(\boldsymbol{\Omega}, \boldsymbol{x})$ for which the ratio $\frac{\|\boldsymbol{\Omega}\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_2}$ can be set arbitrarily small.

Remark 1. The value of $\widehat{E}_p$ provides an upper-bound on the gap between U_δ and $\delta\big(\mathcal{D}(f, \boldsymbol{x})\big)$. However, its normalized value (e.g., $\frac{\widehat{E}_p}{n}$ in the ℓ_1 minimization) is important in determining the phase transition curve. As we discussed earlier, the normalized value is vanishing in the three cases of ℓ_1, $\ell_{1,2}$ and nuclear-norm minimization. However, we do not observe this vanishing property in some cases of ℓ_1-analysis. In these cases, U_δ might not be a good approximation of $\delta(\mathcal{D}(f, \boldsymbol{x}))$.

Type of anal. operator $\boldsymbol{\Omega}$	p	n	r	s	$\kappa(\boldsymbol{\Omega})$	β	$\mathrm{Num}_{\mathrm{E}}$	$\frac{\hat{E}_p}{n}$	$\frac{\hat{E}_n}{n}$	$\frac{E_\delta}{n}$
Random 1	1000	999	944	10	∞	1.3	85.14	139.13	0.12	0.005
Random 1	1000	999	999	2	100	1	63.04	4.85	0.051	0.006
Random 1	2000	1990	1990	11	50	1	43	2.04	0.07	0.001
Random 1*	2000	1995	1995	10	1.94	1.05	82.5	0.08	0.05	0.002
Random 2	1000	995	995	6	483	1.2	59.3	2.72	0.03	0.002
High-pass Daubechies wavelet (2 decom. level)	2048	1024	1023	600	∞	1.3	101.64	22.7	0.1	0.005

Table III: We examine the function $f = \|\boldsymbol{\Omega} \cdot \|_1$ when $\boldsymbol{\Omega}$ is a tall analysis operator. Three types of operators are considered: Random 1, Random 2 and wavelet. These operators are constructed using the procedure proposed in Items 1, 2, and 3 in Section VI, respectively. For the wavelet case, we construct a 3072×1024 Daubechies wavelet transform where we only retain its high-pass coefficients (of size 2048×1). The number of decomposition levels in the wavelet transformation is two. Except for Random 1*, we construct cosparse signals according to the procedure explained in Section VI. For Random 1* we use $\boldsymbol{x} = \boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{S}})}\boldsymbol{c}$, where $\boldsymbol{c}$ is uniformly distributed on the unit sphere $\mathbb{S}^{n-1}$. In this table, $\mathrm{Num}_{\mathrm{E}}$ denotes the numerator of $\hat{E}_p$ and is equal to $2\sup_{s \in \partial f(\boldsymbol{x})} \|s\|_2$.

Type of Analysis operator $\boldsymbol{\Omega}$	p	n	r	s	β	$\mathrm{Num}_{\mathrm{E}}$	$\frac{\hat{E}_p}{n}$	$\frac{\hat{E}_n}{n}$	$\frac{E_\delta}{n}$
Random 1	1490	1500	1490	5	1	71.76	51.93	0.056	0.001
Random 1	1999	2000	1999	10	1	83.06	5.8	0.0136	0.008
Random 1	1999	2000	1997	10	1	82.72	18.38	0.04	0.001
Random 2	1999	2000	1999	10	1.05	124.3	16.15	0.008	0.002
TV	1999	2000	1999	350	1	124.76	9.94	0.08	0.001
TV	1999	2000	1999	10	1	123.72	891.3	0.003	0.001
TV*	1999	2000	1999	10	1	123.12	0.2	0.002	0.001
Low-pass Daubechies wavelet (1 decom. level)	2048	2048	2047	10	1.05	85.8	130.36	0.071	0.002
High-pass Daubechies wavelet (1 decom. level)	1024	1024	1023	8	1.08	63.4	13.9	0.1	0.005

Table IV: We examine the function $f = \|\boldsymbol{\Omega} \cdot \|_1$ when $\boldsymbol{\Omega}$ is a fat analysis operators. Four types of operators are considered: Random 1, Random 2, finite difference, and wavelet. The cases of Random 1 and 2 are constructed using the procedure explained in Items 1 and 2 in Section VI, respectively. We build the wavelet matrices with Daubechies structure where we retain low- and high-pass coefficients. The considered number of decomposition levels is 1. As it is clear, our error bound $\hat{E}_n$ outperforms the previous error estimate (21) denoted by $\hat{E}_p$ in all cases. Except for TV*, we construct cosparse signals according to the procedure explained in Section VI. For TV* we use $\boldsymbol{x} = \boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{S}})}\boldsymbol{c}$, where $\boldsymbol{c}$ is uniformly distributed on the unit sphere $\mathbb{S}^{n-1}$. In this table, $\mathrm{Num}_{\mathrm{E}}$ denotes the numerator of $\hat{E}_p$ and is equal to $2\sup_{s \in \partial f(\boldsymbol{x})} \|s\|_2$.

B. Weak decomposability condition

For functions $f = \{\|\cdot\|_1, \|\cdot\|_{1,2}, \|\cdot\|_*\}$, one can always find a vector $\boldsymbol{z}_0 \in \partial f(\boldsymbol{x})$ such that the weak decomposability assumption (22) holds. More precisely, one can use [18, Definition 2]:

$$\boldsymbol{z}_0 = \mathrm{sgn}(\boldsymbol{x}) \quad \text{for } f = \|\cdot\|_1,$$
$$\boldsymbol{z}_0 = \frac{\boldsymbol{x}_{\mathcal{V}_b}}{\|\boldsymbol{x}_{\mathcal{V}_b}\|_2} \quad \text{for } f = \|\cdot\|_{1,2},$$
$$\boldsymbol{Z}_0 = \boldsymbol{U}_{n_1 \times r}\boldsymbol{V}_{n_2 \times r}^H \quad \text{for } f = \|\cdot\|_*, \tag{29}$$

where $\boldsymbol{U}_{n_1 \times r}$ and $\boldsymbol{V}_{n_2 \times r}$ are the bases corresponding to the reduced singular value decomposition of the ground-truth matrix $\boldsymbol{X} := \boldsymbol{U}_{n_1 \times r}\boldsymbol{\Sigma}_{r \times r}\boldsymbol{V}_{r \times n_2}^H$. The sets $\{\mathcal{V}_b\}_{b=1}^q$ stand for a partitioning of $\{1, ..., n\}$ into q blocks of equal length. In [15], it is shown that $f = \|\cdot\|_{\mathrm{TV}}$ satisfies the weak decomposability assumption (22), i.e., there exists $\boldsymbol{z}_0 \in \partial\|\cdot\|_{\mathrm{TV}}(\boldsymbol{x})$ which satisfies (22). In Section V-A, we show that the weak decomposability condition does not necessarily hold for the general family of $f = \|\boldsymbol{\Omega} \cdot \|_1$; in particular, we construct counter-examples for the case of full-rank tall analysis operators.

V. MAIN RESULTS

Our main results which are stated in the following theorem, estimate the distance between $\delta(\mathcal{D}(f, \boldsymbol{x}))$ and its corresponding upper-bound.

Theorem 2. *Let f be a proper convex function that promotes the structure of $\boldsymbol{x} \neq \boldsymbol{0} \in \mathbb{R}^n$ and let $\boldsymbol{g} \in \mathbb{R}^n$ be a standard i.i.d Gaussian vector. Suppose $\partial f(x)$ satisfies*

$$\exists \boldsymbol{z}_0 \quad \text{s.t.} \ \langle \boldsymbol{z} - \boldsymbol{z}_0, \boldsymbol{z}_0 \rangle = 0, \quad \forall \boldsymbol{z} \in \partial f(\boldsymbol{x}). \tag{30}$$

Then for any positive values of λ, ζ, we have that

$$0 \leqslant E_\delta \leqslant (4\lambda\beta + \gamma)\ \omega(\mathcal{D}(f, \boldsymbol{x}) \cap \mathbb{B}^n) + \gamma(\zeta + 2\lambda\beta) + 4\lambda^2\beta^2$$
$$:= \hat{E}_n, \tag{31}$$

and

$$0 \leqslant \inf_{t \geqslant 0} \mathbb{E}\ \mathrm{dist}(\boldsymbol{g}, t\partial f(\boldsymbol{x})) - \omega(\mathcal{D}(f, \boldsymbol{x}) \cap \mathbb{B}^n) \leqslant 1.6 + 4\beta, \tag{32}$$

where γ is the constant

$$\gamma = \sqrt{72}\sqrt{\ln \frac{3}{1 - 4e^{-\frac{\lambda^2}{2}} - 2e^{-\frac{\zeta^2}{2}}}}, \tag{33}$$

and β is given by

$$\beta = \frac{\|\boldsymbol{z}_1\|_2}{\|\boldsymbol{z}_0\|_2}, \tag{34}$$

where

$$\boldsymbol{z}_1 = \arg\min_{\boldsymbol{z} \in \partial f(\boldsymbol{x})} \|\boldsymbol{z}\|_2. \tag{35}$$

Proof. See Appendices A and B.

Remark 2. The error estimate in (31) is the main result of Theorem 2, while (32) can be thought of as an extension of Result 2 to more general structure-inducing functions (including ℓ_1-analysis). Despite the similarities between $\inf_{t \geqslant 0} \mathbb{E}\ \mathrm{dist}(\boldsymbol{g}, t\partial f(\boldsymbol{x}))$ and $\inf_{t \geqslant 0} \sqrt{\mathbb{E}\ \mathrm{dist}^2(\boldsymbol{g}, t\partial f(\boldsymbol{x}))}$, these two terms are different, and (32) cannot be considered as an error bound. We should add that (32) is used in our proof of (31).

Remark 3. If β is bounded, as $\omega(\mathcal{D}(f,\boldsymbol{x})\cap\mathbb{B}^n)\leqslant\sqrt{n}$, $\frac{\widehat{E}_n}{n}$ asymptotically tends to 0. We numerically observe in Section V-A that for functions $f=\{\|\cdot\|_1,\|\cdot\|_{1,2},\|\cdot\|_*,\|\boldsymbol{\Omega}\cdot\|_1\}$, β is bounded in most cases of $\boldsymbol{\Omega}$. In case of ℓ_1 analysis, β can be upper-bounded by a function of the *generalized sign vector* of $\boldsymbol{x}$ (i.e. $\boldsymbol{\Omega}^T\mathrm{sgn}(\boldsymbol{\Omega}\boldsymbol{x})$).

When β is bounded in the ℓ_1-analysis case, our bound in Theorem 2 implies that the normalized error gap is vanishing asymptotically; equivalently, it implies that U_δ is a good estimate of $\delta\big(\mathcal{D}(\|\boldsymbol{\Omega}\cdot\|_1,\boldsymbol{x})\big)$. Our result holds for various analysis operators including the ones that have non-trivial linear dependencies among their rows. It should be noted that we do not guarantee the boundedness of β; in case β fails to remain bounded, our error estimate is no longer effective.

A. Evaluation of β

To compute β, we need both $\boldsymbol{z}_1$ in (35) and $\boldsymbol{z}_0$ in (30). It is not difficult to see that for functions $f=\{\|\cdot\|_1,\|\cdot\|_{1,2},\|\cdot\|_*\}$, $\boldsymbol{z}_1$ in (35) is obtained by

$$
\begin{aligned}
\boldsymbol{z}_1 &= \mathrm{sgn}(\boldsymbol{x}) \quad \text{for } f=\|\cdot\|_1,\\
\boldsymbol{z}_1 &= \frac{\boldsymbol{x}_{\mathcal{V}_b}}{\|\boldsymbol{x}_{\mathcal{V}_b}\|_2} \quad \text{for } f=\|\cdot\|_{1,2},\\
\boldsymbol{Z}_1 &= \boldsymbol{U}_{n_1\times r}\boldsymbol{V}_{n_2\times r}^H \quad \text{for } f=\|\cdot\|_*,
\end{aligned}
\tag{36}
$$

where $\boldsymbol{U}_{n_1\times r}$ and $\boldsymbol{V}_{n_2\times r}$ are bases corresponding to the reduced singular value decomposition of the ground-truth matrix $\boldsymbol{X}:=\boldsymbol{U}_{n_1\times r}\boldsymbol{\Sigma}_{r\times r}\boldsymbol{V}_{r\times n_2}^H$.

Choosing $\boldsymbol{z}_0=\boldsymbol{z}_1$ in the subdifferential is common for functions $f=\{\|\cdot\|_1,\|\cdot\|_{1,2},\|\cdot\|_*\}$ [17], [18]. Unlike these simple choices, obtaining $\boldsymbol{z}_0$ for the cosparse vectors is more involved. In the following proposition, we discuss this issue when $f=\|\boldsymbol{\Omega}\cdot\|_1$, where $\boldsymbol{\Omega}$ is either a tall or a fat analysis operator.

Proposition 2. *Consider the cosparse vector $\boldsymbol{x}\in\mathbb{R}^n$ in the analysis domain $\boldsymbol{\Omega}\in\mathbb{R}^{p\times n}$ with support $\mathcal{S}$. Then,*

$$
\boldsymbol{z}_0 = \boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})}\boldsymbol{\Omega}^T\mathrm{sgn}(\boldsymbol{\Omega}\boldsymbol{x}),
\tag{37}
$$

satisfies (30).

Proof. See Appendix D.

In what follows, we examine some special and important implications of this proposition.

Remark 4. In case of $\boldsymbol{\Omega}=\boldsymbol{I}$ (i.e. $f=\|\cdot\|_1$), $\boldsymbol{z}_0=\mathrm{sgn}(\boldsymbol{x})$. This supports the fact that choosing $\boldsymbol{z}_0$ in the subdifferential is reasonable and efficient.

Remark 5. (Upper-bound on β) Employing (37), we can express β as

$$
\begin{aligned}
\beta &:= \frac{\inf_{\|\tilde{\boldsymbol{z}}_{\overline{\mathcal{S}}}\|_\infty\leqslant 1}\|\boldsymbol{\Omega}^T\mathrm{sgn}(\boldsymbol{\Omega}\boldsymbol{x})+\boldsymbol{\Omega}_{\overline{\mathcal{S}}}^H\tilde{\boldsymbol{z}}_{\overline{\mathcal{S}}}\|_2}{\|\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})}\boldsymbol{\Omega}^T\mathrm{sgn}(\boldsymbol{\Omega}\boldsymbol{x})\|_2}\\
&\leqslant \frac{\|\boldsymbol{\Omega}_{\mathcal{S}}^T\mathrm{sgn}(\boldsymbol{\Omega}_{\mathcal{S}}\boldsymbol{x})\|_2}{\|\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})}\boldsymbol{\Omega}_{\mathcal{S}}^T\mathrm{sgn}(\boldsymbol{\Omega}_{\mathcal{S}}\boldsymbol{x})\|_2}.
\end{aligned}
\tag{38}
$$

The boundedness of β is consistently observed in our numerical results (see Tables III and IV). We also prove the boundedness for some special cases of $\boldsymbol{\Omega}\in\mathbb{R}^{p\times n}$. For

instance, for fat matrices with orthogonal rows (a special case of which is investigated in Item 1 of Section VI), β equals 1. (see Appendix H for the proof). Another example is when the elements of $\boldsymbol{\Omega}$ are drawn from an i.i.d. Gaussian distribution. In this case, under the assumption $\frac{|\mathcal{S}|+1}{n}\leqslant\rho<1$, we have that

$$
\beta\leqslant\frac{1}{\sqrt{1-\rho}}
$$

with high probability in high dimensions (see Appendix G for the proof).

VI. Numerical experiments

In this section, we numerically compare the new error bound of (31) against the bound (21) derived using the existing approach for various low-dimensional structures. For each test, we optimize λ and ζ to minimize the right-hand side of (31). Figures 1, 2, and 3 show the proposed error bound (31) and the error estimates (24), (25), and (27), for $\|\cdot\|_1$, $\|\cdot\|_{1,2}$ and $\|\cdot\|_*$, respectively. In all cases, the sparsity/rank values are set very small. To compute $\omega(\mathcal{D}(f,\boldsymbol{x})\cap\mathbb{B}^n)$ in (31), we used its upper-bound obtained via (16), [1, Equations D.6, D.10], and [19, Lemma 1]. It is clear from these figures that the new error bound outperforms the previous error bound (21) in very low sparsity/rank regimes; it should be emphasized that the curves depict the upper-bound of (31). Notice that in these three cases, both $\frac{\widehat{E}_p}{n}$ and $\frac{E_n}{n}$ tend to zero at large n.

Due to the varying nature of the ℓ_1 analysis case, we construct three kinds of analysis operators as follows:

1) Random 1: We first generate a $p\times n$ Gaussian matrix with i.i.d. elements. Then, we compute its SVD as $\boldsymbol{U}_{p\times p}\boldsymbol{\Sigma}_{p\times n}\boldsymbol{V}_{n\times n}^H$. Then, $\boldsymbol{\Sigma}$ is replaced with the matrix

$$
\boldsymbol{\Sigma}_1 := \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{0}_{r\times n-r}\\ \boldsymbol{0}_{p-r\times r} & \boldsymbol{0}_{p-r\times n-r} \end{bmatrix}
\tag{39}
$$

to get

$$
\boldsymbol{\Omega} = \boldsymbol{U}_{p\times p}\boldsymbol{\Sigma}_1\boldsymbol{V}_{n\times n}^H.
\tag{40}
$$

When $r=n\leqslant p$, the constructed $\boldsymbol{\Omega}$ in (40) is a tight frame. To have an analysis operator with more varied singular values, we proceed with

$$
\boldsymbol{\Omega} = \boldsymbol{D}_{p\times p}\boldsymbol{U}_{p\times p}\boldsymbol{\Sigma}_1\boldsymbol{V}_{n\times n}^H,
\tag{41}
$$

where $\boldsymbol{D}_{p\times p}$ is a diagonal matrix. This type of matrices is widely used as a benchmark in [11], [13], [14]. The above approach was directly adopted from [14].

2) Random 2: In this case, we simply use the Gaussian ensemble by constructing a $p\times n$ matrix with i.i.d. elements that each follows $\mathcal{N}(0,\sigma^2)$. Here, σ^2 implicitly specifies the range of the singular values.

3) Wavelet: We choose a redundant wavelet transform from the package SPOT [20] to construct an analysis operator. The wavelet filter is chosen from the Daubechies family and has length 8. In some cases, we retain only the high-pass or low-pass coefficients. The rows of the wavelet operator $\boldsymbol{\Omega}$ might have non-trivial linear dependencies.

For a general analysis operator $\boldsymbol{\Omega}_{p\times n}$, we first randomly select $\mathcal{S}$ among the subsets of $\{1,...,p\}$ with size s. Then, we

check whether $\boldsymbol{\Omega}_{\mathcal{S}}\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})}$ is the trivial $\mathbf{0}$ operator or not. In the trivial case, we regenerate $\mathcal{S}$ and repeat the test; otherwise, we form $\boldsymbol{x}$ via

$$\boldsymbol{x} = \boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})}\boldsymbol{w},$$

where $\boldsymbol{w}$ is such that $\boldsymbol{\Omega}\boldsymbol{x} \neq \mathbf{0}$. In this paper, as we would like to highlight the difference between the new error bound and the existing ones, we focus on a subclass of analysis sparse vectors. More specifically, whenever there are non-trivial linear dependencies among the columns of the analysis operator (fat or tall), we generate $\boldsymbol{x}$ according to the procedure explained in Section IV-A. Whenever the analysis operator is of type Random 2 (with highly coherent rows), we generate $\boldsymbol{x}$ as

$$\boldsymbol{x} = \boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})}(\boldsymbol{w} + \alpha\boldsymbol{v}_r), \tag{42}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ are arbitrary quantities and $\boldsymbol{v}_r$ is the right singular vector of $\boldsymbol{\Omega}_{\mathcal{S}}\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})}$ corresponding to the minimum non-zero singular value (indeed, $r = \mathrm{rank}(\boldsymbol{\Omega}_{\mathcal{S}}\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})})$). With this choice, the denominator of the error bound (21) simplifies to

$$\frac{\|\boldsymbol{\Omega}_{\mathcal{S}}\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})}\boldsymbol{w} + \alpha\sigma_r\boldsymbol{u}_r\|_1}{\|\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})}\boldsymbol{w} + \alpha\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})}\boldsymbol{v}_r\|_2}, \tag{43}$$

where $\boldsymbol{u}_r$ is the rth left singular vector of $\boldsymbol{\Omega}_{\mathcal{S}}\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})}$. For highly coherent analysis operators where the minimum singular value is very small, by increasing α, the denominator of the error bound (21) is likely to decrease.

For the tall analysis operators of type Random 1 in (41), we construct pairs of $(\boldsymbol{\Omega}, \boldsymbol{x})$ as follows. We set

$$\boldsymbol{\Omega}_{\mathcal{S}} = \boldsymbol{D}_1\,\boldsymbol{U}_{|\mathcal{S}|\times n}\boldsymbol{V}^{\mathrm{H}},$$
$$\boldsymbol{\Omega}_{\overline{\mathcal{S}}} = \boldsymbol{D}_2\,\boldsymbol{U}_{|\overline{\mathcal{S}}|\times n}\boldsymbol{V}^{\mathrm{H}}, \tag{44}$$

where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are arbitrary diagonal matrices with non-negative values of size $|\mathcal{S}| \times |\mathcal{S}|$ and $|\overline{\mathcal{S}}| \times |\overline{\mathcal{S}}|$, respectively, and $\boldsymbol{U}_{|\mathcal{S}|\times n}$ is a sub-matrix of $\boldsymbol{U}_{p\times p}$ restricted to the rows and columns in $\mathcal{S}$ and $[n]$, respectively. We further generate $\boldsymbol{x}$ as

$$\boldsymbol{x} = \boldsymbol{P}_{\mathrm{null}(\boldsymbol{U}_{|\overline{\mathcal{S}}|\times n}\boldsymbol{V}^{\mathrm{H}})}(\boldsymbol{w} + \alpha\boldsymbol{v}'_{\mathrm{min}}), \tag{45}$$

where $\alpha > 0$ is an arbitrary real, $\boldsymbol{w}$ is an arbitrary vector in $\mathbb{R}^n$ and $\boldsymbol{v}'_{\mathrm{min}}$ is the right singular vector corresponding to the minimum singular value σ_{min} of $\boldsymbol{U}_{|\mathcal{S}|\times n}\boldsymbol{V}^{\mathrm{H}}\boldsymbol{P}_{\mathrm{null}(\boldsymbol{U}_{|\overline{\mathcal{S}}|\times n}\boldsymbol{V}^{\mathrm{H}})}$. Then, the denominator of the error bound (21) becomes

$$\frac{\|\boldsymbol{\Omega}_{\mathcal{S}}\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_2} =$$
$$\frac{\|\boldsymbol{D}_1\boldsymbol{U}_{|\mathcal{S}|\times n}\boldsymbol{V}^{\mathrm{H}}\boldsymbol{P}_{\mathrm{null}(\boldsymbol{U}_{|\overline{\mathcal{S}}|\times n}\boldsymbol{V}^{\mathrm{H}})}\boldsymbol{w}+\alpha\sigma_{\mathrm{min}}\boldsymbol{D}_1\boldsymbol{u}'_{\mathrm{min}}\|_1}{\|\boldsymbol{P}_{\mathrm{null}(\boldsymbol{U}_{|\overline{\mathcal{S}}|\times n}\boldsymbol{V}^{\mathrm{H}})}\boldsymbol{w}+\alpha\boldsymbol{P}_{\mathrm{null}(\boldsymbol{U}_{|\overline{\mathcal{S}}|\times n}\boldsymbol{V}^{\mathrm{H}})}\boldsymbol{v}'_{\mathrm{min}}\|_2}, \tag{46}$$

where $\boldsymbol{u}'_{\mathrm{min}}$ is the left singular vector corresponding to σ_{min}. As we have full control over α and the diagonal elements of $\boldsymbol{D}_1$, we can make the error bound (21) arbitrarily large (decreasing the diagonal elements of $\boldsymbol{D}_1$ while increasing α).

Table III compares the two error bounds for various examples of tall analysis operators. We observe that our error bound (31) is considerably superior to the error estimate (21). We use three kinds of analysis operators: Random 1, 2 and Daubechies wavelet for different sparsity levels and dimensions. Notice
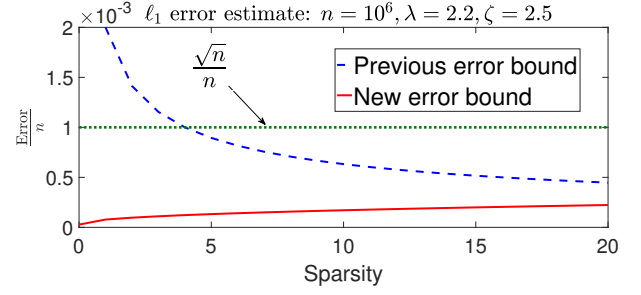

Figure 1: Two strategies of obtaining the error of $\delta(\mathcal{D}(f,\boldsymbol{x}))$ from (5) in case of $f = \|\cdot\|_1$. The previous and new error bounds come from (24) and (31), respectively.
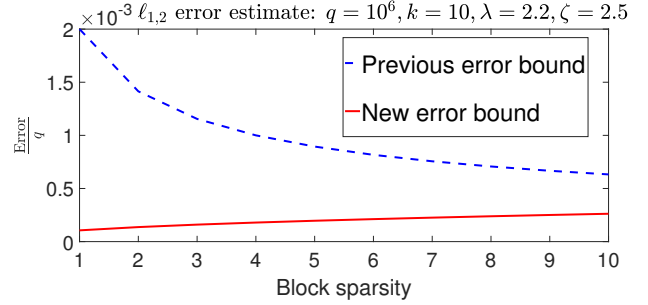

Figure 2: Comparison of the error (6) in case of $f = \|\cdot\|_{1,2}$. The previous and new error bounds come from (25) and (31), respectively.

that the Daubechies wavelet of size 2048×1024 is constructed by retaining the high-pass components of a 2-level Daubechies wavelet of size 3072×1024. The wavelet transformation is computed by the SPOT package [20]. The procedure of computing $\omega(\mathcal{D}(\|\boldsymbol{\Omega}\cdot\|_1,\boldsymbol{x}) \cap \mathbb{B}^n)$ in (31) is explained in Appendix E.

In Table IV, we examine fat analysis operators including Random 1 and 2 structures, TV and Daubechies wavelet. Again, our bound (31) confirms that U_δ is close to $\delta(\mathcal{D}(\|\boldsymbol{\Omega}\cdot\|_1,\boldsymbol{x}))$, while the error bound (21) is inconclusive.

Different from the above mentioned strategies for generating analysis-sparse signals, we also construct signals (shown by Random 1* in Table III and TV* in Table IV) as $\boldsymbol{x} = \boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})}\boldsymbol{c}$ where $\boldsymbol{c}$ is uniformly distributed on the unit sphere. In these cases, we observe that the error estimate (21) is effective.

VII. CONCLUSION

In this work, we presented an error estimate bound for the statistical dimension. This new bound shows that the statistical dimension is well described by its common upper-bound (5) in some settings of TV structure and ℓ_1 analysis.

APPENDIX A
PROOF OF THEOREM 2 (31)

Before beginning the proof, we define some parameters and provides a proof sketch to enhance the readability. Given $\lambda > 0$, we define the parameters

$$\alpha := \mathbb{E}[t_{\boldsymbol{g}}] + \frac{\lambda}{\|\boldsymbol{z}_0\|_2}, \tag{47}$$

$$t_{\boldsymbol{g}} := \arg\min_{t\geqslant 0}\mathrm{dist}(\boldsymbol{g}, t\partial f(\boldsymbol{x})), \tag{48}$$
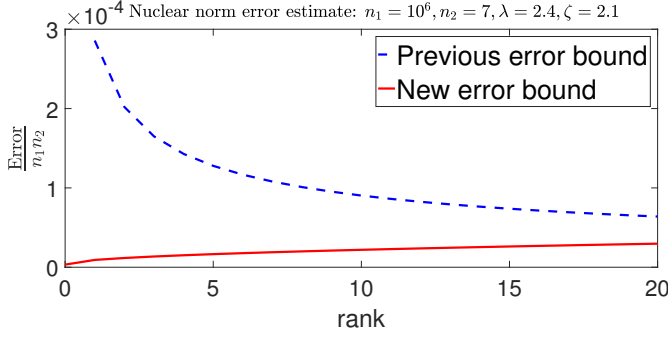
Figure 3: Comparison of the error (6) in case of $f = \|\cdot\|_*$. The previous and new error bounds come from (27) and (31), respectively.

and the function

$$\phi(\boldsymbol{g}) := \mathrm{dist}^2(\boldsymbol{g}, \alpha\partial f(\boldsymbol{x})) - \mathrm{dist}^2(\boldsymbol{g}, \mathrm{cone}(\partial f(\boldsymbol{x}))). \quad (49)$$

Notice that due to [1, Lemma C.1], whenever f is a proper convex function (as is the case in this paper), $t_{\boldsymbol{g}}$ is well-defined and unique. Define the event

$$\mathcal{E} = \left\{ |t_{\boldsymbol{g}} - \mathbb{E}[t_{\boldsymbol{g}}]| < \frac{\lambda}{\|\boldsymbol{z}_0\|_2} \right\}. \quad (50)$$

For fixed $\boldsymbol{g}$, by considering the condition (30), $t_{\boldsymbol{g}}$ is a $\frac{1}{\|\boldsymbol{z}_0\|_2}$ Lipschitz function of $\boldsymbol{g}$ [17, Lemma 3][5]. Hence, by a concentration inequality for Lipschitz functions of Gaussian vectors (see, for example, [21, Theorem 8.40]), we get that

$$\mathbb{P}\{\mathcal{E}\} \geqslant p_0 := 1 - 2e^{-\frac{\lambda^2}{2}}. \quad (51)$$

Proof skech . The goal is to find an upper-bound for E_δ. Instead of bounding E_δ, we bound the expression

$$\mathbb{E}\,\mathrm{dist}^2(\boldsymbol{g}, \alpha\partial f(\boldsymbol{x})) - \delta(\mathcal{D}(f, \boldsymbol{x})) = $$
$$\mathbb{E}\,\mathrm{dist}^2(\boldsymbol{g}, \alpha\partial f(\boldsymbol{x})) - \mathbb{E}\,\mathrm{dist}^2(\boldsymbol{g}, \mathrm{cone}(\partial f(\boldsymbol{x}))) = \mathbb{E}\phi(\boldsymbol{g}),$$

which is counted as its upper-bound. To reach this goal, we do the following steps:

1) When $\mathcal{E}$ holds, we find that

$$\phi_1(\boldsymbol{g}) := \phi(\boldsymbol{g}) - 4\lambda\beta\mathrm{dist}(\boldsymbol{g}, \mathrm{cone}(\partial f(\boldsymbol{x}))) \leqslant 4\lambda^2\beta^2.$$

2) We obtain a lower-bound for the probability of the event $\phi_1(\boldsymbol{g}) \leqslant 4\lambda^2\beta^2$.

3) Then, we obtain a concentration inequality for the expression $\phi_1(\boldsymbol{g})$ which is associated with $\mathbb{E}\phi_1(\boldsymbol{g})$ (see Lemma 1).

4) Combining the concentration inequality in Item 3 and the lower-bound in Item 2, we reach a contradiction unless we have that $\mathbb{E}\phi_1(\boldsymbol{g})$ is bounded above by a certain expression.

5) Finally, the upper-bound on $\mathbb{E}\phi(\boldsymbol{g})$ (and thus bound on E_δ) is directly obtained using the upper-bound on $\mathbb{E}\phi_1(\boldsymbol{g})$.

We now prove each of the above mentioned parts in details. Suppose that $\mathcal{E}$ holds. Define $\boldsymbol{z}^*$ such that

$$\mathrm{dist}^2(\boldsymbol{g}, t_{\boldsymbol{g}}\partial f(\boldsymbol{x})) = \|\boldsymbol{g} - t_{\boldsymbol{g}}\boldsymbol{z}^*\|_2^2. \quad (52)$$

[5]The proof of [17, Lemma 3] does not need $\boldsymbol{z}_0$ to be an element of $\partial f(\boldsymbol{x})$.

Take

$$\boldsymbol{z} = \frac{t_{\boldsymbol{g}}}{\alpha}\boldsymbol{z}^* + (1 - \frac{t_{\boldsymbol{g}}}{\alpha})\boldsymbol{z}_1 \in \partial f(\boldsymbol{x}), \quad (53)$$

where $\boldsymbol{z}_1$ is defined in (35). That this is an element of $\partial f(\boldsymbol{x})$ follows from the fact that both $\boldsymbol{z}_1$ and $\boldsymbol{z}^*$ are in $\partial f(\boldsymbol{x})$, and that $\mathcal{E}$ implies that $t_{\boldsymbol{g}}/\alpha < 1$. Then, we can find an upper-bound for $\phi(\boldsymbol{g})$ as follows:

$$\mathrm{dist}^2(\boldsymbol{g}, \alpha\partial f(\boldsymbol{x})) \leqslant \|\boldsymbol{g} - \alpha\boldsymbol{z}\|_2^2 = \|\boldsymbol{g} - t_{\boldsymbol{g}}\boldsymbol{z}^* + t_{\boldsymbol{g}}\boldsymbol{z}^* - \alpha\boldsymbol{z}\|_2^2$$
$$= \|\boldsymbol{g} - t_{\boldsymbol{g}}\boldsymbol{z}^*\|_2^2 + \|t_{\boldsymbol{g}}\boldsymbol{z}^* - \alpha\boldsymbol{z}\|_2^2 + 2\langle \boldsymbol{g} - t_{\boldsymbol{g}}\boldsymbol{z}^*, t_{\boldsymbol{g}}\boldsymbol{z}^* - \alpha\boldsymbol{z}\rangle$$
$$\leqslant \mathrm{dist}^2(\boldsymbol{g}, \mathrm{cone}(\partial f(\boldsymbol{x})))$$
$$\quad + (t_{\boldsymbol{g}} - \alpha)^2\|\boldsymbol{z}_1\|_2^2 + 2|t_{\boldsymbol{g}} - \alpha|\langle \boldsymbol{g} - t_{\boldsymbol{g}}\boldsymbol{z}^*, \boldsymbol{z}_1\rangle$$
$$\leqslant \mathrm{dist}^2(\boldsymbol{g}, \mathrm{cone}(\partial f(\boldsymbol{x})))$$
$$\quad + 4\lambda^2\beta^2 + 4\lambda\beta\mathrm{dist}(\boldsymbol{g}, \mathrm{cone}(\partial f(\boldsymbol{x}))), \quad (54)$$

where for the last inequality we use the fact that $\mathcal{E}$ holds, the definition of β (34), and the Cauchy-Schwartz inequality. In the following, we obtain a lower-bound for the probability of the event $\phi_1(\boldsymbol{g}) \leqslant 4\lambda^2\beta^2$.

$$\mathbb{P}\left\{\phi_1(\boldsymbol{g}) \leqslant 4\lambda^2\beta^2\right\} = \mathbb{P}\left\{\phi_1(\boldsymbol{g}) \leqslant 4\lambda^2\beta^2|\mathcal{E}\right\}\mathbb{P}\{\mathcal{E}\}+$$
$$\mathbb{P}\left\{\phi_1(\boldsymbol{g}) \leqslant 4\lambda^2\beta^2|\overline{\mathcal{E}}\right\}\mathbb{P}\{\overline{\mathcal{E}}\} \geqslant \mathbb{P}\{\mathcal{E}\} \geqslant p_0, \quad (55)$$

where we used (54), which implies $\mathbb{P}\left\{\phi_1(\boldsymbol{g}) \leqslant 4\lambda^2\beta^2|\mathcal{E}\right\} = 1$, and (51).

In what follows, we propose a lemma that provides a relation between $\phi_1(\boldsymbol{g})$ and $\mathbb{E}\phi_1(\boldsymbol{g})$.

Lemma 1. *Let $\boldsymbol{g} \in \mathbb{R}^n$ be a standard normal i.i.d. vector. Then, for given $\lambda, \zeta > 0$,*

$$\mathbb{P}\{\phi_1(\boldsymbol{g}) - \mathbb{E}\phi_1(\boldsymbol{g}) \leqslant -\gamma(\zeta + \mathbb{E}\,\mathrm{dist}(\boldsymbol{g}, \mathrm{cone}\partial f(\boldsymbol{x}))$$
$$+2\lambda\beta)\} \leqslant p_0, \quad (56)$$

where

$$\gamma := \sqrt{72}\sqrt{\ln\frac{3}{1 - 4e^{-\frac{\lambda^2}{2}} - 2e^{-\frac{\zeta^2}{2}}}},$$

as defined in (33).

The proof of this Lemma is postponed to Appendix C.

By considering (55) and (56), we reach a contradiction unless

$$\mathbb{E}[\phi_1(\boldsymbol{g})] \leqslant \gamma(\zeta + 2\lambda\beta + \mathbb{E}\mathrm{dist}(\boldsymbol{g}, \mathrm{cone}(\partial f(\boldsymbol{x})))) + 4\lambda^2\beta^2. \quad (57)$$

By expressing ϕ_1 in terms of ϕ and identifying the expected distance to the subdifferential cone as the Gaussian width, we reach the right-hand side of (31). The left-hand side is obtained by applying the Jensen's inequality on the infimum of an affine function (which is always concave).

APPENDIX B
PROOF OF THEOREM 2 (32)

Our approach in this part is to a great extent, similar to the proof of Result 2. However, the difference lies in the fact that [17, Proposition 1] needs the condition (22), while our proof needs the condition (30) which holds for more general

structure-inducing functions including ℓ_1 analysis. However, this bound is not an error estimate for the task of predicting the phase transition (see the explanations in Remark 2) and is used in our analysis in proving (31). We proceed with an overview of the proof. We first define

$$\phi_2(\boldsymbol{g}) := \text{dist}(\boldsymbol{g}, \alpha\partial f(\boldsymbol{x})) - \text{dist}(\boldsymbol{g}, \text{cone}\partial f(\boldsymbol{x})). \quad (58)$$

Proof skech . The goal is to find an upper-bound for the expression

$$\inf_{t\geqslant 0} \mathbb{E}\ \text{dist}(\boldsymbol{g}, t\partial f(\boldsymbol{x})) - \omega(\mathcal{D}(f,\boldsymbol{x}) \cap \mathbb{B}^n). \quad (59)$$

We instead intend to find an upper-bound for $\mathbb{E}\phi_2(\boldsymbol{g})$. To reach this goal, we follow the below steps:

1) Under the assumption that the event $\mathcal{E}$ holds, we find that $\phi_2(\boldsymbol{g}) \leqslant 2\lambda\beta$.
2) We obtain a lower-bound for the probability of the event $\phi_2(\boldsymbol{g}) \leqslant 2\lambda\beta$.
3) We obtain a concentration inequality for the expression $\phi_2(\boldsymbol{g})$ which is a 2-Lipschitz function of $\boldsymbol{g}$.
4) The concentration inequality in Item 3 and the lower-bound in Item 2 contradict each other unless we have that $\mathbb{E}\phi_2(\boldsymbol{g})$ is bounded above by a certain expression.

We now provide the details of the proof. Suppose that (50) holds. Define $\boldsymbol{z}^*$ such that

$$\text{dist}(\boldsymbol{g}, t_{\boldsymbol{g}}\partial f(\boldsymbol{x})) = \|\boldsymbol{g} - t_{\boldsymbol{g}}\boldsymbol{z}^*\|_2. \quad (60)$$

By recalling (53) and (47), we have that

$$\text{dist}(\boldsymbol{g}, \alpha\partial f(\boldsymbol{x})) \leqslant \|\boldsymbol{g} - \alpha\boldsymbol{z}\|_2 = \|\boldsymbol{g} - t_{\boldsymbol{g}}\boldsymbol{z}^* + t_{\boldsymbol{g}}\boldsymbol{z}^* - \alpha\boldsymbol{z}\|_2$$
$$\leqslant \|\boldsymbol{g} - t_{\boldsymbol{g}}\boldsymbol{z}^*\|_2 + \|t_{\boldsymbol{g}}\boldsymbol{z}^* - \alpha\boldsymbol{z}\|_2 \leqslant \|\boldsymbol{g} - t_{\boldsymbol{g}}\boldsymbol{z}^*\|_2 +$$
$$|t_{\boldsymbol{g}} - \alpha|\|\boldsymbol{z}_1\|_2 \leqslant \text{dist}(\boldsymbol{g}, \text{cone}\partial f(\boldsymbol{x})) + 2\lambda\beta. \quad (61)$$

Since $\mathbb{P}\{\mathcal{E}\} \geqslant p_0$, we have that

$$\mathbb{P}\{\phi_2(\boldsymbol{g}) \leqslant 2\lambda\beta\} \geqslant p_0, \quad (62)$$

by the argument in (55). Moreover, since $\phi_2(\boldsymbol{g})$ is a 2-Lipschitz function of $\boldsymbol{g}$, the concentration inequality for Lipschitz functions [21, Theorem 8.40], implies that

$$\mathbb{P}\left\{\phi_2(\boldsymbol{g}) - \mathbb{E}[\phi_2(\boldsymbol{g})] \leqslant -r\right\} \leqslant e^{-\frac{r^2}{8}}. \quad (63)$$

With a change of variables, we reach:

$$\mathbb{P}\left\{\phi_2(\boldsymbol{g}) - \mathbb{E}[\phi_2(\boldsymbol{g})] \leqslant -\sqrt{8\ln\frac{1}{p_0}}\right\} \leqslant p_0. \quad (64)$$

Note that (62) and (64) contradict each other unless

$$\mathbb{E}\{\phi_2(\boldsymbol{g})\} \leqslant \sqrt{8\ln\frac{1}{p_0}} + 2\lambda\beta. \quad (65)$$

Finally, by setting $\lambda = 2$ we reach (32).

APPENDIX C
PROOF OF LEMMA 1

Define the functions

$$\begin{aligned}
f_1(\boldsymbol{g}) &:= \text{dist}(\boldsymbol{g}, \alpha\partial f(\boldsymbol{x})), \\
f_2(\boldsymbol{g}) &:= \text{dist}(\boldsymbol{g}, \text{cone}(\partial f(\boldsymbol{x}))), \\
h_1(\boldsymbol{g}) &:= f_1^2(\boldsymbol{g}), \\
h_2(\boldsymbol{g}) &:= f_2^2(\boldsymbol{g}), \quad (66)
\end{aligned}$$

and the event

$$\mathcal{E}_1 := \left\{ f_2(\boldsymbol{g}) - \mathbb{E}f_2(\boldsymbol{g}) \leqslant \zeta \right\}. \quad (67)$$

Suppose that $\mathcal{E}$ and $\mathcal{E}_1$ hold. Then,

$$|h_1(\boldsymbol{g}) - h_1(\boldsymbol{g}')| = |f_1(\boldsymbol{g}) - f_1(\boldsymbol{g}')||f_1(\boldsymbol{g}) + f_1(\boldsymbol{g}')| \leqslant$$
$$2\|\boldsymbol{g} - \boldsymbol{g}'\|_2(\zeta + \mathbb{E}f_2(\boldsymbol{g}) + 2\lambda\beta), \quad (68)$$

where the second inequality comes from the fact that f_1 is 1-Lipschitz function of $\boldsymbol{g}$. The last inequality is the result of $f_1(\boldsymbol{g}) \leqslant f_2(\boldsymbol{g}) + 2\lambda\beta$ (61). Now suppose that only $\mathcal{E}_1$ holds, Then, with the same reasoning, we have:

$$|h_2(\boldsymbol{g}) - h_2(\boldsymbol{g}')| = |f_2(\boldsymbol{g}) - f_2(\boldsymbol{g}')||f_2(\boldsymbol{g}) + f_2(\boldsymbol{g}')| \leqslant$$
$$\|\boldsymbol{g} - \boldsymbol{g}'\|_2(|f_2(\boldsymbol{g})| + |f_2(\boldsymbol{g}')|) \leqslant 2\|\boldsymbol{g} - \boldsymbol{g}'\|_2(\zeta + \mathbb{E}f_2(\boldsymbol{g})), \quad (69)$$

$$\mathbb{P}\left\{h_1 - \mathbb{E}[h_1] \leqslant -\frac{r}{3}\Big|\mathcal{E}, \mathcal{E}_1\right\} \leqslant e^{-\frac{r^2}{72(\zeta + \mathbb{E}f_2(\boldsymbol{g}) + 2\lambda\beta)^2}},$$
$$\mathbb{P}\left\{h_2 - \mathbb{E}[h_2] \geqslant \frac{r}{3}\Big|\mathcal{E}_1\right\} \leqslant e^{-\frac{r^2}{72(\zeta + \mathbb{E}f_2(\boldsymbol{g}))^2}}. \quad (70)$$

Consequently,

$$\mathbb{P}\{\phi_1(\boldsymbol{g}) - \mathbb{E}[\phi_1(\boldsymbol{g})] \leqslant -r\} =$$
$$\mathbb{P}\left\{h_1 - \mathbb{E}[h_1] - h_2 + \mathbb{E}[h_2] - 4\lambda\beta f_2 + 4\lambda\beta\mathbb{E}[f_2] \leqslant -r\right\}$$
$$\leqslant \mathbb{P}\{h_1 - \mathbb{E}[h_1] \leqslant -\frac{r}{3}\} + \mathbb{P}\{h_2 - \mathbb{E}[h_2] \geqslant \frac{r}{3}\}+$$
$$\mathbb{P}\{f_2 - \mathbb{E}[f_2] \geqslant \frac{r}{12\lambda\beta}\} \leqslant \mathbb{P}\left\{h_1 - \mathbb{E}[h_1] \leqslant -\frac{r}{3}\Big|\mathcal{E}_1\right\}\mathbb{P}\{\mathcal{E}_1\}$$
$$+ \mathbb{P}\left\{h_1 - \mathbb{E}[h_1] \leqslant -\frac{r}{3}\Big|\overline{\mathcal{E}}_1\right\}\mathbb{P}\{\overline{\mathcal{E}}_1\}+$$
$$\mathbb{P}\left\{h_2 - \mathbb{E}[h_2] \geqslant \frac{r}{3}\Big|\mathcal{E}_1\right\}\mathbb{P}\{\mathcal{E}_1\}+$$
$$\mathbb{P}\left\{h_2 - \mathbb{E}[h_2] \geqslant \frac{r}{3}\Big|\overline{\mathcal{E}}_1\right\}\mathbb{P}\{\overline{\mathcal{E}}_1\} + \mathbb{P}\{f_2 - \mathbb{E}[f_2] \geqslant \frac{r}{12\lambda\beta}\} \leqslant$$
$$e^{-\frac{r^2}{72(\zeta + \mathbb{E}f_2(\boldsymbol{g}) + 2\lambda\beta)^2}} + 2e^{-\frac{\lambda^2}{2}} + e^{-\frac{\zeta^2}{2}} + e^{-\frac{r^2}{72(\zeta + \mathbb{E}f_2(\boldsymbol{g}))^2}} + e^{-\frac{\zeta^2}{2}}$$
$$+ e^{-\frac{r^2}{72 \times 4\lambda^2\beta^2}} \leqslant 3e^{-\frac{r^2}{72(\zeta + \mathbb{E}f_2(\boldsymbol{g}) + 2\lambda\beta)^2}} + 2e^{-\frac{\lambda^2}{2}} + 2e^{-\frac{\zeta^2}{2}}, \quad (71)$$

where in the third inequality, we used

$$\mathbb{P}\left\{h_1 - \mathbb{E}[h_1] \leqslant -\frac{r}{3}\Big|\mathcal{E}_1\right\} = \mathbb{P}\left\{h_1 - \mathbb{E}[h_1] \leqslant -\frac{r}{3}\Big|\mathcal{E}_1, \mathcal{E}\right\}\mathbb{P}\{\mathcal{E}\}$$
$$+ \mathbb{P}\left\{h_1 - \mathbb{E}[h_1] \leqslant -\frac{r}{3}\Big|\mathcal{E}_1, \overline{\mathcal{E}}\right\}\mathbb{P}\{\overline{\mathcal{E}}\}, \quad (72)$$

and (70). With a change of variable, we reach (56).

APPENDIX D
PROOF OF PROPOSITION 2

The condition (30) for function $f = \|\mathbf{\Omega} \cdot \|_1$ can be stated as:

$$\exists \boldsymbol{z}_0 : \quad \langle \mathbf{\Omega}^T \boldsymbol{w} - \boldsymbol{z}_0, \boldsymbol{z}_0 \rangle = 0 \ : \forall \boldsymbol{w} \in \partial \| \cdot \|_1(\mathbf{\Omega} \boldsymbol{x}). \quad (73)$$

By setting $\boldsymbol{w} = \text{sgn}(\mathbf{\Omega} \boldsymbol{x}) + \boldsymbol{v}_{\overline{\mathcal{S}}}$ where $\|\boldsymbol{v}_{\overline{\mathcal{S}}}\|_\infty \leqslant 1$, we have:

$$\langle \mathbf{\Omega}^T \text{sgn}(\mathbf{\Omega} \boldsymbol{x}) + \mathbf{\Omega}_{\overline{\mathcal{S}}}^T \widetilde{\boldsymbol{v}}_{\overline{\mathcal{S}}}, \boldsymbol{z}_0 \rangle = \|\boldsymbol{z}_0\|_2^2,$$
$$\forall \widetilde{\boldsymbol{v}}_{\overline{\mathcal{S}}} \text{ with } \|\widetilde{\boldsymbol{v}}_{\overline{\mathcal{S}}}\|_\infty \leqslant 1. \quad (74)$$

Since $\boldsymbol{z}_0$ and $\text{sgn}(\mathbf{\Omega} \boldsymbol{x})$ are fixed and both $\widetilde{\boldsymbol{v}}_{\overline{\mathcal{S}}}$ and $-\widetilde{\boldsymbol{v}}_{\overline{\mathcal{S}}}$ satisfy $\|\widetilde{\boldsymbol{v}}_{\overline{\mathcal{S}}}\|_\infty \leqslant 1$, it holds that:

$$\langle \mathbf{\Omega}_{\overline{\mathcal{S}}}^T \widetilde{\boldsymbol{v}}_{\overline{\mathcal{S}}}, \boldsymbol{z}_0 \rangle = 0. \quad (75)$$

The expressions (74) and (75) lead to:

$$\boldsymbol{z}_0 \in \text{null}(\mathbf{\Omega}_{\overline{\mathcal{S}}}), \quad (76)$$
$$\langle \mathbf{\Omega}^T \text{sgn}(\mathbf{\Omega} \boldsymbol{x}), \boldsymbol{z}_0 \rangle = \|\boldsymbol{z}_0\|_2^2. \quad (77)$$

To satisfy (76) and (77) simultaneously, we choose the vector

$$\boldsymbol{z}_0 = \frac{A_1}{A_2} \boldsymbol{P}_{\text{null}(\mathbf{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{c}_0, \quad (78)$$

where

$$A_1 = \langle \mathbf{\Omega}^T \text{sgn}(\mathbf{\Omega} \boldsymbol{x}), \boldsymbol{P}_{\text{null}(\mathbf{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{c}_0 \rangle,$$
$$A_2 = \|\boldsymbol{P}_{\text{null}(\mathbf{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{c}_0\|_2^2,$$

and $\boldsymbol{c}_0$ is an arbitrary vector. By choosing $\boldsymbol{c}_0 = \mathbf{\Omega}^T \text{sgn}(\mathbf{\Omega} \boldsymbol{x})$, we have:

$$\boldsymbol{z}_0 = \boldsymbol{P}_{\text{null}(\mathbf{\Omega}_{\overline{\mathcal{S}}})} \mathbf{\Omega}^T \text{sgn}(\mathbf{\Omega} \boldsymbol{x}) \quad (79)$$

APPENDIX E
NUMERICAL COMPUTATION OF THE GAUSSIAN WIDTH

The Gaussian width $\omega(\mathcal{D}(f, \boldsymbol{x}) \cap \mathbb{B}^n)$ plays a key role in our proposed error estimate in Theorem 2. In this appendix, we explain how to compute this quantity numerically. This approach is adapted from [11, Section B.2]. Recall that this quantity is defined as

$$\omega(\mathcal{D}(f, \boldsymbol{x}) \cap \mathbb{B}^n) := \mathbb{E} \sup_{\substack{\boldsymbol{y} \in \mathcal{D}(f, \boldsymbol{x}) \\ \|\boldsymbol{y}\|_2 \leqslant 1}} \langle \boldsymbol{y}, \boldsymbol{g} \rangle. \quad (80)$$

By choosing a sufficiently small t, e.g. $t = 0.01$ in (11) (the rationale of this choice is discussed in [11, Section B.2]), the expression inside $\mathbb{E}$ can be simplified as the simple convex program:

$$\omega(\mathcal{D}(f, \boldsymbol{x}) \cap \mathbb{B}^n) \approx \mathbb{E} \sup_{\substack{f(\boldsymbol{x}+t\boldsymbol{y}) \leqslant f(\boldsymbol{x}) \\ \|\boldsymbol{y}\|_2 \leqslant 1}} \langle \boldsymbol{y}, \boldsymbol{g} \rangle, \quad (81)$$

that can be solved using the CVX package [22].

APPENDIX F
THE WEAK DECOMPOSABILITY CONDITION FOR ℓ_1-ANALYSIS

With a counterexample, we show that the mentioned decomposability condition does not hold in general. Let $\mathbf{\Omega}_{p \times n}$ be a tall analysis operator in general position and $\boldsymbol{x}_{n \times 1}$ be an analysis-sparse vector such that $\mathcal{S} = \text{supp}(\mathbf{\Omega} \boldsymbol{x})$ with $|\mathcal{S}| \geqslant p - n$ (see [14, Section 2.1]). For the weak decomposability condition of [17] to hold for $(\mathbf{\Omega}, \mathcal{S}, \boldsymbol{x})$, we shall have that

$$\exists \boldsymbol{w}_0 \in \partial \| \cdot \|_1(\mathbf{\Omega} \boldsymbol{x}),$$
$$\forall \boldsymbol{w} \in \partial \| \cdot \|_1(\mathbf{\Omega} \boldsymbol{x}) : \quad \langle \mathbf{\Omega}^T(\boldsymbol{w} - \boldsymbol{w}_0), \mathbf{\Omega}^T \boldsymbol{w}_0 \rangle = 0. \quad (82)$$

In particular, we can set $\boldsymbol{w} = \text{sgn}(\mathbf{\Omega} \boldsymbol{x}) + \boldsymbol{v}_{\overline{\mathcal{S}}}$, where $\boldsymbol{v}_{p \times 1}$ is an arbitrary vector with $\|\boldsymbol{v}\|_\infty \leqslant 1$. Since $(\mathbf{\Omega} \boldsymbol{x})_{\overline{\mathcal{S}}} = \boldsymbol{0}_{p \times 1}$, we can write that

$$\forall \boldsymbol{v}, \|\boldsymbol{v}\|_\infty \leqslant 1 :$$
$$\langle \mathbf{\Omega}^T \text{sgn}(\mathbf{\Omega} \boldsymbol{x}) + \mathbf{\Omega}_{\overline{\mathcal{S}}}^T \widetilde{\boldsymbol{v}}_{\overline{\mathcal{S}}}, \mathbf{\Omega}^T \boldsymbol{w}_0 \rangle = \|\mathbf{\Omega}^T \boldsymbol{w}_0\|_2^2. \quad (83)$$

By setting $\boldsymbol{v} = \boldsymbol{0}$ and applying the result for general $\boldsymbol{v}$, we obtain

$$\langle \mathbf{\Omega}_{\overline{\mathcal{S}}}^T \widetilde{\boldsymbol{v}}_{\overline{\mathcal{S}}}, \mathbf{\Omega}^T \boldsymbol{w}_0 \rangle = 0, \quad (84)$$

for arbitrary $\boldsymbol{v}$ with $\|\boldsymbol{v}\|_\infty \leqslant 1$, or equivalently, for arbitrary $\boldsymbol{v}$. This implies that

$$\langle \widetilde{\boldsymbol{v}}_{\overline{\mathcal{S}}}, \mathbf{\Omega}_{\overline{\mathcal{S}}} \mathbf{\Omega}^T \boldsymbol{w}_0 \rangle = 0, \quad (85)$$

or equivalently

$$\mathbf{\Omega}_{\overline{\mathcal{S}}} \mathbf{\Omega}^T \boldsymbol{w}_0 = 0. \quad (86)$$

As $\boldsymbol{w}_0 \in \partial \| \cdot \|_1(\mathbf{\Omega} \boldsymbol{x})$, we know that $\boldsymbol{w}_0 = \text{sgn}(\mathbf{\Omega} \boldsymbol{x}) + \boldsymbol{v}_{0\overline{\mathcal{S}}}$ for some $\|\boldsymbol{v}_0\|_\infty \leqslant 1$. Therefore,

$$\mathbf{\Omega}_{\overline{\mathcal{S}}} \mathbf{\Omega}_{\overline{\mathcal{S}}}^T \widetilde{\boldsymbol{v}}_{0\overline{\mathcal{S}}} = -\mathbf{\Omega}_{\overline{\mathcal{S}}} \mathbf{\Omega}^T \text{sgn}(\mathbf{\Omega} \boldsymbol{x}) = -\mathbf{\Omega}_{\overline{\mathcal{S}}} \mathbf{\Omega}_{\mathcal{S}}^T \text{sgn}(\mathbf{\Omega}_{\mathcal{S}} \boldsymbol{x}). \quad (87)$$

Because $\mathbf{\Omega}$ is in general position, $\mathbf{\Omega}_{\overline{\mathcal{S}}} \mathbf{\Omega}_{\overline{\mathcal{S}}}^T$ is invertible and we can express $\widetilde{\boldsymbol{v}}_{0\overline{\mathcal{S}}}$ as

$$\widetilde{\boldsymbol{v}}_{0\overline{\mathcal{S}}} = -(\mathbf{\Omega}_{\overline{\mathcal{S}}} \mathbf{\Omega}_{\overline{\mathcal{S}}}^T)^{-1} \mathbf{\Omega}_{\overline{\mathcal{S}}} \mathbf{\Omega}_{\mathcal{S}}^T \text{sgn}(\mathbf{\Omega}_{\mathcal{S}} \boldsymbol{x}). \quad (88)$$

Now, the contradiction comes from the fact that the entries $\boldsymbol{v}_0$ in the above equation are not necessarily confined to the interval $[-1, 1]$. We show this by a numerical example:

$$\boldsymbol{x} = \begin{bmatrix} -.4472 \\ .8944 \end{bmatrix}, \mathbf{\Omega} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 1 & 2 \end{bmatrix}, \mathcal{S} = \{1, 3\}, \overline{\mathcal{S}} = \{2\}. \quad (89)$$

For the latter signal, it holds that

$$\widetilde{v}_{0\overline{\mathcal{S}}} = -1.4, \quad (90)$$

which obviously contradicts the constraint $\|v_{0\overline{\mathcal{S}}}\|_\infty \leqslant 1$.

APPENDIX G
ASYMPTOTIC BEHAVIOR OF β WHEN $\boldsymbol{\Omega}$ IS GAUSSIAN ENSEMBLE

We find an upper-bound on β when the analysis operator is a Gaussian ensemble (whether fat or tall). Recall that an upper-bound for β is obtained in (38) as follows:

$$\beta \leqslant \frac{\|\boldsymbol{\Omega}_{\mathcal{S}}^T \mathrm{sgn}(\boldsymbol{\Omega}_{\mathcal{S}} \boldsymbol{x})\|_2}{\|\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{\Omega}_{\mathcal{S}}^T \mathrm{sgn}(\boldsymbol{\Omega}_{\mathcal{S}} \boldsymbol{x})\|_2}. \tag{91}$$

Since $\boldsymbol{\Omega}_{\mathcal{S}}$ is statistically independent of $\boldsymbol{\Omega}_{\overline{\mathcal{S}}}$, $\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})}$ defines a projection onto an $n - |\mathcal{S}|$ subspace that is independently and uniformly oriented with respect to $\boldsymbol{v} = \frac{\boldsymbol{\Omega}_{\mathcal{S}}^T \mathrm{sgn}(\boldsymbol{\Omega}_{\mathcal{S}} \boldsymbol{x})}{\|\boldsymbol{\Omega}_{\mathcal{S}}^T \mathrm{sgn}(\boldsymbol{\Omega}_{\mathcal{S}} \boldsymbol{x})\|_2}$. In addition

$$\beta \leqslant \frac{1}{\|\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{v}\|_2}. \tag{92}$$

As the relative orientation of $\boldsymbol{v}$ with respect to $\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})$ determines the upper-bound for β, we can fix $\boldsymbol{v}$ on the unit sphere and randomly rotate $\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})$ with a uniform Haar measure. Equivalently, we can fix $\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})$ and randomly select $\boldsymbol{v}$ with a uniform distribution on the unit sphere. In fact, the choice $\boldsymbol{v} = \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$ where $\boldsymbol{g}$ is an i.i.d. random vector with standard normal distribution independent of $\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})$ fulfills this requirement. With this choice, we can rewrite the upper-bound on β as

$$\beta \leqslant \frac{\|\boldsymbol{g}\|_2}{\|\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{g}\|_2}. \tag{93}$$

It is straightforward to check that $\|\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})} \cdot \|_2$ and $\| \cdot \|_2$ are 1-Lipschitz functions. Therefore, in high dimensions, we know that $\|\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{g}\|_2$ and $\|\boldsymbol{g}\|_2$ are concentrated around $\mathbb{E}\|\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{g}\|_2$ and $\mathbb{E}\|\boldsymbol{g}\|_2$, respectively. Also, by [23, Corollary 3.2] and [21, Proposition 8.1], it holds that

$$\mathbb{E}\|\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{g}\|_2 \geqslant \sqrt{\mathbb{E}\|\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{g}\|_2^2 - 1},$$
$$\mathbb{E}\|\boldsymbol{g}\|_2 \leqslant \sqrt{n}. \tag{94}$$

Let $[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n-|\mathcal{S}|}]$ be a basis for $\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})$. Then, we have

$$\mathbb{E}\|\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{g}\|_2^2 = \sum_{i=1}^{n-|\mathcal{S}|} \mathbb{E}|\langle \boldsymbol{u}_i, \boldsymbol{g}\rangle|^2 = \sum_{i=1}^{n-|\mathcal{S}|} \boldsymbol{u}_i^T \mathbb{E}\boldsymbol{g}\boldsymbol{g}^T \boldsymbol{u}_i. \tag{95}$$

Because $\mathbb{E}\boldsymbol{g}\boldsymbol{g}^{\mathrm{H}} = \boldsymbol{I}$, we can now simplify (95) as

$$\mathbb{E}\|\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{g}\|_2^2 = \sum_{i=1}^{n-|\mathcal{S}|} \underbrace{\boldsymbol{u}_i^T \boldsymbol{u}_i}_{1} = n - |\mathcal{S}|. \tag{96}$$

As a result, due to (94) and (96), we reach

$$\beta \leqslant \frac{\sqrt{n}}{\sqrt{n - |\mathcal{S}| - 1}}.$$

APPENDIX H
ANALYSIS OPERATORS WITH ORTHOGONAL ROWS

In this section, we consider a fat analysis operator $\boldsymbol{\Omega} \in \mathbb{R}^{p \times n}$ that is constructed via the recipe in Item 1 of Section VI. The result, however, holds for all analysis operators for which $\boldsymbol{\Omega}\boldsymbol{\Omega}^T$ is diagonal.

When $\boldsymbol{\Omega}$ in (41) is fat with full row-rank, we have $r = p$ and $\boldsymbol{\Omega}$ can be expressed as

$$\boldsymbol{\Omega} = \boldsymbol{D}_{p \times p} \boldsymbol{U}_{p \times p} \boldsymbol{V}_{n \times p}^{\mathrm{T}}. \tag{97}$$

Using MATLAB matrix notations, we have:

$$\boldsymbol{\Omega}_{\overline{\mathcal{S}}} = \boldsymbol{D}(\overline{\mathcal{S}}, \overline{\mathcal{S}}) \boldsymbol{U}(\overline{\mathcal{S}}, [p]) \boldsymbol{V}([n], [p])^{\mathrm{T}},$$
$$\boldsymbol{\Omega}_{\mathcal{S}} = \boldsymbol{D}(\mathcal{S}, \mathcal{S}) \boldsymbol{U}(\mathcal{S}, [p]) \boldsymbol{V}([n], [p])^{\mathrm{T}}. \tag{98}$$

As a consequence, it holds that

$$\boldsymbol{\Omega}_{\overline{\mathcal{S}}} \boldsymbol{\Omega}_{\mathcal{S}}^{\mathrm{T}} = \boldsymbol{D}(\overline{\mathcal{S}}, \overline{\mathcal{S}}) \boldsymbol{U}(\overline{\mathcal{S}}, [p]) \boldsymbol{U}^{\mathrm{T}}(\mathcal{S}, [p]) \boldsymbol{D}(\mathcal{S}, \mathcal{S}). \tag{99}$$

Since $\boldsymbol{U}(\overline{\mathcal{S}}, [p]) \boldsymbol{U}^{\mathrm{T}}(\mathcal{S}, [p])$ is a submatrix of $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}_p$, we have that

$$\boldsymbol{\Omega}_{\overline{\mathcal{S}}} \boldsymbol{\Omega}_{\mathcal{S}}^{\mathrm{T}} = \boldsymbol{0}. \tag{100}$$

Thus, the denominator of β becomes

$$\|\boldsymbol{P}_{\mathrm{null}(\boldsymbol{\Omega}_{\overline{\mathcal{S}}})} \boldsymbol{\Omega}^T \mathrm{sgn}(\boldsymbol{\Omega}\boldsymbol{x})\|_2 = \|(\boldsymbol{I}_n - \boldsymbol{\Omega}_{\overline{\mathcal{S}}}^{\dagger} \boldsymbol{\Omega}_{\overline{\mathcal{S}}}) \boldsymbol{\Omega}_{\mathcal{S}}^T \mathrm{sgn}(\boldsymbol{\Omega}_{\mathcal{S}}\boldsymbol{x})\|_2$$
$$= \|\boldsymbol{\Omega}_{\mathcal{S}}^T \mathrm{sgn}(\boldsymbol{\Omega}_{\mathcal{S}}\boldsymbol{x})\|_2. \tag{101}$$

By checking the gradient of the cost in the optimization in the definition of β (numerator), we can check that

$$\boldsymbol{z}_{\overline{\mathcal{S}}} = -(\boldsymbol{\Omega}_{\overline{\mathcal{S}}} \boldsymbol{\Omega}_{\overline{\mathcal{S}}}^T)^{-1} \boldsymbol{\Omega}_{\overline{\mathcal{S}}} \boldsymbol{\Omega}_{\mathcal{S}}^T \mathrm{sgn}(\boldsymbol{\Omega}_{\mathcal{S}}\boldsymbol{x}) = \boldsymbol{0} \tag{102}$$

is the unique minimizer (the gradient of the cost is zero at $\boldsymbol{z}_{\overline{\mathcal{S}}}$, and $\boldsymbol{z}_{\overline{\mathcal{S}}}$ satisfies the constraints). Also, the minimum value of the cost (value of the numerator) becomes $\|\boldsymbol{\Omega}^T \mathrm{sgn}(\boldsymbol{\Omega}\boldsymbol{x})\|_2$ with this choice of $\boldsymbol{z}_{\overline{\mathcal{S}}}$. This reveals that the numerator of β equals its denominator, i.e., $\beta = 1$.

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: A Journal of the IMA*, vol. 3, no. 3, pp. 224–294, 2014.

[2] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

[3] D. Donoho and J. Tanner, "Counting faces of randomly projected polytopes when the projection radically lowers dimension," *Journal of the American Mathematical Society*, vol. 22, no. 1, pp. 1–53, 2009.

[4] M. Rudelson and R. Vershynin, "On sparse reconstruction from fourier and gaussian measurements," *Communications on Pure and Applied Mathematics*, vol. 61, no. 8, pp. 1025–1045, 2008.

[5] M. Bayati, M. Lelarge, and A. Montanari, "Universality in polytope phase transitions and message passing algorithms," *The Annals of Applied Probability*, vol. 25, no. 2, pp. 753–822, 2015.

[6] M. Stojnic, "Various thresholds for ℓ_1-optimization in compressed sensing," *arXiv preprint arXiv:0907.3666*, 2009.

[7] D. L. Donoho, I. Johnstone, and A. Montanari, "Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising," *IEEE transactions on information theory*, vol. 59, no. 6, pp. 3396–3433, 2013.

[8] S. Oymak and B. Hassibi, "Sharp mse bounds for proximal denoising," *Foundations of Computational Mathematics*, vol. 16, no. 4, pp. 965–1029, 2016.

[9] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," *Inverse problems*, vol. 23, no. 3, p. 947, 2007.

[10] Y. Gordon, "On milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$," in *Geometric Aspects of Functional Analysis*, pp. 84–106, Springer, 1988.

[11] M. Genzel, G. Kutyniok, and M. März, "ℓ_1-analysis minimization and generalized (co-) sparsity: When does recovery succeed?," *arXiv preprint arXiv:1710.04952*, 2017.

[12] P. G. Casazza, G. Kutyniok, and F. Philipp, "Introduction to finite frame theory," in *Finite Frames*, pp. 1–53, Springer, 2013.

[13] M. Kabanava and H. Rauhut, "Analysis ℓ_1-recovery with frames and Gaussian measurements," *Acta Applicandae Mathematicae*, vol. 140, no. 1, pp. 173–195, 2015.

[14] S. Nam, M. E. Davies, M. Elad, and R. Gribonval, "The cosparse analysis model and algorithms," *Applied and Computational Harmonic Analysis*, vol. 34, no. 1, pp. 30–56, 2013.

[15] B. Zhang, W. Xu, J.-F. Cai, and L. Lai, "Precise phase transition of total variation minimization," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 4518–4522, IEEE, 2016.

[16] R. T. Rockafellar, *Convex analysis*. Princeton university press, 2015.

[17] R. Foygel and L. Mackey, "Corrupted sensing: Novel guarantees for separating structured signals," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 1223–1247, 2014.

[18] E. Candes and B. Recht, "Simple bounds for recovering low-complexity models," *Mathematical Programming*, vol. 141, no. 1-2, pp. 577–589, 2013.

[19] S. Daei, F. Haddadi, and A. Amini, "Exploiting prior information in block sparse signals," *arXiv preprint arXiv:1804.08444*, 2018.

[20] E. Van den Berg and M. Friedlander, "Spot-a linear-operator toolbox," *URL http://www. cs. ubc. ca/labs/scl/spot*, 2014.

[21] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, vol. 1. Birkhäuser Basel, 2013.

[22] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," Mar. 2014.

[23] M. Ledoux, *The concentration of measure phenomenon*. No. 89, American Mathematical Soc., 2001.