Compressed Sensing with Prior Information: Strategies, Geometry, and Bounds

João F. C. Mota, *Member, IEEE*, Nikos Deligiannis, *Member, IEEE*, Miguel R. D. Rodrigues, *Senior Member, IEEE*

***Abstract*—We address the problem of compressed sensing (CS) with prior information: *reconstruct a target CS signal with the aid of a similar signal that is known beforehand, our prior information*. We integrate the additional knowledge of the similar signal into CS via ℓ_1-ℓ_1 and ℓ_1-ℓ_2 minimization. We then establish bounds on the number of measurements required by these problems to successfully reconstruct the original signal. Our bounds and geometrical interpretations reveal that if the prior information has good enough quality, ℓ_1-ℓ_1 minimization improves the performance of CS dramatically. In contrast, ℓ_1-ℓ_2 minimization has a performance very similar to classical CS and brings no significant benefits. In addition, we use the insight provided by our bounds to design practical schemes to improve prior information. All our findings are illustrated with experimental results.**

Index Terms — Compressed sensing, prior information, basis pursuit, ℓ_1-ℓ_1 and ℓ_1-ℓ_2 minimization, Gaussian width.

I. Introduction

Nearly a decade ago, compressed sensing (CS) emerged as a new paradigm for signal acquisition [2], [3]. By assuming that signals are compressible rather than bandlimited, CS enables signal acquisition using far less measurements than classical acquisition schemes [4], [5]. Since most signals of interest are indeed compressible, CS has found many applications, including medical imaging [6], radar [7], camera design [8], and sensor networks [9].

We show that whenever a signal similar to the signal to reconstruct is available, the number of measurements can be reduced even further. Such additional knowledge is often called *prior* [10]–[20] or *side* [21]–[23] *information*.

Compressed Sensing (CS). Let $x^\star \in \mathbb{R}^n$ be an unknown s-sparse signal, i.e., with at most s nonzero entries. Assume we have m linear measurements $y = Ax^\star$, where the matrix $A \in \mathbb{R}^{m\times n}$ is known. CS answers two fundamental questions: *How to reconstruct the signal $x^\star$ from the measurements y? And how many measurements m are required for successful reconstruction?* A remarkable result states that if A satisfies a restricted isometry [24]–[26] or nullspace [27] property, then $x^\star$ can be reconstructed perfectly by solving *Basis Pursuit* (BP) [28]:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \|x\|_1 \\ \text{subject to} \quad & Ax = y\,, \end{aligned} \tag{BP}$$

where $\|x\|_1 := \sum_{i=1}^n |x_i|$ is the ℓ_1-norm of x; see [24]–[27]. For example, if $m > 2s\log(n/s) + (7/5)s$, and the entries of $A \in \mathbb{R}^{m\times n}$ are drawn independently and identically distributed (i.i.d.) from the Gaussian distribution, then A satisfies a nullspace property (and thus BP recovers $x^\star$) with high probability [27]. See [2], [3], [29]–[36] for related results.

CS with prior information. Consider that, in addition to the set of measurements $y = Ax^\star$, we also have access to *prior information*, that is, to a signal $w \in \mathbb{R}^n$ similar to the original signal $x^\star$. This occurs in many scenarios: for example, in video acquisition [21], [37]–[41], tracking [42], [43], and medical imaging [6], [11], [20], [44], past signals can be used to create an estimate of the target signal; concretely, if $x^\star$ is a sparse representation of the target signal, then w can be a sparse representation of an estimate of $x^\star$, created from past reconstructed signals, e.g., via extrapolation. Similarly, signals captured by nearby sensors in sensor networks [45] and images in multiview camera systems [46] are (or can be made) similar and, hence, used as prior information. The goal of this paper is to answer the following two key questions:

- *How to reconstruct the signal $x^\star$ from the measurements $y = Ax^\star$ and the prior information w?*
- *And how many measurements m are required for successful reconstruction?*

A. Overview of Our Approach and Main Results

We address CS with prior information by solving an appropriate modification of BP. Suppose $g : \mathbb{R}^n \to \mathbb{R}$ is a function that measures the similarity between $x^\star$ and the prior information w, in the sense that $g(x^\star - w)$ is expected to be small. Then, given $y = Ax^\star$ and w, we solve

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \|x\|_1 + \beta\, g(x - w) \\ \text{subject to} \quad & Ax = y\,, \end{aligned} \tag{1}$$

where $\beta > 0$ establishes a tradeoff between signal sparsity and fidelity to prior information. We consider two specific, convex

This work was supported by the EPSRC grant EP/K033166/1, the VUB Research Programme M3D2, the FWO grant G0A2617N, the VUB-UGent-UCL-Duke International Joint Research Group, and by Heriot-Watt University. Part of this work was presented at the GlobalSIP 2014 conference [1].

J. F. C. Mota is with the Institute of Sensors, Signals, and Systems at Heriot-Watt University, Edinburgh, EH14 4AS, U.K. He was with the Department of Electronic and Electrical Engineering, University College London, London, WC1E 6BT, U.K. (e-mail: j.mota@hw.ac.uk).

N. Deligiannis is with the Department of Electronics and Informatics, Vrije Universiteit Brussel, Brussels 1050, Belgium. He was with the Department of Electronic and Electrical Engineering, University College London, London, WC1E 6BT, U.K. (e-mail: ndeligia@etrovub.be).

M. R. D. Rodrigues is with the Electronic & Electrical Engineering Department at University College London, London, WC1E 6BT, U.K. (e-mail: m.rodrigues@ucl.ac.uk).

models for g: $g_1 := \|\cdot\|_1$ and $g_2 := \frac{1}{2}\|\cdot\|_2^2$, where $\|z\|_2 := \sqrt{z^\top z}$ is the ℓ_2-norm. Then, problem (1) becomes

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \|x\|_1 + \beta\|x - w\|_1 \\ \text{subject to} & Ax = y \end{array} \tag{2}$$

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \|x\|_1 + \frac{\beta}{2}\|x - w\|_2^2 \\ \text{subject to} & Ax = y\,, \end{array} \tag{3}$$

which we will refer to as ℓ_1-ℓ_1 and ℓ_1-ℓ_2 minimization, respectively. The use of the constraints $Ax = y$ implicitly assumes that y was acquired without noise. However, our results also apply to the noisy scenario, i.e., when the constraints are $\|Ax - y\|_2 \leq \sigma$ instead of $Ax = y$.

Overview of results. Problems (2) and (3), as well as their Lagrangian versions, have rarely appeared in the literature (see Section II). For instance, [11], [20] (resp. [12]) considered problems very similar to (2) (resp. (3)). Yet, to the best of our knowledge, no CS-type results have ever been provided for either (2), (3), their variations in [11], [12], [20], or their Lagrangian versions.

Our goal is to establish bounds on the number of measurements that guarantee that (2) and (3) reconstruct $x^\star$ with high probability, when A has i.i.d. Gaussian entries. Our bounds are a function of the prior information "quality" and the tradeoff parameter β. Hence, they not only help us understand what "good" prior information is, but also to select a β that minimizes the number of measurements. The main elements of our contribution can be summarized as follows:

- Our bound for (2) is minimized when $\beta = 1$, a value independent of w, $x^\star$, or any other problem parameter. We will see that the best β in practice is indeed very close to 1. In contrast, the optimal β for (3) depends on several parameters, including the unknown entries of $x^\star$.
- We also establish sharper versions of our bounds, which have to be computed numerically, but precisely describe the experimental performance of (2) and (3). Our analyses of the bounds, sharp and non-sharp, reveal that, typically, (2) requires much fewer measurements than both BP (classical CS) and (3). This superior performance is also observed experimentally, and we interpret it in terms of the underlying geometry of the problem.
- Based on the measures for the quality of prior information revealed by our bounds, we propose schemes that modify prior information in order to improve its quality. The schemes are validated with simulations, which also show that (2) outperforms Modified-CS [12], another strategy for integrating prior information.

A representative result. To give an example of our results, we state a simplified version of Theorem 12 from Section IV-C, which establishes bounds on the number of measurements for successful ℓ_1-ℓ_1 reconstruction. Here, we rewrite it for $\beta = 1$, which gives not only the simplest result, but also the best bound. Define

$$\begin{aligned} \overline{h} &:= \big|\{i : x_i^\star > 0,\ x_i^\star > w_i\} \cup \{i : x_i^\star < 0,\ x_i^\star < w_i\}\big| \\ \xi &:= \big|\{i : w_i \neq x_i^\star = 0\}\big| - \big|\{i : w_i = x_i^\star \neq 0\}\big|\,, \end{aligned}$$

where $|\cdot|$ denotes the cardinality of a set. Note that $\overline{h}$ is defined on the support $I := \{i : x_i^\star \neq 0\}$ of $x^\star$. Recall that $s = |I|$. Later, we will call $\overline{h}$ the *number of bad components* of w. For example, if $x^\star = (0, 3, -2, 0, 1, 0, 4)$ and $w = (0, 4, 3, 1, 1, 0, 0)$, then $\overline{h} = 2$ (due to 3rd and last components) and $\xi = 1 - 1 = 0$ (4th and 5th components).

Theorem 1 (ℓ_1-ℓ_1 minimization: *simplified*)**.** *Let $x^\star \in \mathbb{R}^n$ be the vector to reconstruct and let $w \in \mathbb{R}^n$ be the prior information. Assume $\overline{h} > 0$ and that there exists at least one index i for which $x_i^\star = w_i = 0$. Let the entries of $A \in \mathbb{R}^{m\times n}$ be i.i.d. Gaussian with zero mean and variance $1/m$. If*

$$m \geq 2\overline{h}\log\Big(\frac{n}{s + \xi/2}\Big) + \frac{7}{5}\Big(s + \frac{\xi}{2}\Big) + 1\,, \tag{4}$$

then, with probability greater than $1 - \exp\big(-\frac{1}{2}(m - \sqrt{m})^2\big)$, $x^\star$ is the unique solution of (2) with $\beta = 1$.

Recall that, with a similar probability, classical CS requires

$$m \geq 2s\log\Big(\frac{n}{s}\Big) + \frac{7}{5}s + 1 \tag{5}$$

measurements to reconstruct $x^\star$ [27]; see also Theorem 4 and Proposition 6 in Section III below. These bounds say that, for large n, (2) requires $O(2\overline{h}\log n)$ measurements whereas classical CS requires $O(2s\log n)$. Recall that, by definition, $\overline{h} \leq s$. Equality holds, i.e., $\overline{h} = s$, only when the supports of $x^\star$ and w are disjoint. This means ℓ_1-ℓ_1 minimization is robust to inaccurate prior information; yet, if $\overline{h}$ is small, (4) can be much smaller than (5). For ℓ_1-ℓ_2 minimization (3), we establish a similar bound: $O(v_\beta \log n)$, where

$$v_\beta \simeq \sum_{i\in I}\big(1 + \beta\,\mathrm{sign}(x_i^\star)(x_i^\star - w_i)\big)^2\,, \tag{6}$$

and $\mathrm{sign}(\cdot)$ returns the sign of a number. The approximation is due to neglecting a term that depends on the disjointness of the supports of $x^\star$ and w; thus, (6) is accurate when $x^\star$ and w have similar supports. Notice that while $\overline{h}$ is independent from β and is determined only by the signs of the entries of $x^\star$ and $x^\star - w$, v_β depends on β and also on the actual values of $x^\star$ and w. Furthermore, as shown by our experiments, in practice, it is much easier to obtain smaller values for $\overline{h}$ than it is for v_β.

A numerical example. We provide a numerical example to illustrate further our results. We generated $x^\star$ with 1000 entries, 70 of which were nonzero, i.e., $n = 1000$ and $s = 70$. The nonzero components of $x^\star$ were drawn from a standard Gaussian distribution. The prior information w was created as $w = x^\star + z$, where z is a 28-sparse vector whose nonzero entries were drawn from a zero-mean Gaussian distribution with standard deviation 0.8. The supports of $x^\star$ and z coincided in 22 positions and differed in 6. This pair of $x^\star$ and w yielded $\overline{h} = 11$ and $\xi = -42$. Plugging the previous values into (4) and (5), we see that ℓ_1-ℓ_1 minimization and classical CS require 136 and 472 measurements for perfect reconstruction with high probability, respectively.

Fig. 1 shows the experimental performance of classical CS and ℓ_1-ℓ_1 and ℓ_1-ℓ_2 minimization, i.e., problems (BP), (2) and (3), respectively. More specifically, it depicts the rate

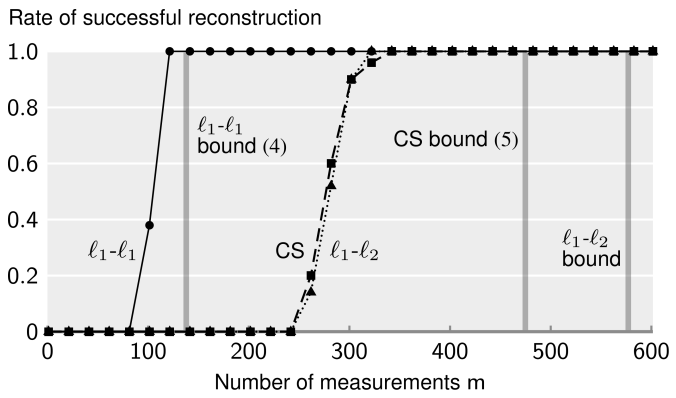


Figure 1. Experimental rate of reconstruction of classical CS (BP), ℓ_1-ℓ_1 minimization, and ℓ_1-ℓ_2 minimization, both with $\beta = 1$. The vertical lines are the bounds for classical CS, and ℓ_1-ℓ_1 and ℓ_1-ℓ_2 minimization.

of success of each problem versus the number of measurements m. For a fixed m, the success rate is the number of times a given problem recovered $x^\star$ with an error smaller than 1% divided by the total number of 50 trials (each trial considered different pairs of A and b). The plot shows that ℓ_1-ℓ_1 minimization required less measurements to reconstruct $x^\star$ successfully than both CS and ℓ_1-ℓ_2 minimization. The curves of the last two, in fact, almost coincide, with ℓ_1-ℓ_2 minimization (line with triangles) having a slightly sharper phase transition. The vertical lines show the bounds (4), (5), and the bound for ℓ_1-ℓ_2 minimization, provided in Section IV. We see that, for this particular example, the bound (4) is quite sharp, while the bound for ℓ_1-ℓ_2 minimization is quite loose (the sharpness of our bounds is discussed in Sections IV and VI). More importantly, this example shows that using prior information properly can improve the performance of CS dramatically.

Our bounds have been used to design an adaptive-rate scheme for state estimation with applications in compressive video background subtraction [40], [41], a reweighted ℓ_1-ℓ_1 minimization scheme [47], [48], and also to design measurements in CS-based communication systems [49].

B. Outline

In Section II, we discuss related work, including the use of other types of "prior information" in CS. Section III introduces fundamental tools in our analysis, which are also used to provide geometrical interpretations of ℓ_1-ℓ_1 and ℓ_1-ℓ_2 minimization. The main results are stated and discussed in Section IV. There, we also provide guidelines on how to improve the prior information in practice. Section V describes experimental results. The main results are proven in Section VI, and the appendix is used for auxiliary results.

II. Related Work

There is a clear analogy of CS with prior information and the distributed source coding problem. Namely, we can view the number of measurements and the reconstruction quality in CS as the information rate and the incurred distortion in coding theory, respectively. As such, CS with prior information at the reconstruction side is reminiscent of the problem of coding with side/prior information at the decoder, a field founded by Slepian and Wolf [50], and Wyner and Ziv [51].

The concept of prior information has appeared in CS under many guises [11], [12], [20], [23], [42]. The work in [11] was apparently the first to consider (1), in particular ℓ_1-ℓ_1 minimization. Specifically, [11] considers dynamic computed tomography, where a prior image helps reconstructing the current one, which is accomplished by solving (2). That work, however, neither provides any kind of analysis nor highlights the benefits of solving (2) with respect to classical CS, i.e., BP. Very recently, [20] considered a variation of (2) where the second term of the objective penalizes differences between x and w, rather than in the sparse domain, in the signals' original domain. Specifically, [20] solves (a Lagrangian version of)

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \|x\|_1 + \beta\|\Psi(x-w)\|_1 \\ \text{subject to} \quad & \Phi\Psi x = y\,, \end{aligned} \tag{7}$$

where A was decomposed as the product of a sensing matrix Φ and a transform matrix Ψ that sparsifies both $x^\star$ and w. Although [20] shows experimentally that (7) requires less measurements than conventional CS to reconstruct MRI images, no analysis or reconstruction guarantees are given for (7).

In [12], prior information refers to an estimate $T \subseteq \{1,\dots,n\}$ of the support of $x^\star$ (see [10], [16], [18] for related approaches). Using the restricted isometry constants of A, [12] provides exact recovery conditions for BP when its objective is modified to $\|x_{T^c}\|_1$, where x_T denotes the components of x indexed by the set T, and T^c is the complement of T in $\{1,\dots,n\}$. The resulting problem is called Modified-CS (Mod-CS), against which we benchmark the performance of (2) and (3) in Section V. When T is a reasonable estimate of the support of $x^\star$, those conditions are shown to be milder than the ones in [24], [25] for standard BP. Then, [12] considers prior information as we do: there is an estimate of the support of $x^\star$ as well as of the value of the respective nonzero components. However, it solves a problem slightly different from (3). Namely, the objective of (3) is replaced with $\|x_{T^c}\|_1 + \beta\|x_T - w_T\|_2^2$. Although some experimental results are presented, no analysis is given for that problem.

A popular modification of BP, of which Mod-CS is a particular instance, considers the weighted ℓ_1-norm $\|x\|_r := \sum_{i=1}^n r_i x_i$, where $r_i \geq 0$ is a known weight. This norm penalizes each component of x according to the magnitude of the corresponding weight and, thus, requires "prior information" about x. The weight r_i associated to the component x_i can, for example, be proportional to the probability of $x_i^\star = 0$. Several works studied weighted ℓ_1-norm minimization [13]–[17], and some [19] used tools similar to ours.

Alternative work has considered

$$\underset{x}{\text{minimize}} \;\; \|x\|_1 + \beta\, g(x-w) + \lambda\|Ax-y\|_2^2\,, \tag{8}$$

with $\lambda > 0$, which can be viewed as a Lagrangian version of

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \|x\|_1 + \beta\, g(x-w) \\ \text{subject to} \quad & \|Ax-y\|_2 \leq \sigma\,. \end{aligned} \tag{9}$$

Problem (9) is a generalization of (1) for noisy scenarios, and we will provide bounds on the number of measurements that

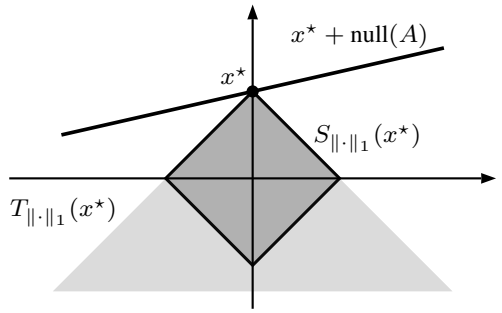


Figure 2. Visualization of the nullspace property in Proposition 2 for BP.

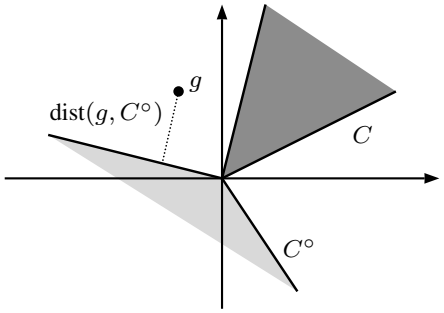


Figure 3. Illustration of how the Gaussian width measures the width of a cone, according to Proposition 3.

it requires for successful reconstruction with $g = \|\cdot\|_1$ and $g = \frac{1}{2}\|\cdot\|_2^2$. Problem (8) has appeared before in [42], in the context of dynamical system estimation. Specifically, the state x^t of a system at time t evolves as $x^{t+1} = f_t(x^t) + \epsilon^t$, where f_t models the system's dynamics at time t, and ϵ^t accounts for modeling errors. Observations of the state x^t are taken as $y^t = A_t x^t + \eta^t$, where A_t is the observation matrix and η^t is noise. The goal is to estimate the state x^t given the observations y^t. The state of the system in the previous instant, x^{t-1}, can be used as prior information by making $w^t = f_{t-1}(x^{t-1})$. If the modeling error ϵ^t is Gaussian and the state x^t is assumed sparse, then x^t can be estimated by solving (8) with $g = \|\cdot\|_2^2$; if the modeling noise is Laplacian, we set $g = \|\cdot\|_1$ instead. Although [42] does not provide any analysis, their experimental results show that, among several strategies for state estimation, including Kalman filtering, (8) with $g = \|\cdot\|_1$ yields the best results. If we take into account the relation between (8) and (9), our theoretical analysis can be used to provide an explanation. Applying the KKT conditions to problem (9) reveals that it has the same solution $\overline{x}$ as (8) if $\|A\overline{x} - y\|_2 = \sigma$ and λ is the *optimal* dual variable of (9). Note that obtaining such λ without first solving (9) is nearly impossible. In contrast, in several applications, it is relatively easy to obtain accurate bounds σ on the magnitude of the acquisition noise. For related approaches, see [23], [52].

Finally, we mention that the phase transition phenomenon in sparse recovery problems was first studied in [33], [53], [54], and that alternative reconstruction problems, such as message passing [55], also have precise phase transitions [55]–[57].

III. The Geometry of ℓ_1-ℓ_1 and ℓ_1-ℓ_2 Minimization

This section introduces concepts and results in CS used in our analysis. We follow the approach of [27], since it leads to the current best CS bounds for Gaussian measurements, and provides the means to understand some of our definitions.

A. Known Results and Tools

The concept of *Gaussian width* plays a key role in [27]. Originally proposed in [58] to quantify the probability of a randomly oriented subspace intersecting a cone, the Gaussian width has been used in several CS-related results [27], [32], [34], [36]. Before defining it, we analyze the optimality conditions of linearly constrained convex optimization problems.

The nullspace property. Consider a real-valued convex function $f : \mathbb{R}^n \to \mathbb{R}$ and the following optimization problem:

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f(x) \\ \text{subject to} & Ax = y\,. \end{array} \tag{10}$$

Assume $Ax = y$ has at least one solution, say, $x^\star$. The set of all solutions of $Ax = y$, i.e., the feasible set of (10), is $\mathcal{A} := x^\star + \text{null}(A)$, where $\text{null}(A) := \{x : Ax = 0\}$ is the *nullspace* of A. To determine if a given $x^\star \in \mathcal{A}$ is a solution of (10), we use the concept of *tangent cone* of f at $x^\star$:

$$T_f(x^\star) := \text{cone}\big(S_f(x^\star) - x^\star\big)\,, \tag{11}$$

where $\text{cone}\, C := \{\alpha c : \alpha \geq 0,\, c \in C\}$ is the *cone generated by the set* C, and $S_f(x^\star) := \{x : f(x) \leq f(x^\star)\}$ is the *sublevel set of* f *at* $x^\star$. See [59, Prop. 5.2.1, Thm. 1.3.4].

Proposition 2 (Prop. 2.1 in [27])**.** *$x^\star$ is the unique optimal solution of* (10) *if and only if* $T_f(x^\star) \cap \text{null}(A) = \{0\}$.

Although this proposition was stated in [27, Prop.2.1] for f equal to an atomic norm, its proof holds for any real-valued convex function. Fig. 2 illustrates it for $f(x) = \|x\|_1$, i.e., for BP. It shows the respective sublevel set $S_{\|\cdot\|_1}(x^\star)$ and tangent cone $T_{\|\cdot\|_1}(x^\star)$ at a "sparse" point $x^\star$. In the figure, $\mathcal{A} = x^\star + \text{null}(A)$ intersects $T_{\|\cdot\|_1}(x^\star)$ at $x^\star$ only, that is, $T_{\|\cdot\|_1}(x^\star) \cap (x^\star + \text{null}(A)) = \{x^\star\}$. Subtracting $x^\star$ to both sides, we obtain the condition in Proposition 2.

Gaussian width. When A is generated randomly, its nullspace has a random orientation, and the condition in Proposition 2 holds or not with a given probability. The smaller the width (or aperture) of $T_f(x^\star)$, the more likely that condition will hold. Such a statement was formalized for Gaussian matrices A by Gordon in [58]. To measure the width of a set $S \in \mathbb{R}^n$, Gordon defined the *Gaussian width*:

$$\overline{w}(S) := \mathbb{E}_g\Big[\sup_{z \in S}\; g^\top z\Big]\,, \tag{12}$$

where $g \sim \mathcal{N}(0, I_n)$ is a vector of n independent, zero-mean, and unit-variance Gaussian random variables, and $\mathbb{E}_g[\cdot]$ is the expected value with respect to g. When the set is a cone C, i.e., $x \in C \Rightarrow \alpha\, x \in C$ for all $\alpha \geq 0$, we have to intersect C with the unit ℓ_2-norm sphere in $\mathbb{R}^n$: $\mathbb{S}_n(0,1) := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$. To simplify notation, we define

$$w(C) := \overline{w}(C \cap \mathbb{S}_n(0,1)) = \mathbb{E}_g\Big[\sup_{z \in C \cap \mathbb{S}_n(0,1)}\; g^\top z\Big]\,. \tag{13}$$

It turns out that $\overline{w}(C \cap \mathbb{S}_n(0,1)) = \overline{w}(C \cap \mathbb{B}_n(0,1))$, where $\mathbb{B}_n(0,1) := \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$ is the unit ℓ_2-norm ball in $\mathbb{R}^n$.[1] As a result, the Gaussian width of a cone C is the expected distance of a Gaussian vector g to the *polar cone of* C, defined as $C^\circ := \{y : y^\top z \leq 0\,, \forall\, z \in C\}$:

Proposition 3 (Example 2.3.1 in [59]; Prop. 3.6 in [27])**.** *The Gaussian width of a cone C can be written as*

$$w(C) = \mathbb{E}_g\Big[\operatorname{dist}(g, C^\circ)\Big]\,, \tag{14}$$

where $\operatorname{dist}(x,S) := \min\{\|z-x\|_2 : z \in S\}$ *denotes the distance of the point x to the set S.*[2]

This follows from the fact that the support function of a "truncated" cone is the distance to its polar cone [59, Ex. 2.3.1]; and can be proved by computing the dual of the optimization problem in (12) [27, Prop. 3.6]. Proposition 3 provides not only a way easier than (12) to compute Gaussian widths of cones, but also a geometrical explanation of why the Gaussian width measures the width of a cone. The wider the cone C, the smaller its polar cone C°. Therefore, the expected distance of a Gaussian vector g to C° increases as C° gets smaller or, equivalently, as C gets wider; see Fig. 3.

From geometry to CS bounds. In [58], Gordon used the concept of Gaussian width to compute bounds on the probability of a cone intersecting a subspace whose orientation is uniformly distributed, e.g., the nullspace of a Gaussian matrix. More recently, [60] showed that those bounds are sharp. Based on Gordon's result, on Proposition 2 (and its generalization for the case where the constraints of (10) are $\|Ax-y\|_2 \leq \sigma$), and a concentration of measure result, [27] establishes:

Theorem 4 (Corollary 3.3 in [27])**.** *Let $A \in \mathbb{R}^{m\times n}$ be a matrix whose entries are i.i.d., zero-mean Gaussian random variables with variance $1/m$. Assume $f : \mathbb{R}^n \to \mathbb{R}$ is convex, and let $\lambda_m := \mathbb{E}_g[\|g\|_2]$ denote the expected length of a zero-mean, unit-variance Gaussian vector $g \sim \mathcal{N}(0, I_m)$ in $\mathbb{R}^m$.*

1) *Suppose $y = Ax^\star$ and let*

$$\hat{x} = \begin{array}{ll} \arg\min\limits_{x} & f(x) \\ \text{s.t.} & Ax = y\,, \end{array} \tag{15}$$

and

$$m \geq w(T_f(x^\star))^2 + 1\,. \tag{16}$$

Then, $\hat{x} = x^\star$ is the unique solution of (15) *with probability greater than* $1 - \exp\big(-\frac{1}{2}\big[\lambda_m - w(T_f(x^\star))\big]^2\big)$.

2) *Suppose $y = Ax^\star + \eta$, where $\|\eta\|_2 \leq \sigma$ and let*

$$\hat{x} \in \begin{array}{ll} \arg\min\limits_{x} & f(x) \\ \text{s.t.} & \|Ax-y\|_2 \leq \sigma\,. \end{array} \tag{17}$$

Define $0 < \epsilon < 1$ and let

$$m \geq \frac{w(T_f(x^\star))^2 + 3/2}{(1-\epsilon)^2}\,. \tag{18}$$

Then, $\|\hat{x} - x^\star\|_2 \leq 2\sigma/\epsilon$ with probability greater than $1 - \exp\big(-\frac{1}{2}\big[\lambda_m - w(T_f(x^\star)) - \epsilon\sqrt{m}\big]^2\big)$.

Theorem 4 was stated in [27] for f equal to an atomic norm. Its proof, however, remains valid when f is any convex function. Note, in particular, that (15) becomes (BP), (2), and (3) when $f(x)$ is $\|x\|_1$, $\|x\|_1 + \beta\|x-w\|_1$, and $\|x\|_1 + \frac{\beta}{2}\|x-w\|_2^2$, respectively; and (17) becomes the noise-robust version of these problems. In this paper, we focus on the noise-free versions of (2) and (3). However, we remark that the bounds we derive also apply to their noise-robust versions because of part 2) of Theorem 4. The quantity λ_m can be sharply bounded as $m/\sqrt{m+1} \leq \lambda_m \leq \sqrt{m}$ [27]. One of the steps of the proof of Theorem 4 shows that (16) implies $w(T_f(x^\star)) \leq \lambda_m$ and (18) implies $w(T_f(x^\star)) + \epsilon\sqrt{m} \leq \lambda_m$. Roughly, the theorem says that, given the noiseless (resp. noisy) measurements $y = Ax^\star$ (resp. $y = Ax^\star + \eta$), we can recover $x^\star$ exactly (resp. with an error of $2\sigma/\epsilon$), provided the number of measurements is larger than a function of the Gaussian width of $T_f(x^\star)$. It is rare, however, to be able to compute Gaussian widths in closed-form; instead, one usually upper bounds it. As proposed in [27], a useful tool to obtain such bounds is Jensen's inequality [59, Thm.B.1.1.8], Proposition 3, and the following proposition. Recall that the *normal cone $N_f(x)$ of a function f at a point x* is the polar of its tangent cone: $N_f(x) := T_f(x)^\circ$. Also, $\partial f(x) := \{d : f(y) \geq f(x) + d^\top(y-x),\ \text{for all } y\}$ is the *subgradient of f at a point x* [59].

Proposition 5 (Theorem 1.3.5, Chapter D, in [59])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and suppose $0 \notin \partial f(x)$ for a given $x \in \mathbb{R}^n$. Then, $N_f(x) = \operatorname{cone} \partial f(x)$.*

Using Propositions 3 and 5, [27] proves:[3]

Proposition 6 (Proposition 3.10 in [27])**.** *Let $x^\star \neq 0$ be an s-sparse vector in $\mathbb{R}^n$. Then,*

$$w\big(T_{\|\cdot\|_1}(x^\star)\big)^2 \leq 2s\log\Big(\frac{n}{s}\Big) + \frac{7}{5}s\,. \tag{19}$$

Together with Theorem 4, this means that if $m \geq 2s\log(n/s) + (7/5)s + 1$, then BP recovers $x^\star$ from m noiseless Gaussian measurements with high probability. A similar result holds for noisy measurements.

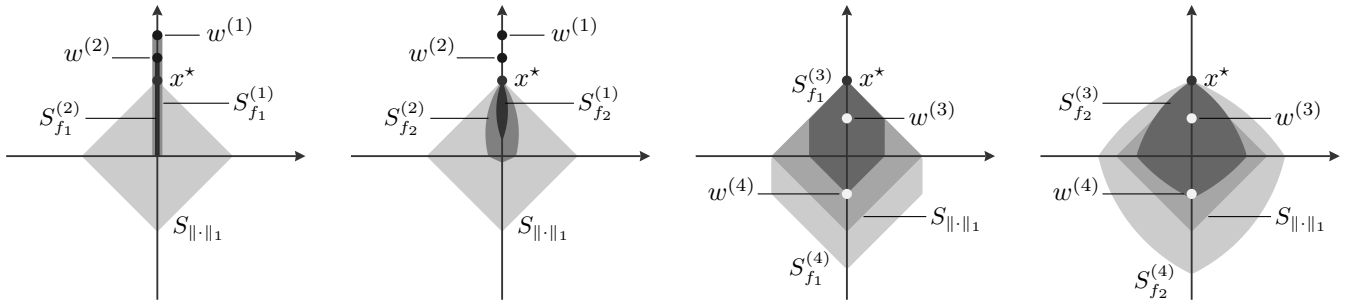
B. The Geometry of ℓ_1-ℓ_1 and ℓ_1-ℓ_2 Minimization

Theorem 4 applies to CS by making $f(x) = \|x\|_1$. Since it is applicable to any convex function f, we will use it to characterize problems (2) and (3), that is, when f is $f_1(x) := \|x\|_1 + \beta\|x-w\|_1$ and $f_2(x) := \|x\|_1 + \frac{\beta}{2}\|x-w\|_2^2$, respectively. In particular, we want to understand the relation between the Gaussian widths of the tangent cones associated with these functions and the one associated with the ℓ_1-norm. If the former is smaller, we may obtain reconstruction bounds for (2) and (3) smaller than the one in (19). In the same way that Proposition 6 bounded the squared Gaussian width of the

[1] That is because the maximizer of the problem in (13) is always in $\mathbb{S}_n(0,1)$. To see that, suppose it is not, i.e., for a fixed g, $z_g := \sup\{g^\top z : z \in C \cap \mathbb{B}_n(0,1)\}$ and $z_g \notin \mathbb{S}_n(0,1)$. This means $\|z_g\|_2 < 1$. Since C is a cone, $\hat{z}_g := z_g/\|z_g\|_2 \in C \cap \mathbb{S}_n(0,1)$. And $g^\top \hat{z}_g = (1/\|\hat{z}_g\|_2) g^\top z_g > g^\top z_g$, contradicting the fact that z_g is optimal.

[2] This result is stated in [27] as an inequality, i.e., with $\leq$ in place of $=$. Because of the previous footnote, the result is in fact an equality.

[3] We noticed an extra factor of $\sqrt{\pi}$ in equation (73) of [27] (proof of Proposition 3.10). Namely, π in (73) should be replaced by $\sqrt{\pi}$. As a consequence, equation (74) in that paper can be replaced, for example, by our equation (59). In that case, the number of measurements in Proposition 3.10 in [27] should be corrected from $2s\log(n/s) + (5/4)s$ to $2s\log(n/s) + (7/5)s$.



(a) ℓ_1-ℓ_1: good components (b) ℓ_1-ℓ_2: good components (c) ℓ_1-ℓ_1: bad components (d) ℓ_1-ℓ_2: bad components

Figure 4. Sublevel sets of functions f_1 and f_2 with $\beta = 1$ for $x^\star = (0, 1)$. In both (a) and (b), the prior information is $w^{(1)} = (0, 1.6)$ and $w^{(2)} = (0, 1.3)$, while in (c) and (d) it is $w^{(3)} = (0, 0.5)$, and $w^{(4)} = (0, -0.5)$. For reference, the sublevel set $S_{\|\cdot\|_1}$ of the ℓ_1-norm at $x^\star$ is also shown in all figures.

ℓ_1-norm in terms of the key parameters n and s, we seek to do the same for f_1 and f_2. To find out the key parameters in this case and also to gain some intuition about the problem, Fig. 4 shows the sublevel sets of f_1 and f_2 with $\beta = 1$ and in two dimensions, i.e., for $n = 2$. Recall that, according to (11), one can estimate tangent cones by observing the sublevel sets that generate them. We set $x^\star = (0, 1)$ in all plots of Fig. 4 and consider four different vectors as prior information w: $w^{(1)} = (0, 1.6)$ and $w^{(2)} = (0, 1.3)$ in Figs. 4(a) and 4(b); and $w^{(3)} = (0, 0.5)$ and $w^{(4)} = (0, -0.5)$ in Figs. 4(c) and 4(d). The sublevel sets are denoted with

$$S_{f_i}^{(j)} := \{x \,:\, \|x\|_1 + g_i(x - w^{(j)}) \le \|x^\star\|_1 + g_i(x^\star - w^{(j)})\}\,,$$

where $i = 1, 2$, $j = 1, 2, 3, 4$, and $g_1 = \|\cdot\|_1$ and $g_2 = \frac{1}{2}\|\cdot\|_2^2$. For reference, we also show the sublevel set $S_{\|\cdot\|_1}$ associated with BP. The sublevel sets of f_1 are shown in Figs. 4(a) and 4(c), whereas the sublevel sets of f_2 are shown in Figs. 4(b) and 4(d). For example, the sublevel set $S_{f_1}^{(1)}$ in Fig. 4(a) can be computed in closed-form as $S_{f_1}^{(1)} = \{(0, x_2) : 0 \le x_2 \le 1.6\}$. The cone generated by this set is the axis $x_1 = 0$. In the same figure, $S_{f_1}^{(2)} = \{(0, x_2) : 0 \le x_2 \le 1.3\}$ and it generates the same cone. Hence, both $S_{f_1}^{(1)}$ and $S_{f_1}^{(2)}$ generate the tangent cone $\{(0, x_2) \,:\, x_2 \in \mathbb{R}\}$, which has Gaussian width smaller than $w(T_{\|\cdot\|_1}(x^\star))$.[4] When we consider f_2 and the same prior information vectors, as in Fig. 4(b), the tangent cones have larger widths, which are still smaller than the width of $T_{\|\cdot\|_1}(x^\star)$. Since small widths are desirable, we say that the nonzero components of the w's in Figs. 4(a) and 4(b) are *good components*. On the other hand, the cones generated by the sublevel sets of Fig. 4(c) coincide with $T_{\|\cdot\|_1}(x^\star)$, and the cones generated by the sublevel sets of Fig. 4(d) have widths larger than $T_{\|\cdot\|_1}(x^\star)$. Therefore, we say that the nonzero components of the w's in Figs. 4(c) and 4(d) are *bad components*. Fig. 4 illustrates the concepts of good and bad components only for $x_i^\star > 0$. For $x_i^\star < 0$, there is geometric symmetry. This motivates the following definition.

[4]Using (13) and denoting $g = (g_1, g_2) \sim \mathcal{N}(0, I_2)$, it can be shown that $w(T_{f_1}(x^\star)) = \mathbb{E}_g[\sup_z\{g^\top z : z = (0, \pm 1)\}] = \mathbb{E}_g[|g_2|] = 2/\sqrt{2\pi} \simeq 0.8$. In contrast, noting that $T_{\|\cdot\|_1}(x^\star)$ is a rotation of the nonnegative orthant and using [60, §3], we have $w(T_{\|\cdot\|_1}(x^\star)) = n/2 = 1 > 0.8$. Note that the difference between the Gaussian widths increases with the ambient dimension.

Definition 7 (Good and bad components)**.** *Let $x^\star \in \mathbb{R}^n$ be the vector to reconstruct and let $w \in \mathbb{R}^n$ be the prior information. For $i = 1, \ldots, n$, a component w_i is considered good if*

$$x_i^\star > 0 \;\text{ and }\; x_i^\star < w_i \qquad \text{or} \qquad x_i^\star < 0 \;\text{ and }\; x_i^\star > w_i\,,$$

and w_i is considered bad if

$$x_i^\star > 0 \;\text{ and }\; x_i^\star > w_i \qquad \text{or} \qquad x_i^\star < 0 \;\text{ and }\; x_i^\star < w_i\,.$$

Note that good and bad components are defined only on the support of $x^\star$ and that the inequalities in the definition are strict. Although good and bad components were motivated geometrically, they arise naturally in our proofs. Notice that Definition 7 (and Fig. 4) consider only components w_i such that $w_i \neq x_i^\star$ and for which $x_i^\star \neq 0$. The other components, of course, also influence the Gaussian width; note the role of ξ in Theorem 1. This will be clear when we present our main results in the next section.

IV. Main Results

In this section we present our main results, namely reconstruction guarantees for ℓ_1-ℓ_1 and ℓ_1-ℓ_2 minimization. After some definitions and preliminary results, we present the results for ℓ_1-ℓ_1 minimization first, and the results for ℓ_1-ℓ_2 minimization next. All proofs are relegated to Section VI.

A. Definitions and Preliminary Results

Definition 8 (Support sets)**.** *Let $x^\star \in \mathbb{R}^n$ be the vector to reconstruct and let $w \in \mathbb{R}^n$ be the prior information. We define*

$$\begin{aligned} I &:= \{i \,:\, x_i^\star \neq 0\} & J &:= \{j \,:\, x_j^\star \neq w_j\} \\ I^c &:= \{i \,:\, x_i^\star = 0\} & J^c &:= \{j \,:\, x_j^\star = w_j\} \\ I_+ &:= \{i \,:\, x_i^\star > 0\} & J_+ &:= \{j \,:\, x_j^\star > w_j\} \\ I_- &:= \{i \,:\, x_i^\star < 0\} & J_- &:= \{j \,:\, x_j^\star < w_j\}\,. \end{aligned}$$

To simplify notation, we denote the intersection of two sets A and B with the product $AB := A \cap B$. Then, the set of good components can be written as $I_+J_- \cup I_-J_+$, and the set of bad components can be written as $I_+J_+ \cup I_-J_-$.

Definition 9 (Cardinality of sets)**.** *The number of good components, the number of bad components, the sparsity of $x^\star$,*

the sparsity of $x^\star - w$, and the cardinality of the union of the supports of $x^\star$ and $x^\star - w$ are represented, respectively, by

$$\begin{aligned} h &:= \big|I_+J_-\big| + \big|I_-J_+\big| \\ \overline{h} &:= \big|I_+J_+\big| + \big|I_-J_-\big| \\ s &:= |I| \\ l &:= |J| \\ q &:= \big|I \cup J\big| . \end{aligned}$$

All these quantities are nonnegative. Before moving to our main results, we present the following useful lemma:

Lemma 10. *For $x^\star$ and w as in Definition 8,*

$$\big|IJ\big| = h + \overline{h} \tag{20}$$
$$|IJ^c| = s - (h + \overline{h}) . \tag{21}$$
$$|I^cJ| = q - s \tag{22}$$
$$\big|I^cJ^c\big| = n - q \tag{23}$$

Proof. Identity (20) is proven by noticing that I_+ and I_- partition I, and J_+ and J_- partition J. Then,

$$\begin{aligned} \big|IJ\big| &= \big|I_+J\big| + \big|I_-J\big| \\ &= \big|I_+J_+\big| + \big|I_+J_-\big| + \big|I_-J_+\big| + \big|I_-J_-\big| \\ &= h + \overline{h} . \end{aligned}$$

To prove (21), we use (20) and the fact that J and J^c are a partition of $\{1, \dots, n\}$:

$$s = \big|I\big| = \big|IJ\big| + \big|IJ^c\big| = (h + \overline{h}) + \big|IJ^c\big| ,$$

from which (21) follows. To prove (22), we use the identity $I \cup J = \big(I^cJ\big) \cup \big(IJ\big) \cup \big(IJ^c\big)$, where I^cJ, IJ and IJ^c are pairwise disjoint. Then, using (20) and (21),

$$q = \big|I \cup J\big| = \big|I^cJ\big| + \big|IJ\big| + \big|IJ^c\big| = \big|I^cJ\big| + s .$$

Finally, (23) holds because

$$n = \big|I\big| + \big|I^c\big| = \big|IJ\big| + \big|IJ^c\big| + \big|I^cJ\big| + \big|I^cJ^c\big| = q + \big|I^cJ^c\big| ,$$

where we used (20), (21), and (22). □

From Lemma 10, we can easily obtain the following identities, which will be frequently used:

$$\big|I^cJ\big| + \big|IJ^c\big| = q - (h + \overline{h}) \tag{24}$$
$$\big|I^cJ\big| + \big|IJ^c\big| + 2\big|I^cJ^c\big| = 2n - (q + h + \overline{h}) \tag{25}$$
$$\big|I^cJ\big| - \big|IJ^c\big| = q + h + \overline{h} - 2s . \tag{26}$$

Finally, note that (23) allows interpreting q as the size of the union of the supports of $x^\star$ and w: since both $x^\star$ and w are zero in I^cJ^c, q is the number of components in which at least one of them is not zero.

B. ℓ_1-ℓ_1 Minimization

We now state our results for ℓ_1-ℓ_1 minimization, which come into two forms of bounds for $w(T_{f_1}(x^\star))^2$: sharp but uninformative bounds, and not so sharp but informative bounds. We start with the latter. To simplify the presentation, we first enumerate some conditions used for $\beta \neq 1$:

Definition 11 (Conditions for $\beta \neq 1$)**.**

$$\frac{q-s}{2n-(q+h+\overline{h})} \leq \frac{1-\beta}{1+\beta}\left(\frac{q+h+\overline{h}}{2n}\right)^{\frac{4\beta}{(\beta+1)^2}} \tag{C1}$$
$$\frac{q-s}{2n-(q+h+\overline{h})} \geq \frac{1-\beta}{1+\beta}\left(\frac{s}{q}\right)^{\frac{4\beta}{(1-\beta)^2}} \tag{C2}$$
$$\frac{s-(h+\overline{h})}{2n-(q+h+\overline{h})} \leq \frac{\beta-1}{\beta+1}\left(\frac{q+h+\overline{h}}{2n}\right)^{\frac{4\beta}{(\beta+1)^2}} \tag{C3}$$
$$\frac{s-(h+\overline{h})}{2n-(q+h+\overline{h})} \geq \frac{\beta-1}{\beta+1}\left(\frac{h+\overline{h}}{s}\right)^{\frac{4\beta}{(\beta-1)^2}} . \tag{C4}$$

Theorem 12 (ℓ_1-ℓ_1 minimization)**.** *Let $x^\star \in \mathbb{R}^n$ be the vector to reconstruct and let $w \in \mathbb{R}^n$ be the prior information. Let $f_1(x) = \|x\|_1 + \beta\|x - w\|_1$ with $\beta > 0$, and assume $x^\star \neq 0$, $w \neq x^\star$, and $q < n$.*

1) *Let $\beta = 1$, and assume there exists at least one bad component, i.e., $\overline{h} > 0$. Then,*

$$w\big(T_{f_1}(x^\star)\big)^2 \leq 2\overline{h}\log\Big(\frac{2n}{q+h+\overline{h}}\Big) + \frac{7}{10}(q+h+\overline{h}) . \tag{27}$$

2) *Let $\beta < 1$.*
   a) *If* (C1) *holds, then*

$$w\big(T_{f_1}(x^\star)\big)^2 \leq 2\Big[\overline{h} + (s-\overline{h})\frac{(1-\beta)^2}{(1+\beta)^2}\Big] \times \times \log\Big(\frac{2n}{q+h+\overline{h}}\Big) + s + \frac{2}{5}(q+h+\overline{h}) . \tag{28}$$

   b) *If $q > s$ and* (C2) *holds, then*

$$w\big(T_{f_1}(x^\star)\big)^2 \leq 2\Big[\overline{h}\frac{(1+\beta)^2}{(1-\beta)^2} + s - \overline{h}\Big]\log\Big(\frac{q}{s}\Big) + \frac{7}{5}s . \tag{29}$$

3) *Let $\beta > 1$.*
   a) *If* (C3) *holds, then*

$$w\big(T_{f_1}(x^\star)\big)^2 \leq 2\Big[\overline{h} + (q+h-s)\frac{(\beta-1)^2}{(\beta+1)^2}\Big] \times \times \log\Big(\frac{2n}{q+h+\overline{h}}\Big) + l + \frac{2}{5}(q+h+\overline{h}) . \tag{30}$$

   b) *If $s > h + \overline{h} > 0$ and* (C4) *holds, then*

$$w\big(T_{f_1}(x^\star)\big)^2 \leq 2\Big[\overline{h}\frac{(\beta+1)^2}{(\beta-1)^2} + q + h - s\Big] \times \times \log\Big(\frac{s}{h+\overline{h}}\Big) + l + \frac{2}{5}(h+\overline{h}) . \tag{31}$$

The proof can be found in Section VI-B. Similarly to Proposition 6, the previous theorem establishes upper bounds on $w(T_{f_1}(x^\star))^2$ that depend only on the key parameters n, s, β, q, h, and $\overline{h}$. Together with Theorem 4, it then provides (useful) bounds on the number of measurements that guarantee ℓ_1-ℓ_1 reconstruction with high probability. The assumption $q < n$ means that the union of the supports of $x^\star$ and w differs from the full set $\{1, \dots, n\}$ or, equivalently, there is at least one index i for which $x^\star_i = w_i = 0$. Assuming $w \neq x^\star$ and $x^\star \neq 0$

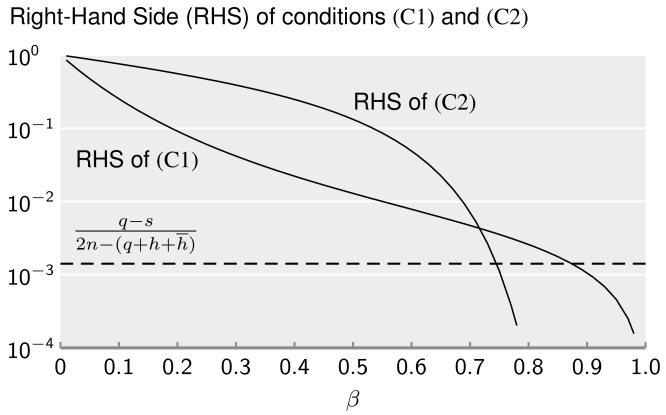


Figure 5. Values of the right-hand side (RHS) of conditions (C1) and (C2) from case 2) of Theorem 12, for the example of Fig. 1.

is equivalent to assuming that the sets J and I are nonempty, respectively.[5]

The theorem is divided into three cases: 1) $\beta = 1$, 2) $\beta < 1$, and 3) $\beta > 1$. We will see that, although rare in practice, the theorem may not cover all possible values of β, due to the conditions imposed in cases 2) and 3). Recall that Theorem 1 in Section I instantiates case 1), i.e., $\beta = 1$, but in a slightly different format. Namely, to obtain (4) from (27), notice that $\xi = |I^c J| - |I J^c|$ and that (26) implies $(q + h + \overline{h})/2 = s + \xi/2$. Therefore, the observations made for Theorem 1 apply to case 1) of the previous theorem. We add to those observations that the assumption that there is at least one bad component, i.e., $\overline{h} > 0$, is necessary to guarantee $0 \notin \partial f_1(x^\star)$ and, hence, that we can use Proposition 5. In fact, it will be shown in part 1) of Lemma 19 that $0 \notin \partial f_1(x^\star)$ if and only if $\overline{h} > 0$ or $\beta \neq 1$. Thus, the assumption $\overline{h} > 0$ can be dropped in cases 2) and 3), where $\beta \neq 1$. Note that the quantities on the right-hand side of (27) are well defined and positive: the assumption that $x^\star \neq 0$ implies $q = |I \cup J| > 0$; and the assumption that $q < n$, i.e., $|I^c J^c| > 0$, and (25) imply $2n > q + h + \overline{h}$.

In case 2), $\beta < 1$ and we have two subcases: when condition (C1) holds, $w(T_{f_1}(x^\star))^2$ is bounded as in (28); when condition (C2) holds, it is bounded as in (29). These subcases are not necessarily disjointed nor are they guaranteed to cover the entire interval $0 < \beta < 1$.[6] Fig. 5 shows how conditions (C1) and (C2) vary with β for the example in Fig. 1. There, we had $n = 1000$, $s = 70$, $h = 11$, $\overline{h} = 11$, and $q = 76$. The right-hand side of conditions (C1) and (C2) vary with β as shown in the figure, and the dashed line represents the left-hand side of (C1) and (C2), which does not vary with β. We can see that (C1) holds for $0 < \beta \lessapprox 0.88$, and (C2) holds for $0.75 \lessapprox \beta < 1$. Therefore, both conditions are valid in the interval $0.75 < \beta < 0.88$. For instance, if $\beta = 0.8$, the bounds in (28) and (29) give 180 and 255 (rounding up), respectively. Both values are larger than the one for $\beta = 1$, which is given by (27) and equal to 135. Indeed, the bound in (27) is almost always smaller than the one in (28): using (26), it can be shown that the linear, non-dominant terms in (27) are smaller than the linear terms in (28) whenever

$$\xi < \frac{2}{5}(q + h + \overline{h})\,. \tag{32}$$

Furthermore, the dominant term in (27), namely the one involving the log, is always smaller than the dominant term in (28). So, even if (32) does not hold, (27) is in general smaller than (28). Curiously, the bound in (28) is minimized for $\beta = 1$ but, in that case, condition (C1) will not hold unless $q = s$ [according to (22), that would mean that $x^\star$ and w have exactly the same support]. The bound in (29), valid only if $q > s$, can be much larger than both (27) and (28) when β is close to 1: this is due to the term $(1+\beta)^2/(1-\beta)^2$ and to the fact that (29) is valid only for values of β near 1 [cf. (C2) and Fig. 5]. From this analysis, we conclude that the bounds given in case 2) will not be sharp near 1. Yet, the bound for $\beta = 1$, i.e., (27), is the sharpest one in the theorem since, as we will see in its proof, is the one with the fewest number of approximations. Case 3) in the theorem is similar to case 2): the expression for both the conditions and the bounds are very similar. The observations made to case 2) then also apply to case 3). Note, for example, that in case 3b) it is assumed $s > h + \overline{h} > 0$. According to (20) and (21), this is equivalent to saying that there is at least one index i for which $x_i^\star \neq 0$ and $w_i \neq x_i^\star$ and another index j for which $x_j^\star \neq 0$ and $w_j = x_j^\star$. The most striking fact about Theorem 12 is that its expressions depend only on the quantities given in Definition 9, which depend on the signs of $x_i^\star$ and $x_i^\star - w_i$, but not on their magnitude. As we will see shortly, that is no longer the case for ℓ_1-ℓ_2 minimization.

A sharper bound. We now present a bound for ℓ_1-ℓ_1 minimization that is sharper than the ones in Theorem 12. However, it is not as informative and has to be computed numerically. We use the following notation for intervals in the real line: for $b \geq 0$, $\mathcal{I}(a, b) := [a - b, a + b]$.

Theorem 13 (ℓ_1-ℓ_1 minimization: shaper bound)**.** *Let $x^\star, w \in \mathbb{R}^n$ be as in Theorem 12. Assume $x^\star \neq 0$, $w \neq x^\star$, and that either $\overline{h} > 0$ or $\beta \neq 1$. Then,*

$$\begin{aligned} w(T_{f_1}(x^\star))^2 \leq \overline{h} + h + \min_{t \geq 0} \Bigg\{ & \Big[\overline{h}(\beta+1)^2 + h(\beta-1)^2\Big] t^2 \\ & + \sum_{i \in I J^c} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i,\, \mathcal{I}(t\,\operatorname{sign}(x_i^\star),\, t\beta)\Big)^2\Big] \\ & + \sum_{i \in I^c J} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i,\, \mathcal{I}(-t\,\beta\,\operatorname{sign}(w_i),\, t)\Big)^2\Big] \\ & + \sum_{i \in I^c J^c} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i,\, \mathcal{I}(0,\, t(\beta+1))\Big)^2\Big] \Bigg\}. \end{aligned} \tag{33}$$

The proof can be found in Section VI-C. The expected distance of a Gaussian scalar random variable to an interval can be computed exactly, as a function of the Q-function; see (52) in Lemma 17, Section VI. Therefore, the right-hand side of (33) can also be computed exactly, although it requires a numerical procedure to solve the optimization

[5] Assuming $x^\star \neq w$ is necessary to guarantee $0 \notin \partial f_1(x^\star)$, as shown in Lemma 19. Of course, taking $w = x^\star$ works well in practice; our theory, however, does not cover that specific case.

[6] If, for example, $n = 20$, $s = 15$, $q = 16$, $h = 10$, and $\overline{h} = 5$, neither (C1) nor (C2) hold for $\beta = 0.9$. In this case, however, $x^\star$ is not "sparse," as 75% of its entries are nonzero. Increasing n to, e.g., 40, makes (C1) hold.

problem in t. The bound in (33) is reminiscent of the bounds in [60, Theorem 4.3] and [36, Proposition 2], which have sharpness guarantees, i.e., there are polynomial expressions on the Gaussian width that upper bound the respective bound. Unfortunately, the proof techniques used in those sharpness results cannot be used in our case, since they only apply to norms and $f_1(x) = \|x\|_1 + \beta\|x - w\|_1$ is not a norm. Yet, as we will see in Section V, (33) describes precisely the experimental performance of ℓ_1-ℓ_1 minimization.

C. ℓ_1-ℓ_2 Minimization

Stating our results for ℓ_1-ℓ_2 minimization requires additional notation. Namely, we use the following subsets of I^cJ:

$$K^{=} := \left\{ i \in I^cJ \,:\, |w_i| = \frac{1}{\beta} \right\} \tag{34}$$

$$K^{\neq} := \left\{ i \in I^cJ \,:\, |w_i| > \frac{1}{\beta} \right\}, \tag{35}$$

where we omit their dependency on β for notational simplicity. The most important parameter in our bounds will be

$$v_\beta := \sum_{i \in I} \left(1 + \beta \operatorname{sign}(x_i^\star)(x_i^\star - w_i)\right)^2 + \sum_{i \in K^{\neq}} (\beta |w_i| - 1)^2 , \tag{36}$$

which is the complete version of (6). We also define $\overline{w} := |w_k|$, where

$$k := \operatorname*{arg\,min}_{i \in I^cJ} \left| |w_i| - \frac{1}{\beta} \right| . \tag{37}$$

In words, $\overline{w}$ is the absolute value of the component of w whose absolute value is closest to $1/\beta$, in the set I^cJ. We will assume $I^cJ \neq \emptyset$, i.e., $s < q$ [cf. (22)], so that k and $\overline{w}$ are well defined.

As in ℓ_1-ℓ_1 minimization, we state our results in two forms: sharp but uninformative bounds, and not so sharp but informative bounds. We start with the latter.

Theorem 14 (ℓ_1-ℓ_2 minimization)**.** *Let $x^\star \in \mathbb{R}^n$ be the vector to reconstruct, $w \in \mathbb{R}^n$ the prior information. Let $f_2(x) = \|x\|_1 + \frac{\beta}{2}\|x - w\|_2^2$ with $\beta > 0$, and assume $0 < s < q < n$. Also, assume that there exists $i \in I^c$ such that $|w_i| > 1/\beta$ or that there exists $i \in IJ$ such that $\beta \neq \operatorname{sign}(x_i^\star)/(w_i - x_i^\star)$.*

1) *If*

$$\frac{q-s}{n-q} \leq |1 - \beta\,\overline{w}| \exp\left(\left((\beta\,\overline{w})^2 - 2\beta\,\overline{w} \right) \log\left(\frac{n}{q} \right) \right), \tag{38}$$

then

$$w\big(T_{f_2}(x^\star)\big)^2 \leq 2v_\beta \log\left(\frac{n}{q}\right) + s + |K^{\neq}| + \frac{1}{2}|K^{=}| + \frac{4}{5}q\,. \tag{39}$$

2) *If*

$$\frac{q-s}{n-q} \geq |1 - \beta\,\overline{w}| \exp\left(4 \frac{(\beta\,\overline{w} - 2)\beta\,\overline{w}}{|1 - \beta\overline{w}|^2} \log\left(\frac{q}{s} \right) \right), \tag{40}$$

then

$$w\big(T_{f_2}(x^\star)\big)^2 \leq \frac{2v_\beta}{(1 - \beta\,\overline{w})^2} \log\left(\frac{q}{s}\right) + |K^{\neq}| + \frac{1}{2}|K^{=}| + \frac{9}{5}s\,. \tag{41}$$

Similarly to Theorem 12 and Proposition 6, this theorem upper bounds $w(T_{f_2}(x^\star))^2$ with expressions that depend on key problem parameters, namely n, q, s, β, v_β, $\overline{w}$, $|K^{\neq}|$, and $|K^{=}|$. Together with Theorem 4, it then provides a sufficient number of measurements that guarantee that (3) reconstructs $x^\star$ with high probability. The assumption $0 < s < q < n$ translates into the sets I, I^cJ, and I^cJ^c being nonempty. It will be shown in Lemma 19 that the remaining assumptions are equivalent to $0 \notin \partial f_2(x^\star)$ and, hence, that we can use Proposition 5. It is relatively easy to satisfy one of these assumptions, namely that there exists $i \in IJ$ such that $\beta \neq \operatorname{sign}(x_i^\star)/(w_i - x_i^\star)$; a sufficient condition is that there are at least two indices i, j in I such that $\operatorname{sign}(x_i^\star)/(w_i - x_i^\star) \neq \operatorname{sign}(x_j^\star)/(w_j - x_j^\star)$. The alternative is to set $\beta > 1/|w_i|$ for all $i \in I^c$. Setting large values for β, however, will not only make the bounds in the theorem very large, but also degrade the performance of ℓ_1-ℓ_2 minimization significantly, as we will see in the experimental results section. Note that if there exists $i \in IJ$ such that $\beta \neq \operatorname{sign}(x_i^\star)/(w_i - x_i^\star)$, the first term of v_β in (36) has at least one nonzero summand; if, on the other hand, there exists $i \in I^c$ such that $|w_i| > 1/\beta$, the second term of v_β has a nonzero summand. We thus conclude that these assumptions are equivalent to $v_\beta > 0$.

The theorem is divided into two cases: 1) if condition (38) is satisfied, the bound in (39) holds; 2) if condition (40) is satisfied, the bound in (41) holds. As in ℓ_1-ℓ_1 minimization, the conditions (38) and (40) are neither necessarily disjointed nor are they guaranteed to cover all the possible values of β (although such a case is rare in practice). Case 1) is the most interesting in practice, since condition (38) holds when n is large with respect to q. In that case, the bound in (39) is mostly determined by the dominant term $2v_\beta \log(n/q)$. We then see that the role played by the number of bad components $\overline{h}$ in ℓ_1-ℓ_1 minimization is now played by v_β in ℓ_1-ℓ_2 minimization. Curiously, the first term of v_β captures the notion of good and bad components: consider $x_i^\star > 0$; clearly, a bad component $w_i < x_i^\star$ yields a larger v_β than a good component $w_i > x_i^\star$ does. The same happens for $x_i^\star < 0$. Finally, note that v_β is the only term in (39) that depends on β. Therefore, that bound is minimized when v_β is minimized, which occurs for

$$\beta^\star = \frac{\|w_{K^{\neq}}\|_1 + 1^\top (x^\star_{I_-} - w_{I_-}) - 1^\top (x^\star_{I_+} - w_{I_+})}{\|x_I^\star - w_I\|_2^2 + \|w_{K^{\neq}}\|_2^2}\,, \tag{42}$$

where z_S denotes the subvector of z whose components are indexed by the set S, and 1 denotes the vector of ones with appropriate dimensions. Selecting β as in (42) leads to

$$v_{\beta^\star} = s + |K^{\neq}| - \frac{\left[1^\top (x^\star_{I_+} - w_{I_+}) - 1^\top (x^\star_{I_-} - w_{I_-}) - \|w_{K^{\neq}}\|_1\right]^2}{\|x_I^\star - w_I\|_2^2 + \|w_{k^{\neq}}\|_2^2}\,. \tag{43}$$

The numerator of the last term of (43) can be written as the square of the inner product $z^\top \left[(x_I^\star - w_I^\star)^\top \quad w_{K^{\neq}}^\top \right]$, where $z_i = 1$ for $i \in I_+$, $z_i = -1$ for $i \in I_-$, and $z_i = -\operatorname{sign}(w_i)$ for $i \in K^{\neq}$. That is, all entries of z are ± 1, and thus $\|z\|_2^2 = s + |K^{\neq}|$. Applying the Cauchy-Schwarz inequality to the last term of (43), we obtain $v_{\beta^\star} \geq 0$. Although this is a trivial identity [see (36)], the conditions under which

it is achieved reveal the type of "good" prior information w for ℓ_1-ℓ_2 minimization. Concretely, the Cauchy-Schwarz inequality becomes an equality when z is a multiple of $\begin{bmatrix}(x^\star_I - w^\star_I)^\top & w^\top_{K^\neq}\end{bmatrix}$ which, in our case, translates into

$$\begin{cases} w_i = x^\star_i + c & , \ i \in I_+ \\ w_i = x^\star_i - c & , \ i \in I_- \\ |w_i| = c & , \ i \in K^\neq , \end{cases} \tag{44}$$

for some positive constant c (positivity is imposed by the last condition). As we had seen before, Theorem 14 does not hold for such a w: at least one of the conditions in (44) must not hold. Yet, asymptotically, the more conditions hold in (44), the better the performance of ℓ_1-ℓ_2 minimization or, in other words, the better the prior information w. To establish a parallel with ℓ_1-ℓ_1 minimization, note that h is the number of components that satisfy the first two conditions of (44) without the requirement that c is the same in all equations. In other words, components considered good for ℓ_1-ℓ_1 minimization (i.e., contributing to h) may not be "good" for ℓ_1-ℓ_2 minimization, since they may not satisfy one of the first two equations of (44). This shows that conditions for "good" prior information are much easier to satisfy in ℓ_1-ℓ_1 minimization than in ℓ_1-ℓ_2 minimization.

Regarding case 2) of Theorem 14, it holds when n is comparable to q, and β is close to $1/\overline{w}$. The bound in that case depends on β through the term $v_\beta/(1-\beta\overline{w})^2$. Although it can be minimized in closed-form, its expression is significantly more complicated than (42).

A sharper bound. Now we present a bound for ℓ_1-ℓ_2 minimization that is sharper than the ones in Theorem 14. Recall the notation $\mathcal{I}(a,b) := [a-b, a+b]$, for $b \geq 0$.

Theorem 15 (ℓ_1-ℓ_2 minimization: sharper bound)**.** *Let $x^\star, w \in \mathbb{R}^n$ be as in Theorem 14. Assume $s := |I| > 0$ and also that there exists $i \in I^c$ such that $|w_i| > 1/\beta$ or that there exists $i \in IJ$ such that $\beta \neq \operatorname{sign}(x^\star_i)/(w_i - x^\star_i)$. Then,*

$$w(T_{f_2}(x^\star))^2 \leq s + \min_{t\geq 0} \left\{ t^2 \sum_{i\in I} (1+\beta \operatorname{sign}(x^\star_i)(x^\star_i - w_i))^2 + \sum_{i\in I^c} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i, \mathcal{I}(-t\beta w_i, t)\Big)^2\Big] \right\}. \tag{45}$$

The proof is in Appendix VI-E. As in the sharper bound for the ℓ_1-ℓ_1 case (Theorem 13), computing (45) requires a numerical procedure. Also, because $f_2(x) = \|x\|_1 + (\beta/2)\|x - w\|_2^2$ is not a norm, the techniques used in [60, Theorem 4.3] and [36, Proposition 2] to prove sharpness of Gaussian width-type of bounds cannot be used in our case. In Section V, we will see that (45) precisely describes the performance of ℓ_1-ℓ_2 minimization.

D. ℓ_1-ℓ_1 minimization versus ℓ_1-ℓ_2 minimization

The bounds in Theorem 12 for ℓ_1-ℓ_1 minimization are minimized when $\beta = 1$, a value that leads to excellent results in practice, as we will see in Section V. In that case and for large n, ℓ_1-ℓ_1 minimization requires $O(2\overline{h}\log n)$ measurements for perfect recovery. Theorem 14, in turn, establishes that ℓ_1-ℓ_2 minimization requires $O(2v_\beta \log n)$ measurements. The optimal value of β in this case depends on $x^\star$ and w. This section starts by analyzing how the dominant factors $\overline{h}$ and v_β compare under additive modeling noise.[7] Then, it establishes a (deterministic) sufficient condition under which the sharp bound for ℓ_1-ℓ_1 minimization in (33) is smaller than the sharp bound for ℓ_1-ℓ_2 minimization in (45).

Dominant factors under additive modeling noise. We consider $w = x^\star + \gamma$, where $\gamma \in \mathbb{R}^n$ is modeling noise. For simplicity, assume γ and $x^\star$ have the same support and each entry of γ is drawn i.i.d. from a distribution symmetric around the origin with finite expected value (which is 0, due to the symmetry). The objective functions of (2) and (3) may lead us to think that ℓ_1-ℓ_1 minimization (2) performs better for a Laplacian γ and ℓ_1-ℓ_2 minimization (3) performs better for a Gaussian γ. This intuition is actually wrong in terms of the dominant parameters $\overline{h}$ and v_β: ℓ_1-ℓ_1 minimization performs better in both cases; in fact, it performs better for any distribution symmetric around the origin. To see why, note that γ having the same support as $x^\star$ implies $I^c J = \emptyset$, and thus $K^\neq = \emptyset$. Denote the variance of the entries of γ with σ. According to our model, both $\overline{h}$ and v_β are random variables. For example, $\overline{h}$ can be written as $\overline{h} = \sum_{i\in I_+} Z_i^- + \sum_{i\in I_-} Z_i^+$, where Z_i^- (resp. Z_i^+) is the indicator of the event "$\gamma_i < 0$" (resp. "$\gamma_i > 0$"). We have $\mathbb{E}[Z_i^-] = \mathbb{P}(\gamma_i < 0) = 1/2$ and $\mathbb{E}[Z_i^+] = \mathbb{P}(\gamma_i > 0) = 1/2$, due to the symmetry of the distribution of γ. The expected values of $\overline{h}$ and v_β are then

$$\mathbb{E}\Big[\overline{h}\Big] = \sum_{i\in I_+} \mathbb{E}[Z_i^-] + \sum_{i\in I_-} \mathbb{E}[Z_i^+] = \frac{s}{2} \tag{46}$$

$$\begin{aligned} \mathbb{E}[v_\beta] &= \sum_{i\in I_+} \mathbb{E}[(1-\beta\,\gamma_i)^2] + \sum_{i\in I_-} \mathbb{E}[(1+\beta\,\gamma_i)^2] \\ &= s(1+\beta^2\sigma^2)\,, \end{aligned} \tag{47}$$

where, in the last step, we used $\mathbb{E}[\gamma_i] = 0$ and $\mathbb{E}[\gamma_i^2] = \sigma^2$, for all i. Under this model, the assumptions of Theorems 12 and 14 [in cases 1)] hold with probability 1.[8] Due to concentration of measure [61], the larger the support I, the more $\overline{h}$ and v_β concentrate around their expected values in (46) and (47). This shows that, under the above model, ℓ_1-ℓ_1 minimization requires about half of the number of measurements than classical CS, whereas ℓ_1-ℓ_2 minimization actually requires more measurements, by a factor of $\beta^2\sigma^2$.

Comparing sharp bounds. We now establish a sufficient condition for the ℓ_1-ℓ_1 sharp bound (33) with $\beta = 1$ being smaller than the ℓ_1-ℓ_2 sharp bound (45) for any $\beta > 0$.

Corollary 16. *Let $x^\star \in \mathbb{R}^n$ be the signal to reconstruct, $w \in \mathbb{R}^n$ the prior information. Assume $\overline{h} > 0$ and $IJ^c = \emptyset$, i.e., $x^\star_i \neq 0$, $w_i \neq 0 \Rightarrow x^\star_i \neq w_i$. Consider ℓ_1-ℓ_1 minimization with $\beta_1 = 1$ and ℓ_1-ℓ_2 minimization with arbitrary $\beta_2 > 0$. If*

$$|w_i| \geq \frac{1}{\beta_2}\,, \qquad \text{for all } i \in I^c J \tag{48a}$$

[7] A more complete analysis for the case of a Laplacian distribution can be found in [41], which analyzes the ℓ_1-ℓ_1 minimization bound (4), and not just its dominant parameter $\overline{h}$.

[8] In reality, the assumption $s < q$ of Theorem 14 does not hold. An inspection of its proof, however, reveals that the role of $s < q$ is just to make $\overline{w}$ well defined. The proof still holds for case 1) if k in (37) is undefined and $\overline{w}$ is set to $+\infty$.

$$|x_i^\star - w_i| \geq \frac{1}{\beta_2}\,, \qquad \textit{for all } i \in I_+J_+ \cup I_-J_-\,, \tag{48b}$$

where (48b) *holds strictly for at least one component, then the right-hand side of* (33) *is always smaller than the right-hand side of* (45)*, i.e., the sharp bound for ℓ_1-ℓ_1 minimization is smaller than the one for ℓ_1-ℓ_2 minimization.*

The proof, in Section VI-F, shows that each term in (33) is smaller than the respective term in (45). This indicates that the assumptions of the corollary are strong and that its conclusions also hold under weaker assumptions.

E. Practical guidelines: improving the prior information

Our results indicate that $\overline{h}$ and v_β are the key parameters in ℓ_1-ℓ_1 and ℓ_1-ℓ_2 minimization, respectively. We now describe how to decrease them by modifying the prior information w.

One way to reduce $\overline{h}$ we find extremely effective in practice is to amplify w by a moderate factor. To give a concrete example, consider $x^\star = (-3, 2, 2, 4, -1)$ and the baseline prior information $w^b = (-2, 3, 1, -1, 0)$. Take $w = c \cdot w^b$, for some $c \geq 1$. If $c = 1$, w has one good component (the 2nd one) and four bad components (the remaining ones), i.e, $h = 1$, $\overline{h} = 4$. When $3/2 < c < 2$, the first component becomes good, i.e., $h = 2$, $\overline{h} = 3$. When $c > 2$, the third component also becomes good, yielding $h = 3$ and $\overline{h} = 2$. So, by amplifying the components of the baseline w^b we reduced its number of bad components to half their initial value. This is the maximum reduction we can get in this example because of the fourth and fifth components: the signs of $w_4 = -1$ and $x_4^\star = 4$ are different so, no matter how large c is, w_4 is always a bad component; similarly, $w_5 = 0$ remains unchanged under multiplication. If no such components exist, i.e., if $\text{sign}(x_i^\star) = \text{sign}(w_i)$ for all $x_i^\star \neq 0$, there is a c above which w has no bad components. In that case, Theorem 12 no longer applies and the number of measurements might actually increase. This is why we recommend a moderate value for c, e.g., 1.3, which should be tuned according to the application.

Applying the same technique to ℓ_1-ℓ_2 minimization does not work as well. Recall that improving the quality of w in this case means satisfying as many conditions in (44) as possible. The previous technique then does not work if the magnitudes of $x_i^\star$ have large variability. So, instead of amplifying w, we recommend adding a small quantity, say c, to the positive components of w and subtracting c from the negative components, i.e., $w_i = w_i^b + c$ if $w_i^b > 0$, and $w_i = w_i^b - c$ if $w_i^b < 0$ [notice the similarity with the first two equations of (44)]. In vector form, $w = w^b + c \cdot \text{sign}(w^b)$, where $\text{sign}(\cdot)$ is applied component-wise. Since a w that satisfies (44) yields $\beta^\star = 1/c$ in (42), we also recommend setting $\beta = 1/c$ in this case. Finally, we note that this technique works for ℓ_1-ℓ_1 minimization as well; of course, we recommend using $\beta = 1$ in that case.

V. Experimental Results

We describe two types of experiments: one that assesses the sharpness of our bounds for a wide range of β's, and another that illustrates the effectiveness of our practical guidelines for improving the prior information.

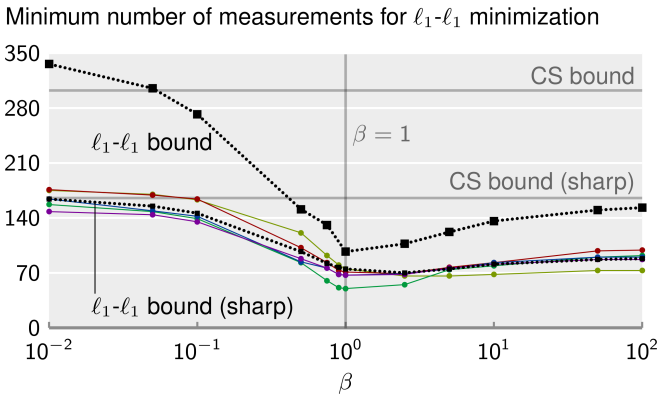


Figure 6. Experimental performance of ℓ_1-ℓ_1 minimization for 5 different Gaussian matrices as a function of β (solid lines). The upper dotted line depicts the bounds of Theorem 12, which are minimized for $\beta = 1$ (vertical line). Horizontal lines: bound in (59) for CS and its sharp version in [60].

A. Sharpness of the bounds

Experimental setup. The data was generated as in Fig. 1, but for smaller dimensions. Namely, $x^\star$ had $n = 500$ entries, $s = 50$ of which were nonzero. The values of these entries were drawn from a zero-mean Gaussian distribution with unit variance. We then generated the prior information as $w = x^\star + z$, where z was 20-sparse and whose support coincided with the one of $x^\star$ in 16 entries and differed in 4 of them. The nonzero entries were zero-mean Gaussian with standard deviation 0.8. This yielded $h = 6$, $\overline{h} = 11$, $q = 53$, and $l = 20$.

The experiments were conducted as follows. We created a square matrix $\overline{A} \in \mathbb{R}^{500\times 500}$ with entries drawn independently from the standard Gaussian distribution. We then set $\overline{y} = \overline{A}x^\star$. Next, for a fixed β, we solved problem (2), first by using only the first row of $\overline{A}$ and the first entry of $\overline{y}$. If the solution of (2), say $\hat{x}_1(\beta)$, did not satisfy $\|\hat{x}_1(\beta) - x^\star\|_2/\|x^\star\|_2 \leq 10^{-2}$, we proceeded by solving (2) with the first two rows of $\overline{A}$ and the first two entries of $\overline{y}$. This procedure was repeated until $\|\hat{x}_m(\beta) - x^\star\|_2/\|x^\star\|_2 \leq 10^{-2}$, where $\hat{x}_m(\beta)$ denotes the solution of (2) when A (resp. y) consists of the first m rows (resp. entries) of $\overline{A}$ (resp. $\overline{y}$). In other words, we stopped when we found the minimum number of measurements that ℓ_1-ℓ_1 minimization requires for successful reconstruction, that is, $\min\{m \,:\, \|\hat{x}_m(\beta) - x^\star\|_2/\|x^\star\|_2 \leq 10^{-2}\}$. The values of β varied between 0.01 and 100. We then repeated the entire procedure for 4 other randomly generated pairs $(\overline{A}, \overline{y})$.

Results for ℓ_1-ℓ_1 minimization. Fig. 6 shows the results of these experiments. It displays the minimum number of measurements for successful reconstruction versus β. The 5 solid lines give the experimental performance of (2) for the 5 different pairs $(\overline{A}, \overline{y})$. The upper dotted line shows the bounds given by Theorem 12. When $\beta \neq 1$, the subcases of cases 2) and 3) of that theorem may give two different bounds, from which we select the smallest one. For reference, we use a vertical line to mark the value that minimizes the bounds in Theorem 12: $\beta = 1$. Another dotted line shows the sharper bound in Theorem 13, computed numerically: it coincides with the experimental curves. Two horizontal lines depict values of two bounds for classical CS: the upper one the simple bound

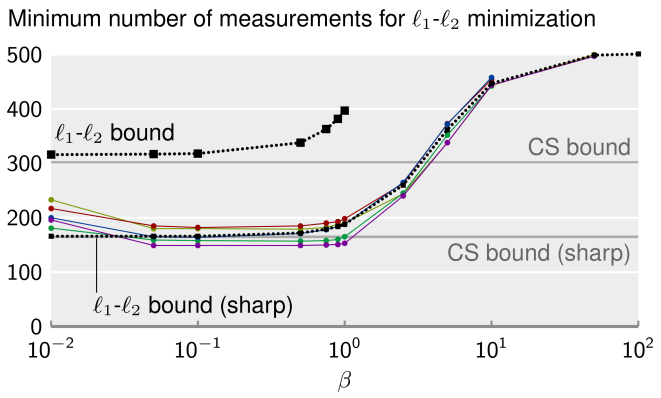


Figure 7. Same as Fig. 6, but for ℓ_1-ℓ_2 minimization. The data is the same as in Fig. 6, but the vertical scales are different. For $\beta > 1$, the bounds given by Theorem 14 are larger than 500 and, hence, are not shown.

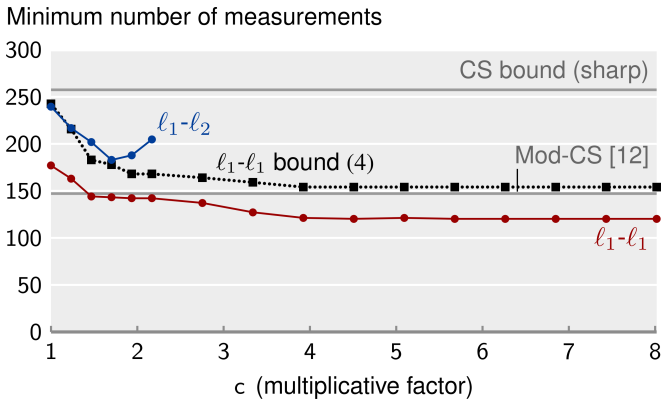


Figure 8. Prior information improvement with a multiplicative factor: $w = c \cdot w^b$, where w^b is the baseline prior information. The vertical axis shows the minimum number of measurements to achieve 1% error. The horizontal lines show the CS bound in [60] and the performance of Mod-CS [12].

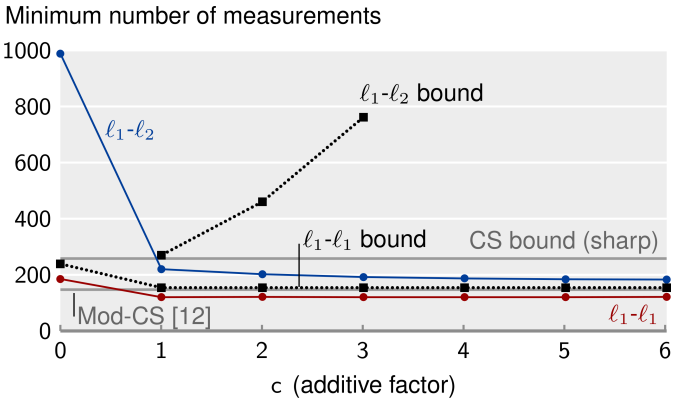


Figure 9. Same as Fig. 8, but for an additive factor. The data is the same, but the prior information is generated as $w = w^b + c \cdot \text{sign}(w^b)$, for $c \geq 0$.

in (19), the lower one the sharper bound in [60], computed numerically. We point out that we removed the point of the ℓ_1-ℓ_1 bound corresponding to $\beta = 0.9$, since it was 576, a value larger than the signal dimensionality, 500. As we had seen before, values of β close to 1 in cases 2.b) and 3.b) yield large bounds in Theorem 12. We had also stated that the bound for $\beta = 1$ is not only the sharpest one in that theorem, but also the smallest one. Fig. 6 also shows that setting β to 1 leads to a performance in practice close to the optimal one. Indeed, three out of the five solid curves in the figure achieved their minimum at $\beta = 1$; the remaining ones achieved it at $\beta = 2.5$. We also observe that the bound for $\beta = 1$ is quite sharp: its value is 97, and the maximum among all of the solid lines for $\beta = 1$ was 75 measurements. The figure also shows that the bounds are looser for $\beta < 1$ and, eventually, become larger than the bound in (19). For $\beta > 1$, the bound is relatively sharp. Regarding the experimental performance of ℓ_1-ℓ_1 minimization, it degrades for small β, towards standard CS, achieving top performance around $\beta = 1$. For large β, the performance also degrades, however, without becoming worse than for small β.

Results for ℓ_1-ℓ_2 minimization. Fig. 7 shows the same experiments, with the same data, for ℓ_1-ℓ_2 minimization. The scale of the vertical axis is different from the one in Fig. 6. We do not show the bounds for $\beta > 1$, because they were larger than 500 (e.g., the bound for $\beta = 2.5$ was 820). The minimum value of the bound was 315 ($\beta = 0.01$), which is slightly larger than the bound for standard CS in (19). In fact, for this example, the bounds given by Theorem 14 were always larger than the one for standard CS; as we will see in the next set of experiments, ℓ_1-ℓ_2 minimization can generally outperform standard CS if we improve the prior information as recommended in Section IV-E. The experimental performance curves in Fig. 7 behaved differently from the ones for ℓ_1-ℓ_1 minimization: from $\beta = 0.01$ to $\beta = 0.05$, they decreased slightly and remained approximately constant until $\beta = 1$. After that point, their performance degraded sharply. For instance, for $\beta = 50$, (3) was able to reconstruct $x^\star$ for one pair $(\overline{A}, \overline{y})$ only; and this required using the full matrix $\overline{A}$. In conclusion, although prior information helped (slightly) for β between 0.01 and 1, the bounds of Theorem 14 were not sharp.

B. *Improving the prior information*

These experiments illustrate the gains obtained by following the guidelines of Section IV-E on how to improve prior information.

Experimental setup. The vector $x^\star$ was generated exactly as before, with $n = 1000$ and $s = 70$. To better illustrate the gains, the prior information was generated differently: the base prior information was created as $w^b = x^\star + z$ with a 104-sparse z, whose support coincided with the one of $x^\star$ in 56 entries and differed in 49. The nonzero components of z were zero-mean Gaussian with variance 0.3. This yielded $h = 32$, $\overline{h} = 25$, $q = 117$, and $l = 104$.

In these experiments, we modified w^b as in Section IV-E: by a multiplicative factor $w = c \cdot w^b$, with c varying between 1 and 7, and by an additive factor $w = w^b + c \cdot \text{sign}(w^b)$, with c varying between 0.01 and 20. For ℓ_1-ℓ_1 minimization, we set always $\beta = 1$. For ℓ_1-ℓ_2 minimization, we set $\beta = 1$ in the multiplicative factor case and $\beta = 1/c$ in the additive case. In contrast with the previous experiments, we generated just one pair $(\overline{A}, \overline{y})$, where $\overline{y} = \overline{A}x^\star$. The experiments consisted of computing the minimum number of rows of $\overline{A}$ that guaranteed a relative error smaller than 1%, for different values of c.

Results for a multiplicative factor. Fig. 8 shows the results for a multiplicative factor improvement. The solid lines represent the experimental performance of ℓ_1-ℓ_1 and ℓ_1-ℓ_2 minimization, the dotted line the bound in (4) (the bound for ℓ_1-ℓ_2 was too large to be displayed), and the horizontal lines the classical CS (sharp) bound in [60] and the performance of Modified-CS (Mod-CS) [12]. Mod-CS integrates prior information in CS via an estimate of the support of $x^\star$; naturally, we used the support of w^b for such estimate. The plot shows that both the experimental performance of ℓ_1-ℓ_1 and its bound are nondecreasing with c, confirming the effectiveness of our strategy to improve the prior information. Both curves decrease monotonically until around $c = 3.5$, after which they reach a plateau: 121 for ℓ_1-ℓ_1 and 154 for the bound. Mod-CS requires 148 measurements to solve this particular problem, a value smaller than the number of measurements required by ℓ_1-ℓ_1 minimization for $c \leq 1.2$. For any $c > 1.2$, ℓ_1-ℓ_1 minimization required less measurements than Mod-CS. Regarding ℓ_1-ℓ_2 minimization, its performance improved for $1 < c \leq 2$: for example, it required 305 measurements for $c = 1.5$. For $c > 2$, its performance degraded quickly. In conclusion, as predicted in Section IV-E, improving the prior information via a multiplicative factor works well for ℓ_1-ℓ_1 minimization, but not as well for ℓ_1-ℓ_2 minimization.

Results for an additive factor. The results for an additive factor are shown in Fig. 9. The curves are the same as in Fig. 8, but now we show the bound for ℓ_1-ℓ_2 minimization. The bound was quite sharp for $c = 1$, but became loose for larger c. In spite of this, the performance of ℓ_1-ℓ_2 minimization improved with increasing c, outperforming classical CS for $c \geq 1$. This improvement was not enough to reach the 148 measurements required by Mod-CS. We can also see that both the performance of ℓ_1-ℓ_1 minimization and the respective bound (4) decreased with c and, in fact, reached the same plateaus as in Fig. 8.

These experiments confirm that improving side information with an additive factor works well for ℓ_1-ℓ_2 minimization, and improving it with both an additive and a multiplicative factor works well for ℓ_1-ℓ_1 minimization. They also show that, in general, ℓ_1-ℓ_1 minimization performs better than ℓ_1-ℓ_2 minimization and, if the prior information has enough quality, also better than state-of-the-art approaches like Mod-CS [12].

VI. Proof of Main Results

In this section, we present the proofs of all results from Section IV. We start with some auxiliary results.

A. Auxiliary Results

The following lemma plays an important role in our proofs. Its first part gives an exact expression for the expected squared distance of a scalar Gaussian random variable to an interval as a function of the Q-function, defined as

$$Q(x) := \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} \exp\Big(-\frac{u^2}{2}\Big)\, du = \int_x^{+\infty} \varphi(u)\, du\,, \quad (49)$$

where

$$\varphi(x) := \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big) \quad (50)$$

is the probability density function of a zero-mean, unit variance scalar Gaussian random variable. To obtain the closed-form bounds in Theorem 12, we will need to (upper) bound the exact expression. That is done in the second part of the lemma. We represent an interval in $\mathbb{R}$ as

$$\mathcal{I}(a,b) := \big\{x \in \mathbb{R} : |x-a| \leq b\big\} = \big[a-b,\, a+b\big]\,. \quad (51)$$

Lemma 17. *Let $g \sim \mathcal{N}(0,1)$ be a scalar, zero-mean Gaussian random variable with unit variance. Let $a, b \in \mathbb{R}$ and $b \geq 0$.*

Part I) Exact expression
There holds

$$\mathbb{E}_g\Big[\mathrm{dist}\big(g, \mathcal{I}(a,b)\big)^2\Big] = (a-b)\varphi(a-b) - (a+b)\varphi(a+b) + \big[1+(a+b)^2\big]Q(a+b) + \big[1+(a-b)^2\big]\big[1-Q(a-b)\big]\,. \quad (52)$$

Part II) Bounds

1) *If $b = 0$, then $\mathcal{I}(a,b) = \{a\}$ and*

$$\mathbb{E}_g\big[\mathrm{dist}(g,a)^2\big] = a^2 + 1\,. \quad (53)$$

2) *If $b > 0$ and $|a| < b$, i.e., $0 \in \mathcal{I}(a,b)$, then*

$$\mathbb{E}_g\Big[\mathrm{dist}\big(g, \mathcal{I}(a,b)\big)^2\Big] \leq \frac{\varphi(b-a)}{b-a} + \frac{\varphi(a+b)}{a+b}\,. \quad (54)$$

3) *If $b > 0$ and $a + b < 0$, then*

$$\mathbb{E}_g\Big[\mathrm{dist}\big(g, \mathcal{I}(a,b)\big)^2\Big] \leq 1 + (a+b)^2 + \frac{\varphi(b-a)}{b-a}\,. \quad (55)$$

4) *If $b > 0$ and $a - b > 0$, then*

$$\mathbb{E}_g\Big[\mathrm{dist}\big(g, \mathcal{I}(a,b)\big)^2\Big] \leq 1 + (a-b)^2 + \frac{\varphi(a+b)}{a+b}\,. \quad (56)$$

5) *If $b > 0$ and $a + b = 0$, then*

$$\mathbb{E}_g\Big[\mathrm{dist}\big(g, \mathcal{I}(a,b)\big)^2\Big] \leq \frac{\varphi(b-a)}{b-a} + \frac{1}{2}\,. \quad (57)$$

6) *If $b > 0$ and $a - b = 0$, then*

$$\mathbb{E}_g\Big[\mathrm{dist}\big(g, \mathcal{I}(a,b)\big)^2\Big] \leq \frac{\varphi(a+b)}{a+b} + \frac{1}{2}\,. \quad (58)$$

The proof can be found in Appendix A. In part II), each case considers a different relative position between the interval $\mathcal{I}(a,b)$ and zero, which is the mean of the random variable g. In case 1), the interval is simply a point. In case 2), $\mathcal{I}(a,b)$ contains zero. In cases 3) and 4), $\mathcal{I}(a,b)$ does not contain zero. And, finally, in cases 5) and 6), zero is one of the endpoints of $\mathcal{I}(a,b)$. Notice that addressing cases 5) and 6) separately from cases 4) and 5) leads to sharper bounds on the former: for example, making $a + b \to 0$ in the right-hand side of (55) gives $1+\varphi(b-a)/(b-a)$, which is larger than the right-hand side of (57). We note that the proof of Proposition 4 in [27] for standard CS uses the bound (54) with $a = 0$. The following result will be used frequently.

Lemma 18. *There holds*

$$\frac{1-\frac{1}{x}}{\sqrt{\pi \log x}} \leq \frac{1}{\sqrt{2\pi}} \leq \frac{2}{5}\,, \quad (59)$$

for all $x > 1$.

The proof can be found in Appendix B. Recall the definitions of functions f_1 and f_2:

$$f_1(x) := \|x\|_1 + \beta\|x - w\|_1 \tag{60}$$

$$f_2(x) := \|x\|_1 + \frac{\beta}{2}\|x - w\|_2^2 . \tag{61}$$

To apply Proposition 5 to these functions, i.e., to say that their normal cones at a given $x^\star$ is equal to the cone generated by their subdifferentials at $x^\star$, we need to guarantee that their subdifferentials do not contain the zero vector: $0 \notin \partial f_j(x^\star)$, $j = 1, 2$. The next two lemmas give a characterization of this condition in terms of the problem parameters in Definition 9. Before that, let us compute $\partial f_1(x^\star)$ and $\partial f_2(x^\star)$. A key property of functions f_1 and f_2, and on which our results deeply rely, is that they admit a component-wise decomposition:

$$f_1(x) = \sum_{i=1}^n f_1^{(i)}(x_i) \qquad f_2(x) = \sum_{i=1}^n f_2^{(i)}(x_i) ,$$

where $f_1^{(i)} = |x_i| + \beta|x_i - w_i|$ and $f_2^{(i)} = |x_i| + \frac{\beta}{2}(x_i - w_i)^2$. Therefore,

$$\partial f_1(x^\star) = \Big(\partial f_1^{(1)}(x_1^\star), \partial f_1^{(2)}(x_2^\star), \dots, \partial f_1^{(n)}(x_n^\star)\Big)$$
$$\partial f_2(x^\star) = \Big(\partial f_2^{(1)}(x_1^\star), \partial f_2^{(2)}(x_2^\star), \dots, \partial f_2^{(n)}(x_n^\star)\Big) .$$

Recall that $\partial|s| = \operatorname{sign}(s)$ for $s \neq 0$, and $\partial|s| = [-1, 1]$ for $s = 0$. The function $\operatorname{sign}(\cdot)$ returns the sign of a number, i.e., $\operatorname{sign}(a) = 1$ if $a > 0$, and $\operatorname{sign}(a) = -1$ if $a < 0$. We then have

$$\partial f_1^{(i)}(x_i^\star) = \begin{cases} \operatorname{sign}(x_i^\star) + \beta \operatorname{sign}(x_i^\star - w_i) & , \ i \in IJ \\ \operatorname{sign}(x_i^\star) + [-\beta, \beta] & , \ i \in IJ^c \\ \beta \operatorname{sign}(x_i^\star - w_i) + [-1, 1] & , \ i \in I^c J \\ \big[-\beta - 1, \beta + 1\big] & , \ i \in I^c J^c \end{cases} \tag{62}$$

and

$$\partial f_2^{(i)}(x_i^\star) = \begin{cases} \operatorname{sign}(x_i^\star) + \beta(x_i^\star - w_i) & , \ i \in I \\ \big[-1, 1\big] - \beta w_i & , \ i \in I^c , \end{cases} \tag{63}$$

for $i = 1, \dots, n$.

Lemma 19. *Assume $x^\star \neq 0$ or, equivalently, that $I \neq \emptyset$. Assume also $w \neq x^\star$ or, equivalently, that $J \neq \emptyset$. Consider f_1 and f_2 in* (60) *and* (61)*, respectively.*

1) $0 \notin \partial f_1(x^\star)$ *if and only if* $\overline{h} > 0$ *or* $\beta \neq 1$.
2) $0 \notin \partial f_2(x^\star)$ *if and only if there is* $i \in IJ$ *such that* $\beta \neq \operatorname{sign}(x_i^\star)/(w_i - x_i^\star)$ *or there is* $i \in I^c$ *such that* $\beta > 1/|w_i|$.

The proof is in Appendix C.

B. Proof of Theorem 12

Proposition 3 establishes that $w(C) = \mathbb{E}_g\big[\operatorname{dist}(g, C^o)\big]$, for a cone C and its polar cone C^o, where $g \sim \mathcal{N}(0, I)$. Using Jensen's inequality [59, Thm. B.1.1.8], $w(C)^2 \leq \mathbb{E}_g\big[\operatorname{dist}(g, C^o)^2\big]$. The polar cone of the tangent cone $T_{f_1}(x^\star)$ is the normal cone $N_{f_1}(x^\star)$ which, according to Proposition 5, coincides with the cone generated by the subdifferential $\partial f_1(x^\star)$ whenever $0 \notin \partial f_1(x^\star)$. In other words, if $0 \notin \partial f_1(x^\star)$, then

$$w\big(T_{f_1}(x^\star)\big)^2 \leq \mathbb{E}_g\Big[\operatorname{dist}\big(g, \operatorname{cone} \partial f_1(x^\star)\big)^2\Big] . \tag{64}$$

Part 1) of Lemma 19 establishes that $0 \notin \partial f_1(x^\star)$ is equivalent to $\beta \neq 1$ or $\overline{h} > 0$. So, provided we assume that $\overline{h} > 0$ for part 1) of the theorem, we can always use (64). The proof is organized as follows. First, we compute a generic upper bound on (64), using the several cases of Lemma 17. This will give us three bounds, each one for a specific case of the theorem, i.e., $\beta = 1$, $\beta < 1$, and $\beta > 1$. These bounds, however, will be uninformative since they depend on unknown quantities and on a free variable. We then address each case separately, selecting a specific value for the free variable and "getting rid" of the unknown quantities. In this last step, we will use the bound in Lemma 18 frequently.

1) Generic Bound: A vector $d \in \mathbb{R}^n$ belongs to the cone generated by $\partial f_1(x^\star)$ if $d = ty$ for some $t \geq 0$ and some $y \in \partial f_1(x^\star)$. According to (62), each component d_i satisfies

$$\begin{cases} d_i = t \operatorname{sign}(x_i^\star) + t\beta \operatorname{sign}(x_i^\star - w_i) & , \text{ if } i \in IJ \\ |d_i - t \operatorname{sign}(x_i^\star)| \leq t\beta & , \text{ if } i \in IJ^c \\ |d_i - t\beta \operatorname{sign}(x_i^\star - w_i)| \leq t & , \text{ if } i \in I^c J \\ |d_i| \leq t(\beta + 1) & , \text{ if } i \in I^c J^c , \end{cases}$$

for some $t \geq 0$. Thus, the right-hand side of (64) is written as

$$\begin{aligned} &\mathbb{E}_g\Big[\operatorname{dist}\big(g, \operatorname{cone} \partial f_1(x^\star)\big)^2\Big] \\ &= \mathbb{E}_g\Bigg[\min_{t \geq 0}\Bigg\{\sum_{i \in IJ} \operatorname{dist}\Big(g_i \, , \, t \operatorname{sign}(x_i^\star) + t\beta \operatorname{sign}(x_i^\star - w_i)\Big)^2 \\ &\qquad + \sum_{i \in IJ^c} \operatorname{dist}\Big(g_i \, , \, \mathcal{I}\big(t \operatorname{sign}(x_i^\star), t\beta\big)\Big)^2 \\ &\qquad + \sum_{i \in I^c J} \operatorname{dist}\Big(g_i \, , \, \mathcal{I}\big(-t\beta \operatorname{sign}(w_i), t\big)\Big)^2 \\ &\qquad + \sum_{i \in I^c J^c} \operatorname{dist}\Big(g_i \, , \, \mathcal{I}\big(0, t(\beta + 1)\big)\Big)^2 \Bigg\}\Bigg] , \end{aligned}$$

where, in the third term, we used $\operatorname{sign}(x_i^\star - w_i) = -\operatorname{sign}(w_i)$, since $x_i^\star = 0$ for $i \in I^c J$. As in the proof of Proposition 6 (in [27]), we fix t now and select a particular value for it later. Our choice for t will not necessarily be optimal, but it will give bounds that can be expressed as a function of the parameters in Definition 9. In other words, if h is a function of t and g, we have

$$\mathbb{E}_g\Big[\min_t \, h(g, t)\Big] \leq \min_t \, \mathbb{E}_g\big[h(g, t)\big] \leq \mathbb{E}_g\big[h(g, t)\big] \, , \forall_t \, . \tag{65}$$

The value we will select for t does not necessarily minimize the second term in (65), but allows deriving useful bounds. For a fixed t, we then have:

$$\mathbb{E}_g\Big[\operatorname{dist}\big(g, \operatorname{cone} \partial f_1(x^\star)\big)^2\Big]$$

$$\leq \sum_{i\in IJ} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\, t\operatorname{sign}(x_i^\star) + t\beta\operatorname{sign}(x_i^\star - w_i)\Big)^2\Big] \tag{66a}$$

$$+ \sum_{i\in IJ^c} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\, \mathcal{I}\big(t\operatorname{sign}(x_i^\star), t\beta\big)\Big)^2\Big] \tag{66b}$$

$$+ \sum_{i\in I^cJ} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\, \mathcal{I}\big(-t\beta\operatorname{sign}(w_i), t\big)\Big)^2\Big] \tag{66c}$$

$$+ \sum_{i\in I^cJ^c} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\, \mathcal{I}\big(0, t(\beta+1)\big)\Big)^2\Big]. \tag{66d}$$

Next, we use Lemma 17 to compute (66a) in closed-form and to upper bound (66b), (66c), and (66d).

Expression for (66a)**.** By partitioning the set IJ into $I_+J_+ \cup I_-J_- \cup I_-J_+ \cup I_+J_-$, we obtain

$$\begin{aligned}(66a) = &\sum_{i\in I_+J_+} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\, t(\beta+1)\Big)^2\Big] \\ &+ \sum_{i\in I_-J_-} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\, -t(\beta+1)\Big)^2\Big] \\ &+ \sum_{i\in I_-J_+} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\, t(\beta-1)\Big)^2\Big] \\ &+ \sum_{i\in I_+J_-} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\, t(1-\beta)\Big)^2\Big].\end{aligned}$$

And using (53) in Lemma 17 and h and $\overline{h}$ in Definition 9,

$$\begin{aligned}(66a) &= \sum_{i\in I_+J_+}\Big[t^2(\beta+1)^2+1\Big] + \sum_{i\in I_-J_-}\Big[t^2(\beta+1)^2+1\Big] \\ &\quad + \sum_{i\in I_-J_+}\Big[t^2(\beta-1)^2+1\Big] + \sum_{i\in I_+J_-}\Big[t^2(\beta-1)^2+1\Big] \\ &= \big|IJ\big| + \Big(\big|I_+J_+\big| + \big|I_-J_-\big|\Big)t^2(\beta+1)^2 \\ &\quad + \Big(\big|I_-J_+\big| + \big|I_+J_-\big|\Big)t^2(\beta-1)^2 \\ &= t^2\Big(\overline{h}(\beta+1)^2 + h(\beta-1)^2\Big) + \big|IJ\big|. \end{aligned} \tag{67}$$

Note that h and $\overline{h}$ appear here naturally, before selecting any t.

Bounding (66b)**.** If we decompose $IJ^c = I_+J^c \cup I_-J^c$, we see that

$$\begin{aligned}(66b) = &\sum_{i\in I_+J^c} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\, \mathcal{I}(t, t\beta)\Big)^2\Big] \\ &+ \sum_{i\in I_-J^c} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\, \mathcal{I}(-t, t\beta)\Big)^2\Big].\end{aligned} \tag{68}$$

There are three cases: $\beta = 1$, $\beta < 1$, and $\beta > 1$.

- If $\beta = 1$, then $\mathcal{I}(t, t\beta) = [0, 2t]$ and $\mathcal{I}(-t, t\beta) = [-2t, 0]$. Applying (58) (resp. (57)) to each summand in the first (resp. second) term of (68) we conclude that
$$(66b) \leq \big|IJ^c\big|\left[\frac{\varphi(2t)}{2t} + \frac{1}{2}\right]. \tag{69}$$
- If $\beta < 1$, then $0 \notin \mathcal{I}(t, t\beta)$ and $0 \notin \mathcal{I}(-t, t\beta)$. If we apply (56) to the summands in the first term of (68) and (55) to the summands in the second term, and take into account that $|I_+J^c| + |I_-J^c| = |IJ^c|$,
$$(66b) \leq \big|IJ^c\big|\left[1 + t^2(1-\beta)^2 + \frac{\varphi(t(\beta+1))}{t(\beta+1)}\right]. \tag{70}$$
- Finally, if $\beta > 1$, then $0 \in \mathcal{I}(t, t\beta)$ and $0 \in \mathcal{I}(-t, t\beta)$. Applying (54) to each summand in both terms of (68) we conclude
$$(66b) \leq \big|IJ^c\big|\left[\frac{\varphi(t(\beta-1))}{t(\beta-1)} + \frac{\varphi(t(\beta+1))}{t(\beta+1)}\right]. \tag{71}$$

Bounding (66c)**.** Decompose $I^cJ = I^cJ_+ \cup I^cJ_-$ and write

$$\begin{aligned}(66c) = &\sum_{i\in I^cJ_+} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\, \mathcal{I}\big(t\beta, t\big)\Big)^2\Big] \\ &+ \sum_{i\in I^cJ_-} \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\, \mathcal{I}\big(-t\beta\,, t\big)\Big)^2\Big].\end{aligned} \tag{72}$$

As before, we have three cases: $\beta = 1$, $\beta < 1$, and $\beta > 1$.

- If $\beta = 1$, then $\mathcal{I}(t\beta, t) = [0, 2t]$ and $\mathcal{I}(-t\beta, t) = [-2t, 0]$. If we apply (58) (resp. (57)) to each summand in the first (resp. second) term of (72), we conclude
$$(66c) \leq \big|I^cJ\big|\left[\frac{\varphi(2t)}{2t} + \frac{1}{2}\right]. \tag{73}$$
- If $\beta < 1$, then $0 \in \mathcal{I}(t\beta, t)$ and $0 \in \mathcal{I}(-t\beta, t)$. Therefore, according to (54),
$$(66c) \leq \big|I^cJ\big|\left[\frac{\varphi(t(1+\beta))}{t(1+\beta)} + \frac{\varphi(t(1-\beta))}{t(1-\beta)}\right]. \tag{74}$$
- If $\beta > 1$, then $0 \notin \mathcal{I}(t\beta, t)$ and $0 \notin \mathcal{I}(-t\beta, t)$. If we apply (56) to each summand in the first term of (72) and (55) to each summand in the second term, we find
$$(66c) \leq \big|I^cJ\big|\left[1 + t^2(\beta-1)^2 + \frac{\varphi(t(\beta+1))}{t(\beta+1)}\right]. \tag{75}$$

Bounding (66d)**.** The interval $\mathcal{I}(0, t(\beta+1))$ contains the origin, so we can apply (54) directly to each summand in (66d):

$$(66d) \leq 2\big|I^cJ^c\big|\frac{\varphi(t(\beta+1))}{t(\beta+1)}. \tag{76}$$

Bounding (66a) **+** (66b) **+** (66c) **+** (66d)**.** Given all the previous bounds, we can now obtain a generic bound for (64). Naturally, there are three cases: $\beta = 1$, $\beta < 1$, and $\beta > 1$.

- For $\beta = 1$, we sum (67) (with $\beta = 1$), (69), (73), and (76) (with $\beta = 1$):
$$\begin{aligned}&\mathbb{E}_g\Big[\operatorname{dist}\big(g, \operatorname{cone}\partial f_1(x^\star)\big)^2\Big] \leq 4\overline{h}t^2 + \big|IJ\big| \\ &+ \frac{1}{2}\big[|IJ^c| + |I^cJ|\big] + \big[|IJ^c| + |I^cJ| + 2|I^cJ^c|\big]\frac{\varphi(2t)}{2t}.\end{aligned} \tag{77}$$
- For $\beta < 1$, we sum (67), (70), (74), and (76):
$$\begin{aligned}&\mathbb{E}_g\Big[\operatorname{dist}\big(g, \operatorname{cone}\partial f_1(x^\star)\big)^2\Big] \leq t^2\Big[\overline{h}(\beta+1)^2 \\ &\quad + (h + |IJ^c|)(\beta-1)^2\Big] + |I| + |I^cJ|\frac{\varphi(t(1-\beta))}{t(1-\beta)}\end{aligned}$$

$$+\Big[|IJ^c|+|I^cJ|+2|I^cJ^c|\Big]\frac{\varphi(t(\beta+1))}{t(\beta+1)}\,. \quad (78)$$

- For $\beta>1$, we sum (67), (71), (75), and (76):

$$\mathbb{E}_g\Big[\mathrm{dist}\big(g,\mathrm{cone}\,\partial f_1(x^\star)\big)^2\Big] \le t^2\Big[\overline{h}(\beta+1)^2 + (h+|I^cJ|)(\beta-1)^2\Big] + |J| + |IJ^c|\frac{\varphi(t(\beta-1))}{t(\beta-1)} + \big[|IJ^c|+|I^cJ|+2|I^cJ^c|\big]\frac{\varphi(t(\beta+1))}{t(\beta+1)}\,. \quad (79)$$

2) Specification of the Bound for Each Case: We now address each one of the cases $\beta=1$, $\beta<1$, and $\beta>1$ individually. Before that, recall from (25) that $|I^cJ|+|IJ^c|+2|I^cJ^c|=2n-(q+h+\overline{h})$, a term that appears in (77), (78), and (79). That term is always positive due to our assumption that $n-q=|I^cJ^c|>0$.

Case 1: $\boldsymbol{\beta=1}$. Notice that, according to (20) and (24),

$$\big|IJ\big|+\frac{1}{2}\big[|IJ^c|+|I^cJ|\big] = h+\overline{h}+\frac{1}{2}q-\frac{1}{2}(h+\overline{h}) = \frac{1}{2}(q+h+\overline{h})\,.$$

This allows rewriting (77) as

$$\mathbb{E}_g\Big[\mathrm{dist}\big(g,\mathrm{cone}\,\partial f_1(x^\star)\big)^2\Big] \le 4\overline{h}t^2+\frac{1}{2}(q+h+\overline{h}) + \frac{1}{2}\Big[2n-(q+h+\overline{h})\Big]\frac{1}{t\sqrt{2\pi}}\exp(-2t^2)\,, \quad (80)$$

where we used the definition of φ in (50). We now select t as

$$t^\star := \sqrt{\frac{1}{2}\log\Big(\frac{2n}{q+h+\overline{h}}\Big)} = \sqrt{\frac{1}{2}\log r}\,,$$

where $r:=2n/(q+h+\overline{h})$. Notice that $t^\star$ is well defined because $2n>q+h+\overline{h}$, i.e., $r>0$. It is also finite, as our assumption that $x^\star\neq 0$, or $|I|>0$, implies $q=|I\cup J|>0$. Replacing $t^\star$ in (80), we obtain

$$\begin{aligned}
&\mathbb{E}_g\Big[\mathrm{dist}\big(g,\mathrm{cone}\,\partial f_1(x^\star)\big)^2\Big]\\
&\le 2\overline{h}\log r+\frac{1}{2}(q+h+\overline{h})\\
&\quad+\frac{1}{2}\Big[2n-(q+h+\overline{h})\Big]\frac{1}{\sqrt{\pi\log r}}\frac{1}{\frac{2n}{q+h+\overline{h}}}\\
&= 2\overline{h}\log r+\frac{1}{2}(q+h+\overline{h})+\frac{1}{2}(q+h+\overline{h})\frac{1-\frac{1}{r}}{\sqrt{\pi\log r}}\\
&\le 2\overline{h}\log r+\frac{1}{2}(q+h+\overline{h})+\frac{1}{5}(q+h+\overline{h})\\
&= 2\overline{h}\log\Big(\frac{2n}{q+h+\overline{h}}\Big)+\frac{7}{10}(q+h+\overline{h})\,,
\end{aligned}$$

where we used (59) in the second inequality. This is (27).

Case 2: $\boldsymbol{\beta<1}$. We rewrite (78) as

$$\mathbb{E}_g\Big[\mathrm{dist}\big(g,\mathrm{cone}\,\partial f_1(x^\star)\big)^2\Big] \le s+F(\beta,t)+G(\beta,t) \quad t^2\Big[\overline{h}(\beta+1)^2+(s-\overline{h})(\beta-1)^2\Big]\,, \quad (81)$$

where we used $s:=|I|$, $|IJ^c|=s-(h+\overline{h})$ (cf. (21)), and

$$F(\beta,t):=(q-s)\frac{\varphi(t(1-\beta))}{t(1-\beta)}$$

$$G(\beta,t):=(2n-(q+h+\overline{h}))\frac{\varphi(t(\beta+1))}{t(\beta+1)}\,. \quad (82)$$

Note that we used (22) and (25) when defining F and G. We will consider two cases: $F(\beta,t)\le G(\beta,t)$ and $F(\beta,t)\ge G(\beta,t)$. Note that

$$\begin{aligned}
&\frac{F(\beta,t)}{G(\beta,t)}\lessgtr 1\\
\iff\quad &\frac{q-s}{2n-(q+h+\overline{h})}\lessgtr\frac{1-\beta}{1+\beta}\exp\big(-2\beta t^2\big)\,. \quad (83)
\end{aligned}$$

- Suppose $F(\beta,t)\le G(\beta,t)$, i.e., (83) is satisfied with $\le$. The bound in (81) implies

$$\begin{aligned}
&\mathbb{E}_g\Big[\mathrm{dist}\big(g,\mathrm{cone}\,\partial f_1(x^\star)\big)^2\Big]\\
&\le t^2\Big[\overline{h}(\beta+1)^2+(s-\overline{h})(\beta-1)^2\Big]+s+2G(\beta,t)\\
&= t^2\Big[\overline{h}(\beta+1)^2+(s-\overline{h})(\beta-1)^2\Big]+s\\
&\quad+2\Big[2n-(q+h+\overline{h})\Big]\frac{\exp\big(-\frac{t^2}{2}(\beta+1)^2\big)}{\sqrt{2\pi}t(\beta+1)}\,,
\end{aligned} \quad (84)$$

where we used the definition of φ. We now select t as

$$t^\star=\frac{1}{\beta+1}\sqrt{2\log\Big(\frac{2n}{q+h+\overline{h}}\Big)}=\frac{1}{\beta+1}\sqrt{2\log r}\,,$$

where $r:=2n/(q+h+\overline{h})$ is as before. Replacing $t^\star$ in (84) yields

$$\begin{aligned}
&\mathbb{E}_g\Big[\mathrm{dist}\big(g,\mathrm{cone}\,\partial f_1(x^\star)\big)^2\Big]\\
&\le 2\Big[\overline{h}+(s-\overline{h})\frac{(\beta-1)^2}{(\beta+1)^2}\Big]\log r+s\\
&\quad+2\Big[2n-(q+h+\overline{h})\Big]\frac{1}{\sqrt{2\log r}}\frac{1}{\sqrt{2\pi}}\frac{1}{\frac{2n}{q+h+\overline{h}}}\\
&= 2\Big[\overline{h}+(s-\overline{h})\frac{(\beta-1)^2}{(\beta+1)^2}\Big]\log r+s\\
&\quad+(q+h+\overline{h})\frac{1-\frac{1}{r}}{\sqrt{\pi\log r}}\\
&\le 2\Big[\overline{h}+(s-\overline{h})\frac{(\beta-1)^2}{(\beta+1)^2}\Big]\log\Big(\frac{2n}{q+h+\overline{h}}\Big)+s\\
&\quad+\frac{2}{5}(q+h+\overline{h})\,,
\end{aligned}$$

which is (28). We used (59) in the last inequality. This bound is valid only when (83) with $\le$ is satisfied with $t=t^\star$, i.e.,

$$\frac{q-s}{2n-(q+h+\overline{h})}\le\frac{1-\beta}{1+\beta}\Big(\frac{q+h+\overline{h}}{2n}\Big)^{\frac{4\beta}{(\beta+1)^2}}\,,$$

which is condition (C1).

- Suppose now that $F(\beta,t)\ge G(\beta,t)$, i.e., (83) is satisfied with $\ge$. Then, (81) becomes

$$\begin{aligned}
&\mathbb{E}_g\Big[\mathrm{dist}\big(g,\mathrm{cone}\,\partial f_1(x^\star)\big)^2\Big]\\
&\le t^2\Big[\overline{h}(\beta+1)^2+(s-\overline{h})(\beta-1)^2\Big]+s+2F(\beta,t)\\
&= t^2\Big[\overline{h}(\beta+1)^2+(s-\overline{h})(\beta-1)^2\Big]+s
\end{aligned}$$

$$+ 2(q-s)\frac{\exp\big(-\frac{t^2}{2}(1-\beta)^2\big)}{\sqrt{2\pi}\,t(1-\beta)}\,. \tag{85}$$

We select t as

$$t^\star = \frac{1}{1-\beta}\sqrt{2\log\Big(\frac{q}{s}\Big)} = \frac{1}{1-\beta}\sqrt{2\log\,r}\,,$$

where r is now $r := q/s$. Since in case 2b) of the theorem, we assume $0 < |I^cJ| = q-s$, we have $t^\star > 0$. Notice that $t^\star$ is finite, because $s > 0$ (given that $x^\star \neq 0$). Replacing $t^\star$ into (85) yields

$$\begin{aligned}
&\mathbb{E}_g\Big[\mathrm{dist}\big(g, \mathrm{cone}\,\partial f_1(x^\star)\big)^2\Big] \\
&\le 2\left[\overline{h}\frac{(1+\beta)^2}{(1-\beta)^2} + s - \overline{h}\right]\log\,r + s \\
&\qquad + 2(q-s)\frac{1}{\sqrt{2\log\,r}}\frac{1}{\sqrt{2\pi}}\frac{1}{\frac{q}{s}} \\
&= 2\left[\overline{h}\frac{(1+\beta)^2}{(1-\beta)^2} + s - \overline{h}\right]\log\,r + s + s\frac{1-\frac{1}{r}}{\sqrt{\pi\log\,r}} \\
&\le 2\left[\overline{h}\frac{(1+\beta)^2}{(1-\beta)^2} + s - \overline{h}\right]\log\,r + s + \frac{2}{5}s \\
&= 2\left[\overline{h}\frac{(1+\beta)^2}{(1-\beta)^2} + s - \overline{h}\right]\log\Big(\frac{q}{s}\Big) + \frac{7}{5}s\,,
\end{aligned}$$

which is (29). Again, we used (59) in the last inequality. This bound is valid only if (83) with $\ge$ is satisfied for $t = t^\star$, i.e.,

$$\frac{q-s}{2n-(q+h+\overline{h})} \ge \frac{1-\beta}{1+\beta}\Big(\frac{s}{q}\Big)^{\frac{4\beta}{(1-\beta)^2}}\,,$$

which is condition (C2).

Case 3: $\beta > 1$. We rewrite (79) as

$$\begin{aligned}
&\mathbb{E}_g\Big[\mathrm{dist}\big(g, \mathrm{cone}\,\partial f_1(x^\star)\big)^2\Big] \\
&\le t^2\Big[\overline{h}(\beta+1)^2 + (q+h-s)(\beta-1)^2\Big] + l + H(\beta,t) \\
&\qquad + G(\beta,t)\,,
\end{aligned} \tag{86}$$

where we used $l := |J|$, $|I^cJ| = q-s$ (cf. (22)), G is defined in (82), and

$$H(\beta,t) := (s-(h+\overline{h}))\frac{\varphi(t(\beta-1))}{t(\beta-1)}\,.$$

Note that we used (21) when defining H. We also consider two cases: $H(\beta,t) \le G(\beta,t)$ and $H(\beta,t) \ge H(\beta,t)$. Note that

$$\begin{aligned}
&\frac{H(\beta,t)}{G(\beta,t)} \lessgtr 1 \\
\iff\quad &\frac{s-(h+\overline{h})}{2n-(q+h+\overline{h})} \lessgtr \frac{\beta-1}{\beta+1}\exp\big(-2\beta t^2\big)\,.
\end{aligned} \tag{87}$$

- Suppose $H(\beta,t) \le G(\beta,t)$, i.e., (87) is satisfied with $\le$. Then, (86) implies

$$\begin{aligned}
&\mathbb{E}_g\Big[\mathrm{dist}\big(g, \mathrm{cone}\,\partial f_1(x^\star)\big)^2\Big] \\
&\le t^2\Big[\overline{h}(\beta+1)^2 + (q+h-s)(\beta-1)^2\Big] + l + 2G(\beta,t) \\
&= t^2\Big[\overline{h}(\beta+1)^2 + (q+h-s)(\beta-1)^2\Big] + l \\
&\qquad + 2(2n-(q+h+\overline{h}))\frac{\exp\Big(-\frac{t^2}{2}(\beta+1)^2\Big)}{\sqrt{2\pi}t(\beta+1)}\,.
\end{aligned} \tag{88}$$

Now we select

$$t^\star = \frac{1}{\beta+1}\sqrt{2\log\Big(\frac{2n}{q+h+\overline{h}}\Big)} = \frac{1}{\beta+1}\sqrt{2\log\,r}\,,$$

where $r := 2n/(q+h+\overline{h})$. Again, note that our assumptions imply that $t^\star$ is well defined and positive. Replacing $t^\star$ into (88) yields

$$\begin{aligned}
&\mathbb{E}_g\Big[\mathrm{dist}\big(g, \mathrm{cone}\,\partial f_1(x^\star)\big)^2\Big] \\
&\le 2\left[\overline{h} + (q+h-s)\Big(\frac{\beta-1}{\beta+1}\Big)^2\right]\log\,r + l \\
&\qquad + (2n-(q+h+\overline{h}))\frac{1}{\sqrt{\pi\log\,r}}\frac{1}{\frac{2n}{q+h+\overline{h}}} \\
&= 2\left[\overline{h} + (q+h-s)\Big(\frac{\beta-1}{\beta+1}\Big)^2\right]\log\,r + l \\
&\qquad + (q+h+\overline{h})\frac{1-\frac{1}{r}}{\sqrt{\pi\log\,r}} \\
&\le 2\left[\overline{h} + (q+h-s)\Big(\frac{\beta-1}{\beta+1}\Big)^2\right]\log\Big(\frac{2n}{q+h+\overline{h}}\Big) + l \\
&\qquad + \frac{2}{5}(q+h+\overline{h})\,.
\end{aligned}$$

This is (30). Again, (59) was used in the last step. This bound is valid only when (87) with $\le$ is satisfied for $t = t^\star$, i.e.,

$$\frac{s-(h+\overline{h})}{2n-(q+h+\overline{h})} \le \frac{\beta-1}{\beta+1}\Big(\frac{q+h+\overline{h}}{2n}\Big)^{\frac{4\beta}{(\beta+1)^2}}\,,$$

which is condition (C3).

- Suppose now that $H(\beta,t) \ge G(\beta,t)$. Then, (86) implies

$$\begin{aligned}
&\mathbb{E}_g\Big[\mathrm{dist}\big(g, \mathrm{cone}\,\partial f_1(x^\star)\big)^2\Big] \\
&\le t^2\Big[\overline{h}(\beta+1)^2 + (q+h-s)(\beta-1)^2\Big] + l + 2H(\beta,t) \\
&= t^2\Big[\overline{h}(\beta+1)^2 + (q+h-s)(\beta-1)^2\Big] + l \\
&\qquad + 2(s-(h+\overline{h}))\frac{\exp\Big(-\frac{t^2}{2}(\beta-1)^2\Big)}{\sqrt{2\pi}t(\beta-1)}\,.
\end{aligned} \tag{89}$$

Given our assumption that $|IJ| = h+\overline{h} > 0$ in case 3b), we can select t as

$$t^\star = \frac{1}{\beta-1}\sqrt{2\log\Big(\frac{s}{h+\overline{h}}\Big)} = \frac{1}{\beta-1}\sqrt{2\log\,r}\,,$$

where $r := s/(h+\overline{h})$. We also assume that $|IJ^c| = s-(h+\overline{h}) > 0$, making $t^\star > 0$. Replacing $t^\star$ into (89) gives

$$\begin{aligned}
&\mathbb{E}_g\Big[\mathrm{dist}\big(g, \mathrm{cone}\,\partial f_1(x^\star)\big)^2\Big] \\
&\le 2\left[\overline{h}\Big(\frac{\beta+1}{\beta-1}\Big)^2 + q + h - s\right]\log\,r + l \\
&\qquad + 2(s-(h+\overline{h}))\frac{1}{\sqrt{2\log\,r}}\frac{1}{\sqrt{2\pi}}\frac{1}{\frac{s}{h+\overline{h}}}
\end{aligned}$$

$$= 2\left[\overline{h}\Big(\frac{\beta+1}{\beta-1}\Big)^2 + q + h - s\right]\log r + l + (h+\overline{h})\frac{1-\frac{1}{r}}{\sqrt{\pi\log r}}$$
$$\leq 2\left[\overline{h}\Big(\frac{\beta+1}{\beta-1}\Big)^2 + q + h - s\right]\log r + l + \frac{2}{5}(h+\overline{h}),$$

which is (31). Again, (59) was employed in the last inequality. This bound is valid only when (87) with $\geq$ holds for $t = t^\star$, that is,

$$\frac{s-(h+\overline{h})}{2n-(q+h+\overline{h})} \geq \frac{\beta-1}{\beta+1}\Big(\frac{h+\overline{h}}{s}\Big)^{\frac{4\beta}{(\beta-1)^2}},$$

which is condition (C4). This concludes the proof. □

Remarks. The bound for case 1), i.e., $\beta = 1$, is clearly the sharpest one, since it does not use inequalities like (83) or (87). Perhaps the "loosest" inequality it uses is (59) in Lemma 18. According to its proof in Appendix B, that bound is exact when $x = 2$, which means $2n/(q+h+\overline{h}) = 2$, for $\beta = 1$. The bounds for $\beta \neq 1$ are not as sharp, due to (83) and (87). Note also that some sharpness is lost by selecting specific values of t and not the optimal ones (cf. (65)).

C. Proof of Theorem 13

The proof follows from taking the minimum over t in the right-hand side of (66) and using (67). □

D. Proof of Theorem 14

The steps to prove the theorem are the same steps as the ones in the proof of Theorem 12. So, we will omit some details. Whenever $0 \notin \partial f_2(x^\star)$, we can use the bound in (64) with f_1 replaced by f_2, i.e.,

$$w\big(T_{f_2}(x^\star)\big)^2 \leq \mathbb{E}_g\Big[\operatorname{dist}\big(g, \operatorname{cone}\partial f_2(x^\star)\big)^2\Big]. \tag{90}$$

Note that this bound results from the characterization of the normal cone provided in Proposition 5 and from Jensen's inequality. Part 2) of Lemma 19 establishes that our assumptions guarantee that $0 \notin \partial f_2(x^\star)$ and, thus, that we can use (90). Next, we express the right-hand side of (90) component-wise, and then we establish bounds for each term.

A vector $d \in \mathbb{R}^n$ belongs to the cone generated by $\partial f_2(x^\star)$ if, for some $t \geq 0$ and some $y \in \partial f_2(x^\star)$, $d = ty$. According to (63), each component d_i satisfies

$$\begin{cases} d_i = t\operatorname{sign}(x_i^\star) + t\beta(x_i^\star - w_i) & ,\ i \in I \\ |d_i + t\beta w_i| \leq t & ,\ i \in I^c, \end{cases}$$

for some $t \geq 0$. This allows expanding the right-hand side of (90) as

$$\mathbb{E}_g\Big[\operatorname{dist}\big(g, \operatorname{cone}\partial f_2(x^\star)\big)^2\Big] = \mathbb{E}_g\left[\min_{t\geq 0}\left\{\sum_{i\in I}\operatorname{dist}\Big(g_i\,,\, t\operatorname{sign}(x_i^\star) + t\beta(x_i^\star - w_i)\Big)^2 + \sum_{i\in I^c}\operatorname{dist}\Big(g_i\,,\,\mathcal{I}(-t\beta w_i, t)\Big)^2\right\}\right].$$

As in the proof of Theorem 12, we fix t and select it later (cf. (65)). Doing so, gives

$$\mathbb{E}_g\Big[\operatorname{dist}\big(g, \operatorname{cone}\partial f_2(x^\star)\big)^2\Big] \leq \sum_{i\in I}\mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\, t\operatorname{sign}(x_i^\star) + t\beta(x_i^\star - w_i)\Big)^2\Big] \tag{91a}$$
$$+ \sum_{i\in I^c}\mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\,\mathcal{I}(-t\beta w_i, t)\Big)^2\Big]. \tag{91b}$$

Next, we use Lemma 17 to derive a closed-form expression for (91a) and establish a bound on (91b).

Expression for (91a)**.** Using (53),

$$\begin{aligned}(91a) &= \sum_{i\in I}\mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\, t\operatorname{sign}(x_i^\star) + t\beta(x_i^\star - w_i)\Big)^2\Big] \\ &= \sum_{i\in I}\Big[\big(t\operatorname{sign}(x_i^\star) + t\beta(x_i^\star - w_i)\big)^2 + 1\Big] \\ &= t^2\Big[\sum_{i\in I_+}(1+\beta(x_i^\star - w_i))^2 + \sum_{i\in I_-}(1-\beta(x_i^\star - w_i))^2\Big] + |I|, \end{aligned} \tag{92}$$

where we decomposed $I = I_+ \cup I_-$.

Bounding (91b)**.** We have

$$(91b) = \sum_{i\in I^cJ}\mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\,\mathcal{I}(-t\beta w_i, t)\Big)^2\Big] + \sum_{i\in I^cJ^c}\mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i\,,\,\mathcal{I}(0,t)\Big)^2\Big]. \tag{93}$$

The second term in the right-hand side of (93) can be bounded according to (54):

$$\sum_{i\in I^cJ^c}\mathbb{E}_{g_i}\Big[\operatorname{dist}(g_i, \mathcal{I}(0,t))^2\Big] \leq 2|I^cJ^c|\frac{\varphi(t)}{t}. \tag{94}$$

The first term, however, is more complicated. Recall that $I^cJ = \{i : w_i \neq x_i^\star = 0\}$. Let us analyze the several possible situations for the interval $\mathcal{I}(-t\beta w_i, t) = [t(-\beta w_i - 1), t(-\beta w_i + 1)]$. It does not contain zero whenever

$$t(-\beta w_i - 1) > 0 \iff t \neq 0 \quad\text{and}\quad w_i < -\frac{1}{\beta},$$

or

$$t(-\beta w_i + 1) < 0 \iff t \neq 0 \quad\text{and}\quad w_i > \frac{1}{\beta}.$$

In addition to the subsets of I^cJ defined in (34)-(35), define

$$K_- := \Big\{i \in I^cJ\,:\, w_i < -\frac{1}{\beta}\Big\}$$
$$K_+ := \Big\{i \in I^cJ\,:\, w_i > \frac{1}{\beta}\Big\}$$
$$K_-^= := \Big\{i \in I^cJ\,:\, w_i = -\frac{1}{\beta}\Big\}$$
$$K_+^= := \Big\{i \in I^cJ\,:\, w_i = \frac{1}{\beta}\Big\}$$
$$L := \Big\{i \in I^cJ\,:\, |w_i| < \frac{1}{\beta}\Big\},$$

where we omit the dependency of these sets on β for notational simplicity. Noticing that $I^cJ = K_- \cup K_+ \cup K_-^= \cup K_+^= \cup L$ and using Lemma 17, we obtain

$$\begin{aligned}
&\sum_{i\in I^cJ} \mathbb{E}_{g_i}\Big[\mathrm{dist}\Big(g_i\,,\,\mathcal{I}(-t\beta w_i,t)\Big)^2\Big] \\
&\le \sum_{i\in K_-}\Big[1+t^2(\beta w_i+1)^2+\frac{\varphi(t(1-\beta w_i))}{t(1-\beta w_i)}\Big] \\
&\quad+\sum_{i\in K_+}\Big[1+t^2(1-\beta w_i)^2+\frac{\varphi(t(1+\beta w_i))}{t(1+\beta w_i)}\Big] \\
&\quad+\sum_{i\in K_-^=}\Big[\frac{1}{2}+\frac{\varphi(t(1-\beta w_i))}{t(1-\beta w_i)}\Big] \\
&\quad+\sum_{i\in K_+^=}\Big[\frac{1}{2}+\frac{\varphi(t(1+\beta w_i))}{t(1+\beta w_i)}\Big] \\
&\quad+\sum_{i\in L}\Big[\frac{\varphi(t(1+\beta w_i))}{t(1+\beta w_i)}+\frac{\varphi(t(1-\beta w_i))}{t(1-\beta w_i)}\Big] \\
&= |K^{\neq}(\beta)|+\frac{1}{2}|K^{=}(\beta)|+t^2\Big[\sum_{i\in K_-}(\beta w_i+1)^2 \\
&\quad+\sum_{i\in K_+}(\beta w_i-1)^2\Big]+\sum_{i\in K_-\cup K_-^=}\frac{\varphi(t(1-\beta w_i))}{t(1-\beta w_i)} \\
&\quad+\sum_{i\in K_+\cup K_+^=}\frac{\varphi(t(1+\beta w_i))}{t(1+\beta w_i)} \\
&\quad+\sum_{i\in L}\Big[\frac{\varphi(t(1+\beta w_i))}{t(1+\beta w_i)}+\frac{\varphi(t(1-\beta w_i))}{t(1-\beta w_i)}\Big] \\
&\le |K^{\neq}(\beta)|+\frac{1}{2}|K^{=}(\beta)|+t^2\Big[\sum_{i\in K_-}(\beta w_i+1)^2 \\
&\quad+\sum_{i\in K_+}(\beta w_i-1)^2\Big] \\
&\quad+\Big[|K_-|+|K_-^=|+|L|\Big]\frac{\varphi(t(1-\beta\,\overline{w}_p))}{t(1-\beta\,\overline{w}_p)} \\
&\quad+\Big[|K_+|+|K_+^=|+|L|\Big]\frac{\varphi(t(1+\beta\,\overline{w}_m))}{t(1+\beta\,\overline{w}_m)}\,, \qquad (95)
\end{aligned}$$

where, in the second step, we used the definitions of $K^=$ and $K^{\neq}$ in (34) and (35), respectively. In the last step, we used $\overline{w}_p := |w_p|$ and $\overline{w}_m := |w_m|$, for

$$\begin{aligned}
p &:= \operatorname*{arg\,min}_{i\in K_-\cup K_-^=\cup L} 1-\beta w_i \\
m &:= \operatorname*{arg\,min}_{i\in K_+\cup K_+^=\cup L} 1+\beta w_i\,.
\end{aligned}$$

Note that $\overline{w} = \big|\arg\min_{w=\overline{w}_p,\overline{w}_m}\{|w|-1/\beta\}\big|$, since the union of the sets K_-, $K_-^=$, L, K_+, and $K_+^=$ gives I^cJ. Therefore, $\varphi(t(1-\beta\,\overline{w}_j))/(t(1-\beta\,\overline{w}_j)) \le \varphi(t(1-\beta\,\overline{w}))/(t|1-\beta\,\overline{w}|)$, for $j=p,m$. Using this in the last two terms of (95), and noticing that

$$|K_-|+|K_-^=|+|K_+|+|K_+^=| = \Big\{i\in I^cJ\,:\,|w_i|\ge\frac{1}{\beta}\Big\} =: K(\beta)\,,$$

we obtain

$$\begin{aligned}
&\sum_{i\in I^cJ}\mathbb{E}_{g_i}\Big[\mathrm{dist}\Big(g_i\,,\,\mathcal{I}(-t\beta w_i,t)\Big)^2\Big] \le |K^{\neq}(\beta)|+\frac{1}{2}|K^{=}(\beta)| \\
&\quad+t^2\Big[\sum_{i\in K_-}(\beta w_i+1)^2+\sum_{i\in K_+}(\beta w_i-1)^2\Big] \\
&\quad+\Big[|K(\beta)|+2|L|\Big]\frac{\varphi(t(1-\beta\,\overline{w}))}{t|1-\beta\,\overline{w}|}\,. \qquad (96)
\end{aligned}$$

Bounding (91a) **+** (91b)**.** Adding up (92), (94), and (96), we obtain

$$\begin{aligned}
&\mathbb{E}_g\Big[\mathrm{dist}\big(g,\mathrm{cone}\,\partial f_2(x^\star)\big)^2\Big] \\
&\le |I|+t^2\Big[\sum_{i\in I_+}(1+\beta(x_i^\star-w_i))^2+\sum_{i\in I_-}(1-\beta(x_i^\star-w_i))^2\Big] \\
&\quad+2|I^cJ^c|\frac{\varphi(t)}{t}+|K^{\neq}(\beta)|+\frac{1}{2}|K^{=}(\beta)| \\
&\quad+t^2\Big[\sum_{i\in K_-}(\beta w_i+1)^2+\sum_{i\in K_+}(\beta w_i-1)^2\Big] \\
&\quad+\Big[|K(\beta)|+2|L|\Big]\frac{\varphi(t(1-\beta\,\overline{w}))}{t|1-\beta\,\overline{w}|} \\
&= v_\beta t^2+|I|+|K^{\neq}(\beta)|+\frac{1}{2}|K^{=}(\beta)|+\Big[|K(\beta)|+2|L|\Big]\times \\
&\quad\times\frac{\varphi(t(1-\beta\,\overline{w}))}{t|1-\beta\,\overline{w}|}+2|I^cJ^c|\frac{\varphi(t)}{t} \\
&\le v_\beta t^2+|I|+|K^{\neq}(\beta)|+\frac{1}{2}|K^{=}(\beta)|+2F(t,\beta,\overline{w})+2G(t)\,, \qquad (97)
\end{aligned}$$

where we used $|K(\beta)|+2|L| \le 2|I^cJ| = 2(q-s)$ (cf. (22)) in the last inequality. Note that v_β is defined in (36) and that we defined

$$\begin{aligned}
F(t,\beta,\overline{w}) &:= (q-s)\frac{\varphi(t(1-\beta\,\overline{w}))}{t|1-\beta\,\overline{w}|} \\
G(t) &:= (n-q)\frac{\varphi(t)}{t}\,.
\end{aligned}$$

We consider two scenarios: $F(t,\beta,\overline{w}) \le G(t)$ and $F(t,\beta,\overline{w}) \ge G(t)$. Note that

$$\begin{aligned}
&\frac{F(t,\beta,\overline{w})}{G(t)} \lessgtr 1 \\
&\iff \quad \frac{q-s}{n-q} \lessgtr |1-\beta\overline{w}|\exp\Big(t^2\beta\overline{w}(\frac{\beta\overline{w}}{2}-1)\Big)\,, \qquad (98)
\end{aligned}$$

- Suppose $F(t,\beta,\overline{w}) \le G(t)$, i.e., (98) is satisfied with $\le$. The bound in (97) implies

$$\begin{aligned}
&\mathbb{E}_g\Big[\mathrm{dist}\big(g,\mathrm{cone}\,\partial f_2(x^\star)\big)^2\Big] \\
&\quad\le v_\beta t^2+s+|K^{\neq}(\beta)|+\frac{1}{2}|K^{=}(\beta)|+4G(t) \\
&\quad= v_\beta t^2+s+|K^{\neq}(\beta)|+\frac{1}{2}|K^{=}(\beta)| \\
&\qquad+4(n-q)\frac{1}{t}\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{t^2}{2}\Big)\,.
\end{aligned}$$

We now select t as

$$t^\star = \sqrt{2\log\Big(\frac{n}{q}\Big)} = \sqrt{2\log\,r}\,,$$

where $r := n/q$. Note that r is well defined, since $x^\star \neq 0$ implies $q > 0$. Also, the assumption $q < n$ implies $t^\star > 0$. Setting t to $t^\star$ and using (59), we get

$$\begin{aligned}
&\mathbb{E}_g\Big[\operatorname{dist}\big(g, \operatorname{cone}\partial f_2(x^\star)\big)^2\Big] \\
&\leq 2v_\beta \log\Big(\frac{n}{q}\Big) + s + |K^{\neq}(\beta)| + \frac{1}{2}|K^{=}(\beta)| \\
&\qquad + 4(n-q)\frac{1}{\sqrt{2\log r}}\frac{1}{\sqrt{2\pi}}\frac{1}{\frac{n}{q}} \\
&= 2v_\beta \log\Big(\frac{n}{q}\Big) + s + |K^{\neq}(\beta)| + \frac{1}{2}|K^{=}(\beta)| \\
&\qquad + 2q\frac{1-\frac{1}{r}}{\sqrt{\pi \log r}} \\
&\leq 2v_\beta \log\Big(\frac{n}{q}\Big) + s + |K^{\neq}(\beta)| + \frac{1}{2}|K^{=}(\beta)| + \frac{4}{5}q\,,
\end{aligned}$$

which is (39). This bound is valid only if (98) with $\leq$ is satisfied for $t^\star$, i.e.,

$$\frac{q-s}{n-q} \leq |1-\beta\,\overline{w}|\exp\Big(2\beta\,\overline{w}\log\Big(\frac{n}{q}\Big)\Big(\frac{\beta\,\overline{w}}{2}-1\Big)\Big)\,,$$

which is condition (38).

- Suppose now that $F(t,\beta,\overline{w}) \geq G(t)$, i.e., (98) is satisfied with $\geq$. The bound in (97) implies

$$\begin{aligned}
&\mathbb{E}_g\Big[\operatorname{dist}\big(g, \operatorname{cone}\partial f_2(x^\star)\big)^2\Big] \\
&\leq v_\beta t^2 + s + |K^{\neq}(\beta)| + \frac{1}{2}|K^{=}(\beta)| + 4F(t,\beta,\overline{w}) \\
&= v_\beta t^2 + s + |K^{\neq}(\beta)| + \frac{1}{2}|K^{=}(\beta)| \\
&\qquad + 4(q-s)\frac{1}{\sqrt{2\pi}}\frac{1}{t|1-\beta\,\overline{w}|}\exp\Big(-\frac{t^2}{2}(1-\beta\,\overline{w})^2\Big)\,.
\end{aligned}$$

And we select t as

$$t^\star = \frac{1}{|1-\beta\,\overline{w}|}\sqrt{2\log\Big(\frac{q}{s}\Big)} = \frac{1}{|1-\beta\,\overline{w}|}\sqrt{2\log r}\,,$$

where $r := q/s$. Again, r is well defined because $s > 0$. Since we assume $q > s$, $t^\star > 0$. Setting t to $t^\star$ and using (59) again, we obtain

$$\begin{aligned}
&\mathbb{E}_g\Big[\operatorname{dist}\big(g, \operatorname{cone}\partial f_2(x^\star)\big)^2\Big] \\
&\leq \frac{2v_\beta}{|1-\beta\,\overline{w}|^2}\log\Big(\frac{q}{s}\Big) + s + |K^{\neq}(\beta)| + \frac{1}{2}|K^{=}(\beta)| \\
&\qquad + 4(q-s)\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{2\log r}}\frac{1}{\frac{q}{s}} \\
&= \frac{2v_\beta}{|1-\beta\,\overline{w}|^2}\log\Big(\frac{q}{s}\Big) + s + |K^{\neq}(\beta)| + \frac{1}{2}|K^{=}(\beta)| \\
&\qquad + 2s\frac{1-\frac{1}{r}}{\sqrt{\pi\log r}} \\
&\leq \frac{2v_\beta}{|1-\beta\,\overline{w}|^2}\log\Big(\frac{q}{s}\Big) + |K^{\neq}(\beta)| + \frac{1}{2}|K^{=}(\beta)| + \frac{9}{5}s\,,
\end{aligned}$$

which is (41). This bound is valid only when (98) with $\geq$ is satisfied for $t^\star$, i.e.,

$$\frac{q-s}{n-q} \geq |1-\beta\overline{w}|\exp\Big(4\frac{(\beta\,\overline{w}-2)\beta\,\overline{w}}{|1-\beta\overline{w}|^2}\log\Big(\frac{q}{s}\Big)\Big)\,,$$

which is condition (40). □

Remarks. Although these bounds were derived using the same techniques as the ones for ℓ_1-ℓ_1 minimization, they are much looser. The main reason is their dependency on the magnitudes of $x^\star$, w, and $x^\star - w$. This forced us to consider a worst-case scenario in the last step of (95).

E. Proof of Theorem 15

The proof follows from taking the minimum over t in the right-hand side of (91) and using (92). □

F. Proof of Corollary 16

First note that the corollary's assumptions imply the assumptions of both Theorems 13 and 15. In particular, if (48b) holds strictly for at least one component k, then $k \in IJ$ and $\beta_2 \neq \operatorname{sign}(x_k^\star)/(w_k - x_k^\star)$. We write the right-hand sides of (33) and (45) as $\min_{t\geq 0}\,\phi_1(t)$ and $\min_{t\geq 0}\,\phi_2(t)$, where

$$\begin{aligned}
\phi_1(t) := \overline{h} + h + 4\overline{h}t^2 &+ \sum_{i\in I^cJ^c}\mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i,\,\mathcal{I}(0,\,2t)\Big)^2\Big] \\
&+ \sum_{i\in I^cJ}\mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i,\,\mathcal{I}(-t\operatorname{sign}(w_i),\,t)\Big)^2\Big] \qquad (99)
\end{aligned}$$

$$\begin{aligned}
\phi_2(t) := s + t^2\sum_{i\in IJ}\Big[1+\beta_2\operatorname{sign}(x_i^\star)(x_i^\star - w_i)\Big]^2 &+ \sum_{i\in I^cJ^c}\mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i,\,\mathcal{I}(0,\,t)\Big)^2\Big] \\
&+ \sum_{i\in I^cJ}\mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i,\,\mathcal{I}(-t\beta_2 w_i,\,t)\Big)^2\Big]\,. \qquad (100)
\end{aligned}$$

Note that we used the assumptions $\beta_1 = 1$ and $IJ^c = \emptyset$. We now compare (99) and (100) term-by-term. First, note that

$$\begin{aligned}
4\overline{h} &= 4\Big(|I_+J_+| + |I_-J_-|\Big) \\
&\leq \sum_{i\in I_+J_+}\Big[1+\beta_2(x_i^\star - w_i)\Big]^2 + \sum_{i\in I_-J_-}\Big[1-\beta_2(x_i^\star - w_i)\Big]^2 \\
&\leq \sum_{i\in IJ}\Big[1+\beta_2\operatorname{sign}(x_i^\star - w_i)\Big]^2\,, \qquad (101)
\end{aligned}$$

where we used (48b) in the first inequality. Furthermore,

$$\overline{h} + h = |IJ| \leq |I| = s \qquad (102)$$

$$\mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i,\,\mathcal{I}(0,\,2t)\Big)^2\Big] \leq \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i,\,\mathcal{I}(0,\,t)\Big)^2\Big] \qquad (103)$$

$$\begin{aligned}
&\mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i,\,\mathcal{I}(-t\operatorname{sign}(w_i),\,t)\Big)^2\Big] \\
&\qquad \leq \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i,\,\mathcal{I}(-t\beta_2\,w_i,\,t)\Big)^2\Big]\,, \quad \forall\, i \in I^cJ\,. \qquad (104)
\end{aligned}$$

In (102), we used (20). To get (103), just notice that $\mathcal{I}(0,t) \subset \mathcal{I}(0,2t)$. To obtain (104), we used (48a) and the fact that

$$\mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i,\,\mathcal{I}(t,\,t)\Big)^2\Big] \leq \mathbb{E}_{g_i}\Big[\operatorname{dist}\Big(g_i,\,\mathcal{I}(c\,t,\,t)\Big)^2\Big]\,, \qquad (105)$$

for any $|c| \geq 1$. To see why (105) holds, write $f_b(a) := \mathbb{E}_g\big[\operatorname{dist}(g,\,\mathcal{I}(a,b))^2\big]$, i.e., with b fixed. Using (52) and the

identities $dQ(x)/dx = -\varphi(x)$, $d\varphi(x)/dx = -x\,\varphi(x)$, and $Q(x) = 1 - Q(-x)$, it can be shown that

$$\frac{d}{da} f_b(a) = 2\big[\varphi(a-b) - \varphi(a+b) + (a+b)Q(a+b) + (a-b)Q(b-a)\big]. \quad (106)$$

The density $\varphi(x)$ is nonincreasing for $x \geq 0$ and nondecreasing for $x \leq 0$. Therefore, all terms of (106) are nonnegative for $a \geq b \geq 0$, meaning that $f_b(a)$ does not decrease with a. This shows (105) for $c \geq 1$. For $c \leq -1$, note that all terms in (106) are nonpositive whenever $-a \geq b \geq 0$, that is, $f_b(a)$ increases with (a negative) a. To conclude the proof, notice that (101)-(104) show $\phi_1(t) \leq \phi_2(t)$, for all $t \geq 0$. Assume t_1 (resp. t_2) is a minimizer of ϕ_1 (resp. ϕ_2), which exists due to the continuity and coercivity of ϕ_1 (the same for ϕ_2). Then, $\phi_1(t_1) \leq \phi_1(t_2) \leq \phi_2(t_2)$, concluding the proof. □

VII. Conclusions

We studied two schemes for integrating prior information in CS: ℓ_1-ℓ_1 and ℓ_1-ℓ_2 minimization. For each scheme, we established bounds on the number of measurements that guarantee successful reconstruction with high probability, under Gaussian measurement matrices. The bounds established for ℓ_1-ℓ_1 minimization are quite sharp and are minimized for $\beta = 1$. In contrast, the bounds for ℓ_1-ℓ_2 minimization can be quite loose, and the β that minimizes them depends on several unknown problem parameters. According to our theory, geometrical interpretations, and experimental results, ℓ_1-ℓ_1 minimization has strong advantages over both standard CS and ℓ_1-ℓ_2 minimization. The insights revealed by our theory also helped us design schemes that improve the quality of prior information. Possible future research directions include extending our bounds to more complex signal structures, for example, block sparsity and the k-support norm.

VIII. Acknowledgments

We would like to thank João Xavier for insightful discussions regarding Proposition 3 and Fig. 3, Volkan Cevher for fruitful discussions about relevant related literature, and the two anonymous reviewers for suggestions that improved the content of the paper.

Appendix A
Proof of Lemma 17

A. Part I) Exact expression

For any a and $b \geq 0$,

$$\mathbb{E}_g\Big[\operatorname{dist}\big(g, \mathcal{I}(a,b)\big)^2\Big] = \mathbb{E}_g\left[\begin{array}{ll}\min\limits_{u} & (u-g)^2 \\ \text{s.t.} & |u-a| \leq b\end{array}\right]$$
$$= \frac{1}{\sqrt{2\pi}}\int_{a+b}^{+\infty}\big(g-(a+b)\big)^2\exp\Big(-\frac{g^2}{2}\Big)\,dg + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a-b}\big(g-(a-b)\big)^2\exp\Big(-\frac{g^2}{2}\Big)\,dg$$
$$= A(a+b) + B(a-b)\,, \quad (107)$$

where

$$A(x) := \frac{1}{\sqrt{2\pi}}\int_x^{+\infty}(g-x)^2\exp\Big(-\frac{g^2}{2}\Big)\,dg$$
$$= \frac{1}{\sqrt{2\pi}}\int_x^{+\infty} g^2\exp\Big(-\frac{g^2}{2}\Big)\,dg \quad (A_1(x))$$
$$- \frac{2x}{\sqrt{2\pi}}\int_x^{+\infty} g\exp\Big(-\frac{g^2}{2}\Big)\,dg \quad (-A_2(x))$$
$$+ \frac{x^2}{\sqrt{2\pi}}\int_x^{+\infty}\exp\Big(-\frac{g^2}{2}\Big)\,dg \quad (A_3(x))$$
$$=: A_1(x) - A_2(x) + A_3(x)\,,$$

and

$$B(x) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}(g-x)^2\exp\Big(-\frac{g^2}{2}\Big)\,dg$$
$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} g^2\exp\Big(-\frac{g^2}{2}\Big)\,dg \quad (B_1(x))$$
$$- \frac{2x}{\sqrt{2\pi}}\int_{-\infty}^{x} g\exp\Big(-\frac{g^2}{2}\Big)\,dg \quad (-B_2(x))$$
$$+ \frac{x^2}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\Big(-\frac{g^2}{2}\Big)\,dg \quad (B_3(x))$$
$$=: B_1(x) - B_2(x) + B_3(x)\,.$$

Using symmetry arguments for even and odd functions, it can be shown that $B_1(x) = A_1(-x)$, $B_2(x) = -A_2(x)$, and $B_3(x) = A_3(-x)$. Therefore,

$$A(x) = \big(A_1(x) + A_3(x)\big) - A_2(x) \quad (108)$$
$$B(x) = \big(A_1(-x) + A_3(-x)\big) + A_2(x)\,. \quad (109)$$

Next, we compute expressions for $A_1(x) + A_3(x)$ and $A_2(x)$. Integrating $A_1(x)$ by parts, we obtain:

$$A_1(x) + A_3(x) = \frac{x}{\sqrt{2\pi}}\exp\Big(-\frac{x^2}{2}\Big) + (1+x^2)\underbrace{\frac{1}{\sqrt{2\pi}}\int_x^{+\infty}\exp\Big(-\frac{g^2}{2}\Big)\,dg}_{:=Q(x)}, \quad (110)$$

where $Q(x)$ is the Q-function, defined in (49). The integral in $A_2(x)$ can be computed in closed-form as

$$A_2(x) = \frac{2x}{\sqrt{2\pi}}\exp\Big(-\frac{x^2}{2}\Big)\,. \quad (111)$$

From (108), (109), (110), and (111), we obtain

$$A(x) = -\frac{x}{\sqrt{2\pi}}\exp\Big(-\frac{x^2}{2}\Big) + (1+x^2)Q(x) \quad (112)$$
$$B(x) = \frac{x}{\sqrt{2\pi}}\exp\Big(-\frac{x^2}{2}\Big) + (1+x^2)Q(-x)\,. \quad (113)$$

Using (107), (112), and (113), and the property $Q(x) = 1 - Q(-x)$, for all x, we obtain

$$\mathbb{E}_g\Big[\operatorname{dist}\big(g, \mathcal{I}(a,b)\big)^2\Big] = (a-b)\varphi(a-b) - (a+b)\varphi(a+b) + \big[1+(a+b)^2\big]Q(a+b) + \big[1+(a-b)^2\big]\big[1-Q(a-b)\big]\,,$$

where we used $\varphi(x) = \exp\big(-x^2/2\big)/\sqrt{2\pi}$ [cf. (50)]. This is exactly (52).

B. Part II) Bounds

Showing (53) is relatively simple: either set $b = 0$ in (52), or just use the linearity of the expected value and the fact that g has zero mean and unit variance:

$$\begin{aligned}\mathbb{E}_g\big[\mathrm{dist}(g,a)^2\big] &= \mathbb{E}_g\big[(a-g)^2\big] = \mathbb{E}_g\big[a^2 - 2ag + g^2\big] \\ &= a^2 + 1\,.\end{aligned}$$

We will now focus on proving cases 2), 3), and 4), which are characterized by $b > 0$ and $|a| \neq b$. These will follow by using bounds on the Q-function. The following bounds, valid for $x > 0$, are sharp for large x [62, Eq. 2.121]: [9]

$$\frac{x}{1+x^2}\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{x^2}{2}\Big) \leq Q(x) \leq \frac{1}{x}\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{x^2}{2}\Big)\,. \tag{114}$$

Now we compute bounds for $A(x)$ and $B(x)$ based on (114) and address the cases $x < 0$ and $x > 0$ separately. We will again use the property $Q(x) = 1 - Q(-x)$. Let us start with $A(x)$. Consider $x < 0$. Then,

$$\begin{aligned}A(x) &= -x\,\varphi(x) + (1+x^2)(1-Q(-x)) \\ &\leq -x\,\varphi(x) + (1+x^2)\bigg(1 + \frac{x}{1+x^2}\varphi(x)\bigg) \\ &= 1 + x^2\,,\end{aligned} \tag{115}$$

where the inequality is due to the lower bound in (114). Now, let $x > 0$. Applying the upper bound in (114) directly, we obtain

$$A(x) \leq -x\varphi(x) + (1+x^2)\frac{1}{x}\varphi(x) = \frac{1}{x}\varphi(x)\,. \tag{116}$$

Now consider $B(x)$ with $x < 0$. Since $Q(-x)$ has a positive argument, we use the upper bound in (114):

$$B(x) \leq x\,\varphi(x) + (1+x^2)\Big(-\frac{1}{x}\varphi(x)\Big) = \frac{1}{|x|}\varphi(x)\,, \tag{117}$$

where, in the inequality, we used the fact that $\varphi(-x) = \varphi(x)$. Assume now $x > 0$. Then,

$$\begin{aligned}B(x) &= x\,\varphi(x) + (1+x^2)(1-Q(x)) \\ &\leq x\,\varphi(x) + (1+x^2)\bigg(1 - \frac{x}{1+x^2}\varphi(x)\bigg) \\ &= 1 + x^2\,,\end{aligned} \tag{118}$$

[9] The lower bound in [62, Eq. 2.121] is actually $((x^2-1)/x^3)\varphi(x)$. The lower bound in (114), however, is tighter and stable near the origin. We found this bound in [63]. Since we were not able to track it to a published reference, we replicate the proof from [63] here. For $x > 0$, there holds

$$\begin{aligned}\Big(1+\frac{1}{x^2}\Big)Q(x) &= \int_x^\infty \Big(1+\frac{1}{x^2}\Big)\varphi(u)\,du \geq \int_x^\infty \Big(1+\frac{1}{u^2}\Big)\varphi(u)\,du \\ &= -\int_x^\infty \frac{u\,d\varphi(u)/du - \varphi(u)}{u^2}\,du = -\int_x^\infty \frac{d}{du}\Big(\frac{\varphi(u)}{u}\Big)\,du = \frac{\varphi(x)}{x}\,,\end{aligned}$$

from which the bound follows. In the third step, we used the property $d\varphi(u)/du = -u\varphi(u)$.

where we used the lower bound in (114). In sum, (115), (116), (117), and (118) tell us that

$$A(x) \leq \begin{cases} 1+x^2 & ,\ x<0 \\ \frac{1}{x}\varphi(x) & ,\ x>0 \end{cases} \tag{119}$$

$$B(x) \leq \begin{cases} \frac{1}{|x|}\varphi(x) & ,\ x<0 \\ 1+x^2 & ,\ x>0 \end{cases}\,. \tag{120}$$

From (107), (119), and (120),

$$\begin{aligned}\mathbb{E}_g\Big[\mathrm{dist}\big(g,\mathcal{I}(a,b)\big)^2\Big] &= A(a+b) + B(a-b) \\ &\leq \begin{cases} \frac{\varphi(a+b)}{a+b} + \frac{\varphi(a-b)}{|a-b|} & ,\ |a|<b \\ 1 + (a+b)^2 + \frac{\varphi(a-b)}{|a-b|} & ,\ a+b<0 \\ \frac{\varphi(a+b)}{a+b} + 1 + (a-b)^2 & ,\ a-b>0\,. \end{cases}\end{aligned}$$

Taking into account that $\varphi(x) = \varphi(-x)$ for any x, this is exactly (54), (55), and (56).

We now address cases 5) and 6). Suppose $a + b = 0$. Since $a - b < 0$ (recall that $b > 0$), (117) applies and tells us that $B(a-b) \leq \varphi(a-b)/(b-a)$. Setting $x = 0$ in (112), we obtain $A(a+b) = A(0) = Q(0) = 1/2$. Therefore, $A(a+b) + B(a-b) = \varphi(a-b)/(b-a) + 1/2$, which is (57). The proof of (58) is identical. □

Appendix B
Proof of Lemma 18

Denote $f(x) := (1 - 1/x)/\sqrt{\log x}$. It can be shown that

$$\begin{aligned}\frac{d}{dx}f(x) &= \frac{2\log x + 1 - x}{2x^2\log^{3/2} x} \\ \frac{d^2}{dx^2}f(x) &= \frac{3(x-1) - 8\log^2 x + 2(x-3)\log x}{4x^3\log^{5/2} x}\,.\end{aligned}$$

The stationary points of f are those for which $\frac{d}{dx}f(x) = 0$, that is, the points that satisfy the equation $2\log x = x - 1$. This equation has only one solution, say $\overline{x}$, for $x > 1$: $\log\overline{x} = (\overline{x}-1)/2$. Using this identity, we can conclude that

$$\begin{aligned}\frac{d^2}{dx^2}f(\overline{x}) &= \frac{3(\overline{x}-1) - 2(\overline{x}-1)^2 + (\overline{x}-3)(\overline{x}-1)}{\frac{1}{\sqrt{2}}\overline{x}^3(\overline{x}-1)^{5/2}} \\ &= \frac{2-\overline{x}}{\frac{1}{\sqrt{2}}\overline{x}^3(\overline{x}-1)^{3/2}} < 0\,,\end{aligned}$$

since $\overline{x} > 2$. This is because $\log 2 > 1/2$ and, e.g., $\log 11 < 5$ or, in other words, $(x-1)/2$ intersects $\log x$ somewhere in the interval $2 < x < 11$, that is, $2 < \overline{x} < 11$. This means that the only stationary point $\overline{x}$ is a local maximum. Since $\lim_{x\downarrow 1} f(x) = 0$ and $\lim_{x\to+\infty} f(x) = 0$ (using for example l'Hôpital's rule), $\overline{x}$ is actually a global maximum. Knowing that $\overline{x}$ satisfies $\log\overline{x} = (\overline{x}-1)/2$, we have

$$f(\overline{x}) = \frac{\overline{x}-1}{\overline{x}\sqrt{\log\overline{x}}} = \sqrt{2}\frac{\overline{x}-1}{\overline{x}\sqrt{\overline{x}-1}} = \sqrt{2}\frac{\sqrt{\overline{x}-1}}{\overline{x}}\,.$$

By equating the derivative of the function $\sqrt{x-1}/x$ to zero, we know that it achieves its maximum at $x = 2$. Therefore, $f(\overline{x}) \leq \sqrt{2}/2 = 1/\sqrt{2}$. Dividing by $1/\sqrt{\pi}$, we obtain (59). □

APPENDIX C
PROOF OF LEMMA 19

A. Proof of 1)

According to (62), $0 \in \partial f_1^{(i)}(x_i^\star)$ is equivalent to either:

$$i \in IJ \text{ and } \operatorname{sign}(x_i^\star) + \beta \operatorname{sign}(x_i^\star - w_i) = 0\,, \text{or} \quad (121a)$$

$$i \in IJ^c \text{ and } \beta \geq 1\,, \text{or} \quad (121b)$$

$$i \in I^cJ \text{ and } \beta \leq 1\,, \text{or} \quad (121c)$$

$$i \in I^cJ^c\,. \quad (121d)$$

Note that (121a) cannot be satisfied whenever $\beta \neq 1$. Hence, conditions (121a)-(121d) can be rewritten as

- $\beta = 1$: $\operatorname{sign}(x_i^\star) + \operatorname{sign}(x_i^\star - w_i) = 0$ for $i \in IJ$, or $i \in I^cJ$, or $i \in IJ^c$, or $i \in I^cJ^c$.
- $\beta > 1$: $i \in IJ^c$ or $i \in I^cJ^c$.
- $\beta < 1$: $i \in I^cJ$ or $i \in I^cJ^c$.

We consider two scenarios: $IJ \neq \emptyset$ and $IJ = \emptyset$.

- Let $IJ \neq \emptyset$. When $\beta = 1$, $0 \notin \partial f_1(x^\star)$ if and only if there is an $i \in IJ$ such that $\operatorname{sign}(x_i^\star) + \operatorname{sign}(x_i^\star - w_i) \neq 0$, i.e., there is at least one bad component: $\overline{h} > 0$. When $\beta \neq 1$, there is at least one $i \in IJ$ for which (121a) is not satisfied, that is, $0 \notin \partial f_1(x^\star)$. We thus conclude that part 1) is true whenever $IJ \neq \emptyset$.
- Let $IJ = \emptyset$ or, equivalently, $x_i^\star = w_i$ for all $i \in I$. Recall from (20) that $|IJ| = h + \overline{h}$. Thus, $IJ = \emptyset$ implies $\overline{h} = 0$. In this case, if $\beta = 1$, then $0 \in \partial f_1(x^\star)$. On the other hand, for $\beta > 1$, $0 \notin \partial f_1(x^\star)$ if and only if $I^cJ \neq \emptyset$; similarly, for $\beta < 1$, $0 \notin \partial f_1(x^\star)$ if and only if $IJ^c \neq \emptyset$. We next show that $IJ = \emptyset$, together with $I \neq \emptyset$ and $J \neq \emptyset$, implies that both I^cJ and IJ^c are nonempty, thus showing that part 1) is also true whenever $IJ = \emptyset$. In fact, $I \neq \emptyset$ implies $IJ^c \neq \emptyset$, because $I = IJ \cup IJ^c = IJ^c$. Also, $J \neq \emptyset$, that is, $x^\star \neq w$, implies $I^cJ \neq \emptyset$. This is because $IJ = \emptyset$ means that $x^\star$ and w coincide on I, and $I^cJ = \{i : 0 = x_i^\star \neq w_i\}$ is the set of nonzero components of w outside I. Since $x^\star$ and w coincide on I, they have to differ outside I, i.e., $I^cJ \neq \emptyset$.

B. Proof of 2)

From (63), $0 \in \partial f_2^{(i)}(x_i^\star)$ is equivalent to either

$$i \in IJ \text{ and } \beta(w_i - x_i^\star) = \operatorname{sign}(x_i^\star)\,, \text{or}$$
$$i \in IJ^c\,, \text{or}$$
$$i \in I^c \text{ and } \beta \leq 1/|w_i|\,.$$

□

REFERENCES

[1] J. F. C. Mota, N. Deligiannis, and M. R. D. Rodrigues, "Compressed sensing with side information: Geometrical interpretation and performance bounds," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 675–679.

[2] D. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Info. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[4] H. Nyquist, "Certain topics in telegraph transmission theory," *Trans. American Institute of Electrical Engineers*, vol. 47, no. 2, pp. 617–644, 1928.

[5] C. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[6] M. Lustig, D. Donoho, and J. Pauly, "Sparse MRI: The application of compressed sensing to rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, pp. 1182–1195, 2007.

[7] R. Baraniuk and P. Steeghs, "Compressive radar imaging," in *IEEE Radar Conf.*, 2007, pp. 128–133.

[8] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Sig. Proc. Mag.*, vol. 25, no. 2, pp. 83–91, 2008.

[9] J. Haupt, W. Bajwa, M. Rabbat, and R. Nowak, "Compressed sensing for networked data," *IEEE Sig. Proc. Mag.*, vol. 25, no. 2, pp. 92–101, 2008.

[10] R. von Borries, C. Miosso, and C. Potes, "Compressed sensing using prior information," in *IEEE Int. Workshop Comput. Advances Multi-Sensor Adaptive Proc.*, 2007, pp. 121–124.

[11] G.-H. Chen, J. Tang, and S. Leng, "Prior image constrained compressed sensing (PICCS): a method to accurately reconstruct dynamic CT images from highly undersampled projection data sets," *Med Phys.*, vol. 35, no. 2, pp. 660–663, 2008.

[12] N. Vaswani and W. Lu, "Modified-CS: Modifying compressive sensing for problems with partially known support," *IEEE Trans. Signal Processing*, vol. 58, no. 9, pp. 4595–4607, 2010.

[13] T. Tanaka and J. Raymond, "Optimal incorporation of sparsity information by weighted ℓ_1 optimization," in *IEEE Intern. Symposium on Information Theory (ISIT)*, 2010, pp. 1598–1602.

[14] M. A. Khajehnejad, W. Xu, A. Avestimehr, and B. Hassibi, "Analyzing weighted ℓ_1 minimization for sparse recovery with nonuniform sparse models," *IEEE Trans. Signal Processing*, vol. 59, no. 5, pp. 1985–2001, 2011.

[15] S. Oymak, M. A. Khajehnejad, and B. Hassibi, "Recovery threshold for optimal weight ℓ_1 minimization," in *IEEE Intern. Symposium on Information Theory (ISIT)*, 2012, pp. 2032–2036.

[16] J. Scarlett, J. Evans, and S. Dey, "Compressed sensing with prior information: Information-theoretic limits and practical decoders," *IEEE Trans. Signal Processing*, vol. 61, no. 2, pp. 427–439, 2013.

[17] M. P. Friedlander, H. Mansour, R. Saab, and O. Yilmaz, "Recovering compressively sampled signals using partial support information," *IEEE Trans. Info. Theory*, vol. 58, no. 2, pp. 1122–1134, 2012.

[18] K. Mishra, M. Cho, A. Kruger, and W. Xu, "Off-the-grid spectral compressed sensing with prior information," in *IEEE Intern. Conf. Acoustics, Speech, and Sig. Processing (ICASSP)*, 2014, pp. 1010–1014.

[19] H. Mansour and R. Saab, "Recovery analysis for weighted ℓ_1-minimization using a null space property," *Applied and Computational Harmonic Analysis*, 2015.

[20] L. Weizman, Y. C. Eldar, and D. Ben Bashat, "Compressed sensing for longitudinal MRI: An adaptive-weighted approach," *Medical Physics*, vol. 42, no. 9, pp. 5195–5207, 2015.

[21] V. Stanković, L. Stanković, and S. Cheng, "Compressive image sampling with side information," in *IEEE Intern. Conf. Image Processing (ICIP)*, 2009, pp. 3037–3040.

[22] M. Rostami, "Compressed sensing in the presence of side information," Master's thesis, University of Waterloo, 2012.

[23] X. Wang and J. Liang, "Side information-aided compressed sensing reconstruction via approximate message passing," 2013, preprint: http://arxiv.org/abs/1311.0576v1.

[24] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Info. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[25] E. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus de l'Academie des Sciences (Paris), Série I*, no. 346, pp. 589–592, 2008.

[26] E. Candès and M. Wakin, "An introduction to compressive sampling," *IEEE Sig. Proc. Mag.*, vol. 25, no. 2, pp. 21–30, 2008.

[27] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky, "The convex geometry of linear inverse problems," *Found. Computational Mathematics*, vol. 12, pp. 805–849, 2012.

[28] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1998.

[29] E. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Info. Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[30] E. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, pp. 969–985, 2007.

[31] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[32] 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.

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

[34] M. Stojnic, "Various thresholds for ℓ_1-optimization in compressed sensing," 2009, preprint: http://arxiv.org/abs/0907.3666.

[35] B. Bah and J. Tanner, "Improved bounds on restricted isometry constants for Gaussian matrices," *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 5, pp. 2882–2898, 2010.

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

[37] L.-W. Kang and C.-S. Lu, "Distributed compressive video sensing," in *IEEE Intern. Conf. Acoustics, Speech, and Sig. Processing (ICASSP)*, 2009, pp. 1169–1172.

[38] A. Sankaranarayanan, C. Studer, and R. Baraniuk, "CS-MUVI: video compressive sensing for spatial multiplexing cameras," in *Intern. Conf. Computation Photography*, 2012, pp. 1–10.

[39] A. C. Sankaranarayanan, L. Xu, C. Studer, Y. Li, K. Kelly, and R. G. Baraniuk, "Video compressive sensing for spatial multiplexing cameras using motion-flow models," 2014, under review at SIAM J. Imaging Sciences.

[40] J. F. C. Mota, N. Deligiannis, A. C. Sankaranarayanan, V. Cevher, and M. R. D. Rodrigues, "Dynamic sparse state estimation using ℓ_1-ℓ_1 minimization: Adaptive-rate measurement bounds, algorithms, and applications," in *IEEE Intern. Conf. Acoustics, Speech, and Sig. Processing (ICASSP)*, 2015, pp. 3332–3336.

[41] J. F. C. Mota, N. Deligiannis, A. C. Sankaranarayanan, V. Cevher, and M. Rodrigues, "Adaptive-rate reconstruction of time-varying signals with application in compressive foreground extraction," *IEEE Trans. Signal Processing*, vol. 64, no. 14, pp. 3651–3666, 2016.

[42] A. Charles, M. Asif, J. Romberg, and C. Rozell, "Sparsity penalties in dynamical system estimation," in *IEEE Conf. Information Sciences and Systems*, 2011, pp. 1–6.

[43] J. Ziniel, L. C. Potter, and P. Schniter, "Tracking and smoothing of time-varying sparse signals via approximate belief propagation," in *Asilomar Conf. Signals, Systems, and Computers*, 2010, pp. 808–812.

[44] H. Jung and J. C. Ye, "Motion estimated and compensated compressed sensing dynamic magnetic resonance imaging: What we can learn from video compression techniques," *Intern. Journal of Imaging Systems and Technology*, vol. 20, no. 2, pp. 81–98, 2010.

[45] D. Baron, M. Wakin, M. Duarte, S. Sarvotham, and R. Baraniuk, "Distributed compressed sensing," ECE06-12, Electrical and Computer Engineering Dept., Rice University, Tech. Rep., 2006.

[46] M. Trocan, T. Maugey, J. Fowler, and B. Pesquet-Popescu, "Disparity-compensated compressed-sensing reconstruction for multiview images," in *IEEE Intern. Conf. Multimedia and Expo (ICME)*, 2010, pp. 1225–1229.

[47] J. F. C. Mota, L. Weizman, N. Deligiannis, Y. C. Eldar, and M. R. D. Rodrigues, "Reference-based compressed sensing: A sample complexity approach," in *IEEE Intern. Conf. Acoustics, Speech, and Sig. Processing (ICASSP)*, 2016, pp. 4687–4691.

[48] ——, "Reweighted ℓ_1-norm minimization with guarantees: An incremental measurement approach to sparse reconstruction," 2017, accepted at Signal Processing with Adaptive Sparse Structured Representations (SPARS).

[49] P. Song, J. F. C. Mota, N. Deligiannis, and M. R. D. Rodrigues, "Measurement matrix design for compressive sensing with side information at the encoder," in *IEEE Statistical Signal Processing Workshop (SSP)*, 2016, pp. 1–5.

[50] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Info. Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[51] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Info. Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976.

[52] S. Oymak, C. Thrampoulidis, and B. Hassibi, "The squared-error of generalized LASSO: A precise analysis," 2013, preprint: http://arxiv.org/abs/1311.0830.

[53] D. Donoho and J. Tanner, "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing," *Phil. Trans. R. Soc. A*, vol. 367, pp. 4273–4293, 2009.

[54] D. L. Donoho and J. Tanner, "Precise undersampling theorems," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 913–924, 2010.

[55] D. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[56] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Info. Theory*, vol. 57, no. 2, pp. 764–785, 2011.

[57] D. L. Donoho, A. Maleki, and A. Montanari, "The noise-sensitivity phase transition in compressed sensing," *IEEE Trans. Info. Theory*, vol. 57, no. 10, pp. 6920–6941, 2011.

[58] Y. Gordon, "On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$," in *Geometric Aspects of Functional Analysis, Israel Seminar 1986-1987. Lecture Notes in Mathematics*, 1988, pp. 84–106.

[59] J. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*. Springer, 2004.

[60] D. Amelunxen, M. Lotz, M. Mccoy, and J. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Information and Inference: A Journal of the IMA*, pp. 1–71, 2014.

[61] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[62] J. Wozencraft and I. Jacobs, *Principles of Communication Engineering*. John Wiley & Sons, 1965.

[63] Wikipedia, "Q-function," https://en.wikipedia.org/wiki/Q-function, retrieved May 21, 2014.

João F. C. Mota received the M.Sc. degree and the Ph.D. degree in Electrical and Computer Engineering from the Technical University of Lisbon, Portugal, in 2008 and 2013, respectively. He also received the Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University, PA, USA, in 2013.

From 2013 to 2016, he was Senior Research Associate at University College London, London, U.K. In 2017, he became Assistant Professor in the School of Engineering and Physical Sciences at Heriot-Watt University, Edinburgh, U.K., where he is also affiliated with the Institute of Sensors, Signals, and Systems.

His current research interests include theoretical and practical aspects of high-dimensional data processing, inverse problems, optimization theory, machine learning, data science, and distributed information processing and control. He was the recipient of the 2015 IEEE Signal Processing Society Young Author Best Paper Award for the paper "Distributed Basis Pursuit", published in IEEE Transactions on Signal Processing.

Nikos Deligiannis received the Diploma degree in electrical and computer engineering from the University of Patras, Greece, in 2006, and the Ph.D. degree (Hons.) in applied sciences from Vrije Universiteit Brussel, Belgium, in 2012. He is currently an Assistant Professor with the Electronics and Informatics Department, Vrije Universiteit Brussel.

From 2012 to 2013, he was a Post-Doctoral Researcher with the Department of Electronics and Informatics, Vrije Universiteit Brussel. From 2013 to 2015, he was a Senior Researcher with the Department of Electronic and Electrical Engineering, University College London, U.K., and also a Technical Consultant on Big Visual Data Technologies with the British Academy of Film and Television Arts, U.K.

His current research interests include big data processing and analysis, machine learning, Internet-of-Things networks, and distributed signal processing. He has authored over 80 journal and conference publications, book chapters, and two patent applications (one owned by iMinds, Belgium and the other by BAFTA, U.K.). He was a recipient of the 2011 ACM/IEEE International Conference on Distributed Smart Cameras Best Paper Award and the 2013 Scientific Prize FWO-IBM Belgium.

Miguel R. D. Rodrigues is currently a Reader in information theory and processing with the Department of Electronic and Electrical Engineering, University College London, U.K. He was with the University of Porto, Portugal, rising through the ranks from Assistant to Associate Professor, where he was also the Head of the Information Theory and Communications Research Group, Instituto de Telecomunicações Porto.

He received the Licenciatura degree in electrical and computer engineering from the University of Porto, Portugal, and the Ph.D. degree in electronic and electrical engineering from the University College London, U.K. He has also held post-doctoral positions and visiting appointments with various institutions worldwide including University College London, Cambridge University, Princeton University, and Duke University from 2003 to 2016.

His research interests are in the general areas of information theory and processing with a current focus on sensing, analysis and processing of high-dimensional data. His work, which has led to over 150 papers in the leading international journals and conferences in the field, has also been honored with the Prestigious IEEE Communications and Information Theory Societies Joint Paper Award 2011.

Dr. Rodrigues was a recipient of fellowships from the Portuguese Foundation for Science and Technology and the Foundation Calouste Gulbenkian. He has served as a Co-Chair of the Technical Program Committee of the IEEE Information Theory Workshop 2016 and also as a Co-Organizer of the Workshop on Sensing and Analysis of High-Dimensional Data 2014 and 2015. He currently serves as an Associate Editor of the IEEE Communications Letters.