Exploiting Prior Information in Block-sparse Signals

Sajad Daei, Farzan Haddadi, Arash Amini

***Abstract*—We study the problem of recovering a block-sparse signal from under-sampled observations. The non-zero values of such signals appear in few blocks, and their recovery is often accomplished using an $\ell_{1,2}$ optimization problem. In applications such as DNA micro-arrays, some extra information about the distribution of non-zero blocks is available; i.e., the number of non-zero blocks in certain subsets of the blocks is known. A typical way to consider the extra information in recovery procedures is to solve a weighted $\ell_{1,2}$ problem. In this paper, we consider a block-sparse model which is accompanied with a partitioning of the blocks; besides the overall block-sparsity level of the signal, we assume to know the block-sparsity of each subset in the partition. Our goal in this work is to minimize the number of required linear measurements for perfect recovery of the signal by tuning the weights of a weighted $\ell_{1,2}$ problem. For this goal, we apply tools from conic integral geometry and derive closed-form expressions for the optimal weights. We show through precise analysis and simulations that the weighted $\ell_{1,2}$ problem with optimal weights significantly outperforms the regular $\ell_{1,2}$ problem. We further show that the optimal weights are robust against the inaccuracies of prior information.**

***Index Terms*—block-sparse, prior information, weighted $\ell_{1,2}$, conic integral geometry.**

I. Introduction

COMPRESSED sensing (CS) aims at recovering a sparse signal $\boldsymbol{x} \in \mathbb{R}^n$ from a few random linear measurements

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is called the measurement matrix, $\boldsymbol{e}$ is the noise vector which is assumed to be i.i.d. Gaussian with variance σ_e^2, and m is much smaller than n. In this work, the signal is assumed to be block-sparse which occurs in many applications such as DNA micro-arrays [1], direction of arrival (DOA) estimation [2], computational neuroscience [3], and multiple measurement vector (MMV) problem [4]. Consider a block-sparse signal $\boldsymbol{x} \in \mathbb{R}^n$ which is a concatenation of q blocks $\mathcal{V}_b, b = 1, ..., q$. The block support is defined as the index set of non-zero blocks, which we denote by $\mathcal{B}$. Figure 1 shows an example of a block-sparse signal composed of $q = 10$ blocks of length $k = 10$. The block support in this figure is the index set $\{3, \dots, 7\}$; i.e., the signal is 5-block-sparse and contains zero blocks except for $\{\mathcal{V}_3, \dots, \mathcal{V}_7\}$. Ideally, a block-sparse signal is reconstructed using the following optimization problem:

$$\begin{aligned}\mathsf{P}_{0,2}^{\eta} : \min_{\boldsymbol{z}\in\mathbb{R}^n} \|\boldsymbol{z}\|_{0,2} &:= \sum_{b=1}^{q} \mathbf{1}_{\|\boldsymbol{z}_{\mathcal{V}_b}\|_2>0} \\ \text{s.t.} \quad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}\|_2 &\leqslant \eta,\end{aligned} \tag{2}$$

where $\mathbf{1}_{\mathcal{E}}$ denotes the indicator function of the event $\mathcal{E}$, and η is an upper-bound for $\|\boldsymbol{e}\|_2$. This problem is computationally intractable in polynomial time and in general, it is NP-hard. Following Donoho [5], $\mathsf{P}_{0,2}^{\eta}$ can be relaxed as an $\ell_{1,2}$ minimization of the form

$$\begin{aligned}\mathsf{P}_{1,2}^{\eta} : \min_{\boldsymbol{z}\in\mathbb{R}^n} \|\boldsymbol{z}\|_{1,2} &:= \sum_{b=1}^{q} \|\boldsymbol{z}_{\mathcal{V}_b}\|_2 \\ \text{s.t.} \quad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}\|_2 &\leqslant \eta.\end{aligned} \tag{3}$$

It is known that $\mathsf{P}_{1,2}^{\eta}$ for $\eta = 0$ (noiseless case) finds the original block-sparse signal with high probability if $\boldsymbol{A}$ comes from a probability law that distributes the null space uniformly with respect to the Haar measure [6]. This includes Gaussian ensembles and partial Fourier matrices. We also assume the same type of random matrices in this paper.

The main challenge in recovering a block-sparse signal is to find the block support. Intuitively, if there exists an additional piece of information about the block-sparsity level of certain subsets of the blocks[1], one can probably solve $\mathsf{P}_{1,2}^{0}$ and $\mathsf{P}_{1,2}^{\eta}$ with fewer measurements or smaller reconstruction error, respectively (see Subsection I-C for more explanations). In some scenarios, such extra information is available. For example, in DNA microarray one might know the occurrence frequency of specific Genes in certain sets of arrays. In computational neuroscience, the behavior of neurons exhibit non-uniform clustered responses [3]. In radar signal processing, an operator might know the range of speed with which an aircraft flies or the range of angular domain it forms in the radar [7]. A typical way to consider such prior information is to apply a weighted $\ell_{1,2}$ minimization formulated as:

$$\begin{aligned}\mathsf{P}_{1,2,\boldsymbol{w}}^{\eta} : \min_{\boldsymbol{z}\in\mathbb{R}^n} \|\boldsymbol{z}\|_{1,2,\boldsymbol{w}} &:= \sum_{b=1}^{q} w_b\|\boldsymbol{z}_{\mathcal{V}_b}\|_2 \\ \text{s.t.} \quad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}\|_2 &\leqslant \eta,\end{aligned} \tag{4}$$

where $\boldsymbol{w} = [w_1, ..., w_q]^T$. This work explicitly takes prior block information into account by optimally tuning the weights in $\mathsf{P}_{1,2,\boldsymbol{w}}^{\eta}$. Specifically, the main objective of this paper is to find optimal weights in $\mathsf{P}_{1,2,\boldsymbol{w}}^{\eta}$ in the sense that they minimize the reconstruction error for $\eta > 0$, and the required number of measurements for $\eta = 0$.

A. Contributions

Although block-sparse signals are widely used in many applications in the literature (see Subsection I-B), the inclusion of prior information about the block-support distribution in improving the performance of the signal recovery is not

[1]Our block-sparse model is deterministic in this paper; the normalized block-sparsity of a subset of blocks (the number of nonzero blocks divided by the total number of blocks in the subset) is occasionally referred to as the non-zero probability or non-zero likelihood of the blocks in this subset.

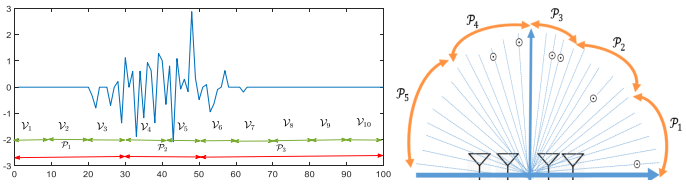

Figure 1: Left image: Illustration of a non-uniform block-sparse signal $\boldsymbol{x} \in \mathbb{R}^n$ with parameters $n = 100$, $q = 10$, $k = 10$, $L = 3$, $\alpha_1 = \frac{1}{3}$, $\alpha_2 = 1$ and $\alpha_3 = \frac{2}{5}$. Right image: Schematic diagram of DOA estimation of far-field sources. The angular half-space is divided into $q = 30$ angular clusters $\{\mathcal{V}_b\}_{b=1}^{q}$ with equal length. $L = 5$ block support estimators with accuracies $\alpha_1 = \frac{1}{6}$, $\alpha_2 = \frac{1}{6}$, $\alpha_3 = \frac{1}{2}$, $\alpha_4 = \frac{1}{3}$ and $\alpha_5 = 0$ are considered.

analyzed yet. In this work, we first consider the standard block-sparse model without prior information and derive tight and non-asymptotic bounds for the required sample complexity of $\mathsf{P}_{1,2}^{0}$. Interestingly, the asymptotic form of our bound matches the conjecture in [8]. Then, we consider the weighted $\mathsf{P}_{1,2}^{\eta}$ problem denoted by $\mathsf{P}_{1,2,\boldsymbol{w}}^{\eta}$ in presence of prior information for which we present a new method for obtaining the optimal weights. In particular, we assume to know a partitioning of $[q] := \{1, \ldots, q\}$ into L disjoint subsets $\{\mathcal{P}_i\}_{i=1}^{L}$, about which we know the number of non-zero blocks in each index set $\mathcal{P}_i$. In simple words, we assume to know the block-sparsity level of each subset in a partitioning of the blocks. These partial block-sparsity levels are commonly expressed via $\alpha_i := \frac{|\mathcal{P}_i \cap \mathcal{B}|}{|\mathcal{P}_i|}$ which are called the accuracy of $\mathcal{P}_i$s. Obviously, if the partition consists of a single subset $[q]$, we are dealing with a conventional block-sparse model without prior information. Note that the measurement matrix combines all the blocks before generating the measurement vector. Hence, we do not have access to measurements that are solely determined by the blocks of a single partition; otherwise, the recovery problem could be simplified to multiple block-sparse recovery in lower dimensions without prior information. We call signals generated based on this model as *non-uniform* block-sparse signals. The left image of Figure 1 shows a sample signal with this model. The strategy of finding optimal weights is to study the asymptotics of the upper and lower bounds of the statistical dimension of a certain convex cone (these bounds differ from each other in an asymptotically vanishing constant term). In this regard, we prove that our method of obtaining optimal weights is robust to slight inaccuracies of the prior knowledge. With optimal weights in $\mathsf{P}_{1,2,\boldsymbol{w}}^{0}$, we also find that the number of measurements required for exact recovery of a block-sparse vector equals to the sum of the required number of measurements in the simpler (smaller size) problems of recovering each subset via $\mathsf{P}_{1,2}^{0}$ separately.

B. Related works and Key Differences

There are extensive works in CS literature that consider reconstructing low-dimensional signals with prior information; for instance see [9]–[14]. Here, we discuss some of them.

The weighted ℓ_1 minimization for sparse signal recovery is likely initiated by the work [15]. While this work reveals the advantages of weighting, it does not consider any prior information about the distribution of the support. In [16], partial knowledge about the support is assumed as prior information. In this work, some of the standard sparse recovery techniques are modified to benefit from the available prior information; however, none of the methods rely on weighting. A non-uniform sparse structure similar to the one that is considered in this paper, is proposed in [9] to incorporate prior information about the partial sparsity level of a given partitioning of the elements in sparse signals. The authors apply a weighed ℓ_1 minimization approach and find an upper-bound for the failure probability of the recovery method. The bound is derived based on the internal and external angles of a certain weighted polytope. The weights could then be optimized to minimize the resulting upper-bound. Unfortunately, this approach does not yield explicit expressions for the weights; in addition, since the probability bound is not necessarily tight, the optimality of such weights is questionable. As mentioned by the authors, the extension of this approach to block-sparse signals is very sophisticated.

Using a different approach, [10] investigates the same model with weighted ℓ_1 minimization. Indeed, the approach consists of dividing the signal elements into two subsets; the elements of the first set are penalized with a variable weight, while the weights of the second set are kept fixed at 1. If the first set encompasses at least 50% of the signal support, it is shown in [10] that the weighting technique can improve the recovery performance.

The bipartitioning approach of [10] can be considered as having two estimates of the support. The extension to multiple estimates for the support with varying degrees of accuracy is studied in [11]. It is again confirmed that a multi-level weighting scheme can enhance the recovery performance. The tuning of the weights in [11] is achieved based on empirical results rather than theoretical findings.

A systematic approach for determining the optimal weights is presented in [17]. Indeed, the optimal weights are derived by minimizing the number of required measurements for almost sure sparse recovery of a weighted ℓ_1 minimization. The resulting optimal weights are expressed in terms of the accuracy level (normalized partial sparsity level) of the sets. Our paper could be considered as an extension of [17] to the block-sparse setting. However, our approach to derive theoretical results is substantially different from [17]. For instance, deriving the optimal weights (and proving their uniqueness) for the $\mathsf{P}_{1,2,\boldsymbol{w}}^{\eta}$ problem is much more challenging than the weighted ℓ_1 minimization. Moreover, the construction of a tight upper-bound for the sample complexity that leads to closed-form expressions is nontrivial.

In many practical scenarios, the low-dimensional signal of interest has block-sparse structure rather than simple sparse structure (see Section I-C and [9, Section I] for illustrative examples). In this work, we consider a non-uniform block-sparse model where we have a partitioning of the blocks and their associated accuracy levels. This model extends the non-uniform sparse model considered in [10], [17] to the non-uniform block-sparse signals which is observed in a few works such as [18]–[20]. A distinctive feature of our work, besides the generality of our model, is the closed-form expressions for the optimal weights in the $\mathsf{P}_{1,2,\boldsymbol{w}}^{\eta}$ problem. In contrast to

the result in [10], we show that the minimum 50% accuracy level is not a pre-requisite for $\mathsf{P}^0_{1,2,\boldsymbol{w}}$ to outperform the unweighted $\mathsf{P}^0_{1,2}$ recovery approach. We study the robustness of the weights under inaccuracies in prior knowledge for the first time, and prove robustness for the introduced optimal weights. The study of robustness has not been accomplished in the past even for the easier sparse models.

C. DOA estimation

DOA estimation is the problem of finding the direction of a few sources from observations on an array of sensors. The location of each source is characterized by the direction of arrival $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ with respect to the array axis. Let $\theta_i := \pi(\frac{i-1}{q-1} - \frac{1}{2}), i = 1, ..., q$ be a discretization of the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$ into q grid points representing the potential direction of the sources (see the right block of Figure 1). Assume that $s \ll q$ far-field narrow-band sources with wavelength λ whose directions exactly match the grid angles incident on a q-element uniform linear array (ULA) of sensors with inter-sensor spacing d. Form the so-called array manifold matrix $\boldsymbol{F} = [\boldsymbol{a}(\theta_1), ..., \boldsymbol{a}(\theta_q)] \in \mathbb{C}^{q\times q}$ where $\boldsymbol{a}(\theta_l) := \mathrm{e}^{\mathrm{j}2\pi(d/\lambda)[0,...,q-1]^T \sin(\theta_l)}$ is the steering vector caused by the propagation delay from ith potential source to each of the q sensors. The observed signal at ULA elements at time t are given by

$$\boldsymbol{y}(t) = \boldsymbol{F}\,\boldsymbol{x}(t) + \boldsymbol{e}(t), \tag{5}$$

where $x_i(t) = \sqrt{p_i}\mathrm{e}^{\mathrm{j}2\pi it}$ is the amplitude of the ith potential source, $\boldsymbol{y}(t) \in \mathbb{C}^q$ is the received vector at time (or snapshot) t at all sensors, and $\boldsymbol{e}(t)$ is a Gaussian noise term with variance σ_e^2. In practice, the number of ULA sensors is limited due to cost limitations or physical constraints; hence, it is crucial in DOA estimation to exploit time redundancies (taking several snapshots, *i.e.*, $\boldsymbol{y}(t_1)$, ..., $\boldsymbol{y}(t_k)$). With the assumption that the sources remain fixed in all snapshots, the received signals at all k time instances could be expressed as

$$\boldsymbol{Y} := [\boldsymbol{y}(t_1), ..., \boldsymbol{y}(t_k)] = \boldsymbol{F}\boldsymbol{X} + \boldsymbol{E}, \tag{6}$$

where $\boldsymbol{X} = [\boldsymbol{x}(t_1), ..., \boldsymbol{x}(t_k)] \in \mathbb{C}^{q\times k}$ and $\boldsymbol{E}$ is similarly defined. In practical scenarios, we can not directly observe the received signal at all of the sensors. Instead, we have access to the outputs of $m \ll q$ sensors formed by

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{F}\boldsymbol{X} + \widetilde{\boldsymbol{E}} \in \mathbb{C}^{m\times k}, \tag{7}$$

where $\boldsymbol{A} \in \mathbb{C}^{m\times q}$ is the partial identity matrix [2], $\boldsymbol{A}\boldsymbol{F}$ represents the effective measurement matrix and $\widetilde{\boldsymbol{E}} := \boldsymbol{A}\boldsymbol{E} \in \mathbb{C}^{m\times k}$. In practical settings, the number of sources are limited. Thus, $\boldsymbol{X}$ consists mainly of zeros. In addition, if $\boldsymbol{X}$ is vectorized by concatenating the rows, we achieve a block-sparse vector. Indeed, each row forms a block in the vectorized signal. It is common in radar applications that the operator has prior information (previous measurement of recorded history) about the partial number of sources in a given angular partitioning (see $\{\mathcal{P}_i\}_{i=1}^5$ in the right block of Figure 1). More precisely, the radar operator might know the accuracy (α_is in the right block of Figure 1) of specific angular bands. It is of considerable importance to exploit these information in order to minimize the number of required sensors for finding the sources. We will turn back to this problem in Section VII.

The paper is organized as follows: some concepts from convex geometry are reviewed in Section II. The signal model and our methodology are stated in Section III. The number of measurements required for $\mathsf{P}^0_{1,2,\boldsymbol{w}}$ is obtained in Section IV. In Section V, the procedure of finding optimal weights is explained. In Section VI, the robustness of optimal weights with respect to inaccuracies in prior information is discussed. Numerical simulations on synthetic data that support the theory are presented in Section VII. Finally, the paper is concluded in Section VIII.

Notation Throughout the paper, scalars are denoted by lowercase letters, vectors by lowercase boldface letters, and matrices by uppercase boldface letters. The ith element of a vector $\boldsymbol{x}$ is shown either by $x(i)$ or x_i. $[n]$ refers to $\{1, ..., n\}$. We reserve calligraphic uppercase letters for sets (e.g. $\mathcal{S}$). The cardinality of a set $\mathcal{S}$ is shown by $|\mathcal{S}|$. $\overline{\mathcal{S}}$ denotes the complement $[n]\backslash\mathcal{S}$ of a set $\mathcal{S} \subset [n]$. For $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x}_{\mathcal{S}}$ is the subvector in $\mathbb{R}^{|\mathcal{S}|}$ consisting of the entries indexed by $\mathcal{S}$, that is, $(\boldsymbol{x}_S)_i = x_{j_i} \ : \ \mathcal{S} = \{j_i\}_{i=1}^{|\mathcal{S}|}$. In this paper, $\mathbf{1}_{\mathcal{E}}$ denotes the indicator of a set $\mathcal{E}$. The null space of linear operators is denoted by $\mathrm{null}(\cdot)$. Given a vector $\boldsymbol{x} \in \mathbb{R}^n$ and a set $\mathcal{C} \subseteq \mathbb{R}^n$, the set obtained by scaling elements of $\mathcal{C}$ by the elements of $\boldsymbol{x}$ is denoted by $\boldsymbol{x} \odot \mathcal{C}$. The polar $\mathcal{K}^\circ$ of a cone $\mathcal{K} \subset \mathbb{R}^n$ is the set of vectors forming non-acute angles with every vector in $\mathcal{K}$, i.e. $\mathcal{K}^\circ = \{\boldsymbol{v} \in \mathbb{R}^n : \langle \boldsymbol{v}, \boldsymbol{z}\rangle \leqslant 0 \ \forall \boldsymbol{z} \in \mathcal{K}\}$. For a matrix $\boldsymbol{A}$, the operator norm is defined as $\|\boldsymbol{A}\|_{p\to q} = \sup_{\|\boldsymbol{x}\|_p \leqslant 1} \|\boldsymbol{A}\boldsymbol{x}\|_q$. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{x} \leqslant \boldsymbol{y}$ stands for component-wise inequality while $\boldsymbol{x} < \boldsymbol{y}$ denotes component-wise inequality with strict inequality in at least one component. $\mathbb{B}^n_\epsilon$ refers to the ϵ-ball $\mathbb{B}^n_\epsilon = \{\boldsymbol{x} \in \mathbb{R}^n : \ \|\boldsymbol{x}\|_2 \leqslant \epsilon\}$. Lastly, the notation $(a)_+$ stands for the positive part of a, i.e., $\max\{a, 0\}$.

II. Preliminaries

In this section, basic concepts of conic integral geometry are reviewed.

A. Subdifferential

The subdifferential of a proper[3] convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ at $\boldsymbol{x} \in \mathbb{R}^n$ is given by:

$$\partial f(\boldsymbol{x}) := \{\boldsymbol{z} \in \mathbb{R}^n : f(\boldsymbol{y}) \geqslant f(\boldsymbol{x}) + \langle \boldsymbol{z}, \boldsymbol{y} - \boldsymbol{x}\rangle \ : \ \forall \boldsymbol{y} \in \mathbb{R}^n\}. \tag{8}$$

Proposition 1. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ be a proper convex function that is 1-homogeneous, i.e. $f(\alpha\boldsymbol{z}) = |\alpha| f(\boldsymbol{z}) \ : \forall \alpha \in \mathbb{R}$ and sub-additive, i.e. $f(\boldsymbol{x} + \boldsymbol{y}) \leqslant f(\boldsymbol{x}) + f(\boldsymbol{y}) \ : \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, then we have a simpler form of subdifferential given by:*

$$\partial f(\boldsymbol{x}) = \left\{ \boldsymbol{z} \in \mathbb{R}^n : \begin{array}{ll} \langle \boldsymbol{z}, \boldsymbol{x}\rangle = f(\boldsymbol{x}), f^*(\boldsymbol{z}) = 1, & \boldsymbol{x} \neq \boldsymbol{0} \\ f^*(\boldsymbol{z}) \leqslant 1, & \boldsymbol{x} = \boldsymbol{0} \end{array} \right\} \tag{9}$$

[2] This means that only m random rows of the identity matrix $\boldsymbol{I}_{q\times q}$ are chosen.

[3] An everywhere defined function taking values in $(-\infty, \infty]$ with at least one finite value in $(-\infty, \infty)$.

where $f^*(\boldsymbol{z}) = \sup_{f(\boldsymbol{y})\leqslant 1} \langle \boldsymbol{z}, \boldsymbol{y}\rangle$ *is the dual function of* $f(\boldsymbol{z})$.

Proof. See Appendix A.

B. Descent Cones

The descent cone of a proper convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ at point $\boldsymbol{x} \in \mathbb{R}^n$ is the conic hull of the directions from $\boldsymbol{x}$ that f does not increase:

$$\mathcal{D}(f, \boldsymbol{x}) = \bigcup_{t\geqslant 0} \{\boldsymbol{z} \in \mathbb{R}^n : f(\boldsymbol{x} + t\boldsymbol{z}) \leqslant f(\boldsymbol{x})\}\cdot \tag{10}$$

The descent cone of a convex function is a convex set. There is also a relation between decent cone and subdifferential of a convex function [21, Chapter 23] which is given by:

$$\mathcal{D}^\circ(f, \boldsymbol{x}) = \mathrm{cone}(\partial f(\boldsymbol{x})) := \bigcup_{t\geqslant 0} t.\partial f(\boldsymbol{x}). \tag{11}$$

C. Statistical Dimension

Definition 1. Statistical Dimension [6]: Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a convex closed cone. Statistical dimension of $\mathcal{C}$ is defined as:

$$\delta(\mathcal{C}) := \mathbb{E}_{\boldsymbol{g}} \|\mathcal{P}_{\mathcal{C}}(\boldsymbol{g})\|_2^2 = \mathbb{E}_{\boldsymbol{g}} \mathrm{dist}^2(\boldsymbol{g}, \mathcal{C}^\circ), \tag{12}$$

where

$$\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) := \arg\min_{\boldsymbol{z}\in\mathcal{C}} \|\boldsymbol{z} - \boldsymbol{x}\|_2, \tag{13}$$

is the projection of $\boldsymbol{x} \in \mathbb{R}^n$ onto the set $\mathcal{C}$, and $\boldsymbol{g}$ is a vector with i.i.d. elements each following a standard normal distribution.

Statistical dimension is a measure of the size of a convex cone and generalizes the concept of dimension for linear subspaces to the class of convex cones.

D. Inverse Problems via Convex Optimization

Convex optimization is a common approach for recovering a structured signal $\boldsymbol{x}_{n\times 1}$ from m linear measurements of the form (1). Let $f(\cdot)$ be a function that promotes the structure of $\boldsymbol{x}$ (e.g., sparsity). Then, one might be able to recover the signal $\boldsymbol{x}$ by solving the problem

$$\mathsf{P}_f^\eta : \quad \min_{\boldsymbol{z}\in\mathbb{R}^n} \ f(\boldsymbol{z}) \quad \text{s.t.} \ \ \|\boldsymbol{A}\boldsymbol{z} - \boldsymbol{y}\|_2 \leqslant \eta, \tag{14}$$

where η is a known upper-bound on the norm of the noise $\boldsymbol{e}$. For the noiseless case of $\boldsymbol{e} = \boldsymbol{0}$ in (1), it is shown in [6] that P_f^0 exhibits a sharp phase-transition behavior (success/failure) as the number of measurements increases. In addition, the boundary of the transition is determined by the statistical dimension of the decent cone of f at $\boldsymbol{x}$, i.e. $\delta(\mathcal{D}(f, \boldsymbol{x}))$.

In Theorem 1, we present a result regarding the recovery performance of (14) from random measurements, in both noiseless and noisy settings. This theorem is an adaptation of [22, Corollary 3.5] and [6, Theorem 2].

Theorem 1. *Let* $f(\cdot)$ *be a proper convex function that promotes the structure of* $\boldsymbol{x}$*. Let* $\boldsymbol{A}_{m\times n}$ *be a random matrix whose null space is uniformly distributed with respect to the Haar measure. Then, if*

$$m \geqslant \delta(\mathcal{D}(f, \boldsymbol{x})) + \sqrt{8\log(\tfrac{4}{\zeta})n}$$

for some $\zeta \in [0, 1]$*, then* P_f^0 *recovers* $\boldsymbol{x}$ *from* $\boldsymbol{y}_{m\times 1} = \boldsymbol{A}\boldsymbol{x}$ *with probability at least* $1 - \zeta$*. Alternatively, in the noisy case of* $\boldsymbol{y}_{m\times 1} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$*, where* $\|\boldsymbol{e}\|_2 \leqslant \eta$*, if* $\widehat{\boldsymbol{x}}_\eta$ *is any solution of* P_f^η*, then*

$$\|\widehat{\boldsymbol{x}}_\eta - \boldsymbol{x}\|_2 \leqslant \frac{2\eta}{(\sqrt{m-1} - \sqrt{\delta(\mathcal{D}(f,\boldsymbol{x}))} - \zeta)_+}, \tag{15}$$

with probability at least $1 - \mathrm{e}^{-\frac{\zeta^2}{2}}$.

Also in [6], the following error bound for the statistical dimension is provided.

Theorem 2. *[6, Theorem 4.3] For any* $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$*:*

$$0 \leqslant \inf_{t\geqslant 0} \mathbb{E}\mathrm{dist}^2(\boldsymbol{g}, t\partial f(\boldsymbol{x})) - \delta(\mathcal{D}(f, \boldsymbol{x})) \leqslant \frac{2\sup_{s\in\partial f(\boldsymbol{x})} \|s\|_2}{f(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2})}. \tag{16}$$

III. Model and Methodology

Suppose $\boldsymbol{x}_{n\times 1}$ is a s-block-sparse signal consisting of q blocks $\{\mathcal{V}_b\}_{b=1}^q$ of equal size k, among which only s blocks are non-zero. We are further given a partition of $\{1, \ldots, q\}$ into L disjoint subsets $\mathcal{P}_i \subseteq \{1, \ldots q\}$, for which we know

$$\alpha_i = \tfrac{|\mathcal{P}_i \cap \mathcal{B}|}{|\mathcal{P}_i|}, \ \rho_i = \tfrac{|\mathcal{P}_i|}{q}, i = 1, ..., L, \tag{17}$$

where $\mathcal{B} \subseteq \{1, \ldots, q\}$ is the block-support of $\boldsymbol{x}_{n\times 1}$. Intuitively, α_i (which is called the accuracy of $\mathcal{P}_i$) stands for the normalized block-sparsity level of $\boldsymbol{x}$ restricted to $\mathcal{P}_i$. We call $\boldsymbol{x}$ a non-uniform block-sparse model with parameters $\{\alpha_i\}_{i=1}^L$ and $\{\rho_i\}_{i=1}^L$; if α_i's are all equal, the non-uniform model $\mathsf{P}_{1,2,\boldsymbol{w}}^\eta$ reduces to the uniform model $\mathsf{P}_{1,2}^\eta$. Each set $\mathcal{P}_i$ is associated with a weight $\omega_i \geqslant 0$ and the resulting weight $\boldsymbol{w}$ in $\mathsf{P}_{1,2,\boldsymbol{w}}^\eta$ is

$$\boldsymbol{w} = \boldsymbol{D}\boldsymbol{\omega}, \tag{18}$$

where $\boldsymbol{D} := [\boldsymbol{1}_{\mathcal{P}_1}, ..., \boldsymbol{1}_{\mathcal{P}_L}] \in \mathbb{R}^{q\times L}$. To better distinguish between ω_i and w_i, note that the former penalizes the index set $\mathcal{P}_i$ while the latter penalizes the index set $\mathcal{V}_i$ (see the illustrations in the left image of Figure 1).

In this work, the following questions are answered about a non-uniform block-sparse model:

1) How many measurements are required for $\mathsf{P}_{1,2}^0$ and $\mathsf{P}_{1,2,\boldsymbol{w}}^0$ to successfully recover a s-block-sparse vector from independent linear measurements?
2) Given extra prior information, what is the optimal choice of weights in $\mathsf{P}_{1,2,\boldsymbol{w}}^\eta$?
3) How close one can get to the optimal weights if the prior information is slightly inaccurate?

In the reminder of this work, we provide the answer to these questions in three sections.

IV. Number of Measurements for Successful Recovery

For a fixed tolerance $\zeta \in [0,1]$, let us denote the normalized number of measurements required for $\mathsf{P}^0_{1,2}$ and $\mathsf{P}^0_{1,2,\boldsymbol{w}}$ to recover a s-block-sparse vector (with probability $1-\zeta$) by $m_{q,s}$ and $m_{q,s,\boldsymbol{w}}$, respectively:

$$m_{q,s} := \tfrac{\delta(\mathcal{D}(\|\cdot\|_{1,2},\boldsymbol{x}))}{q}, \quad m_{q,s,\boldsymbol{w}} := \tfrac{\delta(\mathcal{D}(\|\cdot\|_{1,2,\boldsymbol{w}},\boldsymbol{x}))}{q}. \tag{19}$$

Below, we obtain an upper-bound for the number of measurements required for $\mathsf{P}^0_{1,2}$ to succeed with probability $1-\zeta$.

Lemma 1. *Let $\boldsymbol{x} \in \mathbb{R}^n$ be a non-uniform s-block-sparse vector in $\mathbb{R}^n$ with parameters $\{\rho_i\}_{i=1}^L$ and $\{\alpha_i\}_{i=1}^L$. Then,*

$$m_{q,s,\boldsymbol{w}} \leqslant \widehat{m}_{q,s,\boldsymbol{w}} \tag{20}$$

for

$$\widehat{m}_{q,s,\boldsymbol{w}} = \inf_{t \geqslant 0} \Psi_{t,\boldsymbol{w}}(\sigma,\boldsymbol{\rho},\boldsymbol{\alpha}), \tag{21}$$

where

$$\Psi_{t,\boldsymbol{w}}(\sigma,\boldsymbol{\rho},\boldsymbol{\alpha}) = \sum_{i=1}^{L} \rho_i\big(\alpha_i(k+t^2\omega_i^2) + \tfrac{(1-\alpha_i)\phi(t\omega_i)}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})}\big),$$
$$\phi(z) := \int_z^\infty (u-z)^2 u^{k-1}\exp(-\tfrac{u^2}{2})\mathrm{d}u, \;\; k = \tfrac{n}{q}. \tag{22}$$

Proof. See Appendix B.

Corollary 1. *By considering $\boldsymbol{w} = \mathbf{1} \in \mathbb{R}^q$, and using the fact that the normalized block-sparsity level is $\sigma := \frac{\|\boldsymbol{x}\|_{0,2}}{q} = \sum_{i=1}^L \rho_i\alpha_i$, we reach an upper-bound $\widehat{m}_{q,s}$ for $m_{q,s}$ as*

$$\widehat{m}_{q,s} = \inf_{t \geqslant 0} \Psi_t(\sigma), \tag{23}$$

where

$$\Psi_t(\sigma) = \sigma(k+t^2) + \tfrac{(1-\sigma)\phi(t)}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})}.$$

Remark. (Prior work) In [8, Lemma 3.2], the same expression as for $\widehat{m}_{q,s}$ is obtained for the normalized minimax MSE of the denoising problem

$$\min_{\boldsymbol{z}\in\mathbb{R}^n} \tau\|\boldsymbol{z}\|_{1,2} + \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{z}\|_2^2, \tag{24}$$

where $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{e}$ is the observed noisy vector. It is further conjectured in [8] that this value is equal to the number of measurements required by $\mathsf{P}^0_{1,2}$ in the asymptotic regime. We show that this formula describes the required number of measurements in $\mathsf{P}^0_{1,2}$ even in the non-asymptotic case (Proposition 2).

In the following Proposition, we demonstrate that the proposed upper-bound in Lemma 1 is asymptotically tight. We use this fact for the optimality of the obtained weights.

Proposition 2. *The normalized number of linear measurements required for $\mathsf{P}^0_{1,2,\boldsymbol{w}}$ and $\mathsf{P}^0_{1,2}$ to successfully recover a non-uniform s-block-sparse vector in $\mathbb{R}^n$ (i.e. $m_{q,s,\boldsymbol{w}}$ and $m_{q,s}$, respectively) satisfy the following error bounds:*

$$\widehat{m}_{q,s,\boldsymbol{w}} - \tfrac{2}{\sqrt{qL}} \leqslant m_{q,s,\boldsymbol{w}} \leqslant \widehat{m}_{q,s,\boldsymbol{w}}, \tag{25}$$
$$\widehat{m}_{q,s} - \tfrac{2}{\sqrt{sq}} \leqslant m_{q,s} \leqslant \widehat{m}_{q,s}. \tag{26}$$

Proof. See Appendix D.

It is interesting that the error bound in (26) is a special case of the error bound of Proposition (25) where one has s sets of blocks with size $\frac{q}{L}$ that each contributes to the block support with probability $\frac{L}{q}$.

V. Optimal Weights

Our strategy in finding the optimal weights is to minimize the reconstruction error (15) in the noisy case (under a fixed number of measurements i.e. m), and the required number of measurements in the noiseless case. Based on Theorem 1, both of these objectives lead to the same optimization problem

$$\boldsymbol{\omega}^* = \operatorname*{arg\,min}_{\boldsymbol{\omega}\in\mathbb{R}^L_+} \; m_{q,s,\boldsymbol{D\omega}} \in \mathbb{R}^L_+. \tag{27}$$

Instead of the latter minimization, we minimize the upper and lower bounds of statistical dimension (i.e. (25)), simultaneously. So

$$\inf_{\boldsymbol{\omega}\in\mathbb{R}^L_+} \widehat{m}_{q,s,\boldsymbol{D\omega}} - \tfrac{2}{\sqrt{qL}} \leqslant \inf_{\boldsymbol{\omega}\in\mathbb{R}^L_+} m_{q,s,\boldsymbol{D\omega}} \leqslant \inf_{\boldsymbol{\omega}\in\mathbb{R}^L_+} \widehat{m}_{q,s,\boldsymbol{D\omega}}, \tag{28}$$

where $\boldsymbol{D} := [\mathbf{1}_{\mathcal{P}_1},...,\mathbf{1}_{\mathcal{P}_L}]_{q\times L}$. In the weighted block sparsity optimization, we call the weight

$$\boldsymbol{\omega}^* = \operatorname*{arg\,min}_{\boldsymbol{\omega}\in\mathbb{R}^L_+} \; \widehat{m}_{q,s,\boldsymbol{D\omega}} \in \mathbb{R}^L_+ \tag{29}$$

optimal since it asymptotically (as $q\to\infty$) minimizes simultaneously the number of measurements required for $\mathsf{P}^0_{1,2,\boldsymbol{w}}$ to succeed, and the reconstruction error of $\mathsf{P}^\eta_{1,2,\boldsymbol{w}}$. In the following lemma, the uniqueness of the optimal weights is shown by proving that $\delta(\mathcal{D}(\|\cdot\|_{1,2,\boldsymbol{D\omega}},\boldsymbol{x}))$ is a strictly convex function of $\boldsymbol{\omega}\in\mathbb{R}^L_+$.

Lemma 2. *Assume $\mathcal{C} := \partial\|\cdot\|_{1,2}(\boldsymbol{x})$ does not contain the origin. We know that $\mathcal{C}$ is compact and $1 \leqslant \|\boldsymbol{z}\|_2 \leqslant \sqrt{q}$ for all $\boldsymbol{z}\in\mathcal{C}$. Also let $\boldsymbol{g}\in\mathbb{R}^n$ be a standard normal vector. Consider the function*

$$J(\boldsymbol{\nu}) := \mathbb{E}_{\boldsymbol{g}}\mathrm{dist}^2(\boldsymbol{g},\boldsymbol{\upsilon}\odot\mathcal{C}) = \mathbb{E}_{\boldsymbol{g}}[J_{\boldsymbol{g}}(\boldsymbol{\nu})]$$
$$\text{with} \;\; \boldsymbol{\upsilon} = \boldsymbol{D}_b\boldsymbol{D}\boldsymbol{\nu} \in \mathbb{R}^n \;\; \text{for} \;\; \boldsymbol{\nu}\in\mathbb{R}^L_+, \tag{30}$$

where $\boldsymbol{D}_b := [\mathbf{1}_{\mathcal{V}_1},...,\mathbf{1}_{\mathcal{V}_q}]_{n\times q}$. Then, the function J is strictly convex on $\boldsymbol{\nu}\in\mathbb{R}^L_{++}$ and J has a unique minimizer.

Proof. See Appendix E.

An analytic expression for the optimal weights is given in the following proposition via solving (29).

Proposition 3. *Let $\boldsymbol{x}$ be a non-uniform s-block-sparse vector in $\mathbb{R}^n$ with parameters $\{\rho_i\}_{i=1}^L$ and $\{\alpha_i\}_{i=1}^L$. Then, there exist unique optimal weights $\boldsymbol{\omega}^*\in\mathbb{R}^L_+$ (up to a positive scaling) that minimize $\widehat{m}_{q,s,\boldsymbol{D\omega}}$. The optimal weights $\boldsymbol{\omega}^*\in\mathbb{R}^L_+$ are obtained via the following integral equations:*

$$\alpha_i\omega_i^* = \tfrac{1}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})}(1-\alpha_i)\int_{\omega_i^*}^\infty (u-\omega_i^*)u^{k-1}\mathrm{e}^{-\frac{u^2}{2}}\mathrm{d}u$$
$$i = 1,...,L. \tag{31}$$

Proof. See Appendix C

The optimal weights in (31) depend only on the accuracy

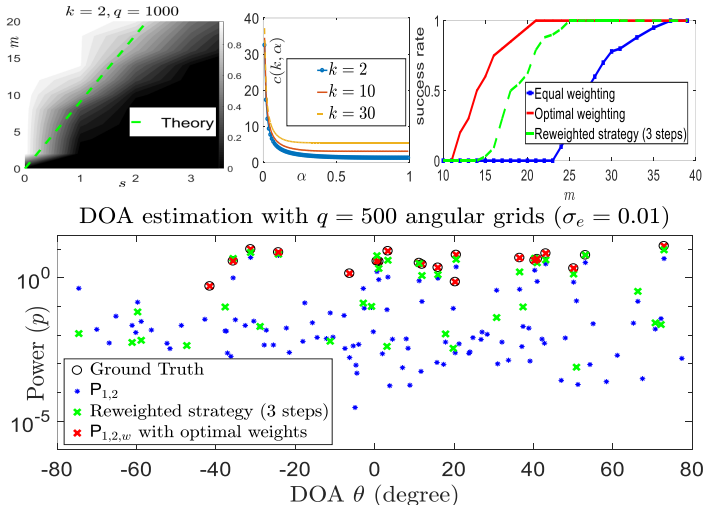

Figure 2: Top left image: This plot shows the empirical probability that $\mathsf{P}^0_{1,2}$ recovers $\boldsymbol{x} \in \mathbb{R}^{2000}$ with s blocks with nonzero ℓ_2 norm from m Gaussian linear measurements. The green dashed line shows the number of measurements obtained by Corollary 1. Top middle image: This plot shows $c(k,\alpha)$ in (34) versus the accuracy α for $k=2,10,30$. Indeed, it demonstrates the sensitivity of the optimal weight against small perturbations of α. Top right image: This plot shows the probability that $\mathsf{P}^0_{1,2,\boldsymbol{w}}$ succeed to recover $\boldsymbol{x} \in \mathbb{R}^{1280}$ from Gaussian linear measurements when $\boldsymbol{w}$ is chosen equally, optimally and via a reweighted strategy (3 steps) proposed in [20]. The parameters we used in this figure, are: $q=128$, $\sigma=\rho_1=\frac{10}{128}$, $\alpha_1=\frac{9}{10}$, $\rho_2=\frac{118}{128}$, $\alpha_2=\frac{1}{118}$. The optimal weights obtained via (31) with the aforementioned parameters are $\boldsymbol{\omega}^*=[.0766,1]^T$. Bottom image: DOA estimation of $s=20$ sources in the angular half-space $[-\frac{\pi}{2},\frac{\pi}{2}]$ with $k=10$ snapshots, $m=24$, $q=500$ and $\frac{d}{\lambda}=1$. The measurements are contaminated with additive Gaussian noise with $\sigma_e=0.01$. Also, we consider four angular bands with accuracies $\alpha_1=\frac{5}{6}$, $\alpha_2=\frac{13}{14}$, $\alpha_3=\frac{2}{3}$ and $\alpha_4=0$ as prior information.

of $\mathcal{P}_i$s, i.e., $\{\alpha_i\}_{i=1}^L$ and not their relative size $\{\rho_i\}_{i=1}^L$. Finally, we prove the significant fact that with optimal weights, $\mathsf{P}^0_{1,2,\boldsymbol{D\omega}^*}$ acts as if $\boldsymbol{x}_{\mathcal{P}_i}$'s are separately recovered via $\mathsf{P}^0_{1,2}$, in the sense of the required number of measurements.

Theorem 3. *Let $\boldsymbol{x}$ be a non-uniform s-block-sparse vector in $\mathbb{R}^n$ with parameters $\{\rho_i\}_{i=1}^L$ and $\{\alpha_i\}_{i=1}^L$. Then, the number of measurements required for $\mathsf{P}^0_{1,2,\boldsymbol{D\omega}^*}$ is exactly equals the total number of measurements required for $\mathsf{P}^0_{1,2}$ to recover each $\{\boldsymbol{x}_{\mathcal{P}_i} \in \mathbb{R}^n\}_{i=1}^L$ separately, up to an asymptotically additive vanishing error term i.e.*

$$-\tfrac{2}{\sqrt{qL}} \leqslant m_{q,s,\boldsymbol{D\omega}^*} - \sum_{i=1}^{L} m_{q,\|\boldsymbol{x}_{\mathcal{P}_i}\|_{0,2}} \leqslant \tfrac{2}{\sqrt{q}} \sum_{i=1}^{L} (\|\boldsymbol{x}_{\mathcal{P}_i}\|_{0,2})^{-\frac{1}{2}}. \tag{32}$$

VI. Robustness Analysis

In this section, we evaluate the robustness of optimal weights if prior information is slightly inaccurate. As stated in the below proposition, under the condition $\alpha \gtrapprox \frac{1}{10}$, the optimal weights $\boldsymbol{\omega}^*$ in (31) are robust to inaccuracies in the prior information.

Proposition 4. *Let α be the accuracy of a set $\mathcal{P}$ which inaccurately assumed α' in practice. Let ω and ω' be the optimal weights, obtained from (31), corresponding to α and α', respectively. Then, there exists a constant $c(k,\alpha)$ such that*

$$|\omega-\omega'| \leqslant c(k,\alpha)|\alpha-\alpha'| \tag{33}$$

for

$$c(k,\alpha) := \frac{\left(\sqrt{2}h(\alpha)\left(\Gamma(\frac{k}{2})-\gamma(\frac{k}{2},\frac{h(\alpha)^2}{2})\right)+2\gamma(\frac{k}{2},\frac{h(\alpha)^2}{2})\right)^2}{2\sqrt{2}\Gamma(\frac{k}{2})\gamma(\frac{k}{2},\frac{h(\alpha)^2}{2})}, \tag{34}$$

where $\gamma(a,z) := \int_z^\infty u^{a-1}\mathrm{e}^{-u}\mathrm{d}u$ is incomplete gamma function and $h(\alpha)$ is the nonlinear function in (31) that relates α to the optimal weight ω.

Proof. See Appendix G.

This proposition shows that our method of finding optimal weights in (31) is robust to inaccuracies in prior knowledge.

VII. Simulations

In this subsection, we numerically verify our theoretical results on the optimal weights. We have employed the CVX MATLAB package [23] to implement optimization problems. First, we investigate the required number of measurements for the scaling of successful recovery of $\mathsf{P}^0_{1,2}$ in terms of the block sparsity. For this purpose, we construct a s-block-sparse $\boldsymbol{x} \in \mathbb{R}^{2000}$. Then, we form the measurements $\boldsymbol{y}_{m\times 1}=\boldsymbol{A}\boldsymbol{x}$ and obtain an estimate $\widehat{\boldsymbol{x}}$ by solving the problem $\mathsf{P}^0_{1,2}$; we repeat this experiment for 100 realizations of $\boldsymbol{A}$. For each experiment, we declare success if $\|\boldsymbol{x}-\widehat{\boldsymbol{x}}\|_2 \leqslant 10^{-4}$. The heatmap in the top left image of Figure 2 shows the empirical probability of success for this procedure (black=0%, white=100%) which is consistent with the theory obtained by (23).

In the second experiment, we set $s=10$ and generate a block-sparse random vector $\boldsymbol{x} \in \mathbb{R}^{1280}$ with $q=128$ blocks of equal size $k=10$. The block support of $\boldsymbol{x}$ is drawn uniformly at random and the values within each non-zero block are drawn from i.i.d. standard normal distribution. Then, we consider two sets $\mathcal{P}_1$, $\mathcal{P}_2$ with $\alpha_1=\frac{9}{10}$, $\alpha_2=\frac{1}{118}$ that partition the set of blocks $\{1,...,128\}$. There are 100 Monte Carlo trials for each m where in each, we solve $\mathsf{P}^0_{1,2,\boldsymbol{D\omega}}$ with optimal weights $\boldsymbol{\omega}^*$, equal weights and the reweighted strategy (3 steps) proposed in [20] and recover $\boldsymbol{x} \in \mathbb{R}^{1280}$ from m Gaussian linear measurements. Note that $\boldsymbol{x}$ is kept fixed in Monte Carlo trials and only $\boldsymbol{A}$ changes. Optimal weights $\boldsymbol{w}^*$ are obtained from (31) by MATLAB function fzero. The top right image of Figure 2 shows that $\mathsf{P}^0_{1,2,\boldsymbol{D\omega}}$ with optimal weights needs fewer measurements than $\mathsf{P}^0_{1,2}$ and the reweighted method. To better investigate the sensitivity of optimal weights against small perturbations of α, we have plotted the robustness constant $c(k,\alpha)$ (see (34)) in the top middle image of Figure 2. These curves confirm the stability of optimal weights against the inaccuracy of prior information in the range $\alpha \gtrapprox \frac{1}{10}$.

In the last experiment, we investigate DOA estimation. We consider $s=20$ sources. As shown in the right block of Figure 1, these sources are scattered in the angular half-space $[-\frac{\pi}{2},\frac{\pi}{2}]$ which is divided into q angular grids. In practical scenarios, it is reasonable for an engineer to know the

likelihood of source appearance in some known angular bands (α_i in our non-uniform block-sparse model). This information can be obtained for instance by considering the statistics of previous estimations or the characteristics of the underlying model. In this experiment, we consider four angular bands with accuracies $\alpha_1 = \frac{5}{6}$, $\alpha_2 = \frac{13}{14}$, $\alpha_3 = \frac{2}{3}$ and $\alpha_4 = 0$. We have implemented $\mathsf{P}^{\eta}_{1,2,\boldsymbol{w}}$ using CVX software when $\boldsymbol{w}$ is chosen equally, optimally and via the reweighted strategy proposed in [20]. The optimal weights are calculated using (31). We follow the model (7) with $d = \lambda$, $q = 500$, $\sigma_e = 0.01$, and $m = 24$. As it turns out from bottom image of Figure 2, $\mathsf{P}^{\eta}_{1,2,\boldsymbol{w}*}$ exactly recovers the direction and the power of the sources in contrast to $\mathsf{P}^{\eta}_{1,2}$ and the reweighted strategy which both fail. This in turn reveals the fact that, with optimal weighting strategy, much less sensors are required for stable recovery.

VIII. Conclusion

In this paper, we studied the recovery of block-sparse vectors using a weighted $\ell_{1,2}$-minimization, when some prior information about the distribution of the block-support is available. We have particularly studied the case where partial block-sparsity level of the vector within a block partitioning is available as prior information. Our goal was to find the optimal weights so as to minimize the number of required measurements for perfect recovery in the noiseless case, and to minimize the reconstruction error in the noisy case. In both cases, we simplified the task to minimizing the statistical dimension of a weighted descent cone. We introduced closed-form expressions that identify the unique optimal weights. We have also shown the robustness of the optimal weights against inaccuracies of the prior information.

Appendix

A. Proof of Proposition 1

Assume first that $\boldsymbol{x} \neq \boldsymbol{0}$. By setting $\boldsymbol{y} = \boldsymbol{0}$ in (8), we have: $\langle \boldsymbol{z}, \boldsymbol{x} \rangle \geqslant f(\boldsymbol{x}) - f(\boldsymbol{0})$ and $f(\boldsymbol{0}) = 0$ due to the homogeneity, $f^*(\boldsymbol{z}) \geqslant 1$. Also, setting $\boldsymbol{y} = \boldsymbol{v} + \boldsymbol{x}$ in (8) under condition $f(\boldsymbol{v}) = 1$ we have:

$$f^*(\boldsymbol{z}) = \sup_{f(\boldsymbol{v})=1} \langle \boldsymbol{z}, \boldsymbol{v} \rangle \leqslant \sup_{f(\boldsymbol{v})=1} (f(\boldsymbol{v}+\boldsymbol{x}) - f(\boldsymbol{x})) \leqslant 1, \tag{35}$$

where we used sub-additivity of f. Hence, we have $f^*(\boldsymbol{z}) = 1$ and subsequently $\langle \boldsymbol{z}, \boldsymbol{x} \rangle = f(\boldsymbol{x})$.

On the other hand, if we have $\boldsymbol{z}$ such that $f^*(\boldsymbol{z}) = 1$ and $\langle \boldsymbol{z}, \boldsymbol{x} \rangle = f(\boldsymbol{x})$, then for each $\boldsymbol{y} \in \mathbb{R}^n$:

$$f(\boldsymbol{x}) + \langle \boldsymbol{z}, \boldsymbol{y} - \boldsymbol{x} \rangle = \langle \boldsymbol{z}, \boldsymbol{x} \rangle + \langle \boldsymbol{z}, \boldsymbol{y} - \boldsymbol{x} \rangle \leqslant f(\boldsymbol{y}) \cdot \tag{36}$$

For $\boldsymbol{x} = \boldsymbol{0}$ in (8), $\langle \boldsymbol{z}, \boldsymbol{y} \rangle \leqslant f(\boldsymbol{y})$ and then $f^*(\boldsymbol{z}) \leqslant 1$. Further, $f^*(\boldsymbol{z}) \leqslant 1$ implies that $\boldsymbol{z} \in \partial f(\boldsymbol{0})$.

B. Proof of Lemma 1

By the definition of statistical dimension for $\mathcal{D}(\|\cdot\|_{1,2,\boldsymbol{w}}, \boldsymbol{x})$ in (12), the fact that infimum of an affine function is concave and Jensen's inequality, we can find an upper-bound for $m_{q,s,\boldsymbol{w}}$ as:

$$m_{q,s,\boldsymbol{w}} \leqslant \inf_{t \geqslant 0} \underbrace{q^{-1}\mathbb{E}_{\boldsymbol{g}} \mathrm{dist}^2(\boldsymbol{g}, t\partial \|\cdot\|_{1,2,\boldsymbol{w}}(\boldsymbol{x}))}_{\Psi_{t,\boldsymbol{w}}(\sigma,\boldsymbol{\rho},\boldsymbol{\alpha})} := \widehat{m}_{q,s,\boldsymbol{w}}. \tag{37}$$

The next step is to calculate $\partial \|\cdot\|_{1,2,\boldsymbol{w}}(\boldsymbol{x})$. From Proposition 1, we have:

$$\partial \|\cdot\|_{1,2,\boldsymbol{w}}(\boldsymbol{x}) = \{\boldsymbol{z} \in \mathbb{R}^p : \ \langle \boldsymbol{z}, \boldsymbol{x} \rangle = \|\boldsymbol{x}\|_{1,2,\boldsymbol{w}}, \ \|\boldsymbol{z}\|^*_{1,2,\boldsymbol{w}} = 1\}. \tag{38}$$

It is not hard to show that,

$$\partial \|\cdot\|_{1,2,\boldsymbol{w}}(\boldsymbol{x}) = \left\{ \boldsymbol{z} \in \mathbb{R}^n : \ \begin{array}{ll} \frac{w_b \ \boldsymbol{x}_{\mathcal{V}_b}}{\|\boldsymbol{x}_{\mathcal{V}_b}\|_2}, & b \in \mathcal{B} \\ \|\boldsymbol{z}_{\mathcal{V}_b}\|_2 \leqslant w_b, & b \in \overline{\mathcal{B}} \end{array} \right\}. \tag{39}$$

Now to calculate $\Psi_{t,\boldsymbol{w}}(\sigma, \boldsymbol{\rho}, \boldsymbol{\alpha})$, regarding (39), we compute the distance of the dilated subdifferential of descent cone of $\ell_{1,2,\boldsymbol{w}}$ norm at $\boldsymbol{x} \in \mathbb{R}^n$ from a standard Gaussian vector $\boldsymbol{g} \in \mathbb{R}^n$ which is given by:

$$\begin{aligned} &\mathrm{dist}^2(\boldsymbol{g}, t\partial \|\cdot\|_{1,2,\boldsymbol{w}}(\boldsymbol{x})) = \inf_{\boldsymbol{z} \in \partial \|\cdot\|_{1,2,\boldsymbol{w}}(\boldsymbol{x})} \|\boldsymbol{g} - t\boldsymbol{z}\|_2^2 = \\ &\sum_{b \in \mathcal{B}} \|\boldsymbol{g}_{\mathcal{V}_b} - t w_b \tfrac{\boldsymbol{x}_{\mathcal{V}_b}}{\|\boldsymbol{x}_{\mathcal{V}_b}\|_2}\|_2^2 + \sum_{b \in \overline{\mathcal{B}}} \inf_{\|\boldsymbol{z}_{\mathcal{V}_b}\|_2 \leqslant w_b} \|\boldsymbol{g}_{\mathcal{V}_b} - t\boldsymbol{z}_{\mathcal{V}_b}\|_2^2 = \\ &\sum_{b \in \mathcal{B}} \|\boldsymbol{g}_{\mathcal{V}_b} - t w_b \tfrac{\boldsymbol{x}_{\mathcal{V}_b}}{\|\boldsymbol{x}_{\mathcal{V}_b}\|_2}\|_2^2 + \sum_{b \in \overline{\mathcal{B}}} (\|\boldsymbol{g}_{\mathcal{V}_b}\|_2 - t w_b)_+^2, \end{aligned} \tag{40}$$

where we used triangle inequality in the second part. By taking expectation from both sides, we reach:

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{g}} \mathrm{dist}^2(\boldsymbol{g}, t\partial \|\cdot\|_{1,2,\boldsymbol{w}}(\boldsymbol{x})) = \\ &ks + \sum_{b \in \mathcal{B}} (t w_b)^2 + \sum_{b \in \overline{\mathcal{B}}} \mathbb{E}(\underbrace{\|\boldsymbol{g}_{\mathcal{V}_b}\|_2}_{\zeta} - t w_b)_+^2, \end{aligned} \tag{41}$$

where $k = \frac{n}{q}$. and $\zeta^2 := \|\boldsymbol{g}_{\mathcal{V}_b}\|_2^2$ is distributed as chi-squared with k degrees of freedom. Moreover,

$$\begin{aligned} &\mathbb{E}(\zeta - t w_b)_+^2 = 2\int_0^\infty a\mathbb{P}(\zeta^2 \geqslant (a + t w_b)^2) da = \\ &\tfrac{2}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \int_0^\infty \int_{(t w_b + a)^2}^\infty a u^{\frac{k}{2}-1} e^{-\frac{u}{2}} du \ da = \\ &\tfrac{2}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \int_{(t w_b)^2}^\infty \int_0^{\sqrt{u} - t w_b} a u^{\frac{k}{2}-1} e^{-\frac{u}{2}} da \ du = \\ &\tfrac{1}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})} \int_{t w_b}^\infty (u - t w_b)^2 u^{k-1} e^{-\frac{u^2}{2}} du := \tfrac{\phi(t w_b)}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})}, \end{aligned} \tag{42}$$

where in the third line, the order of integration is changed and in the forth line, a change of variable is used. As a consequence, (41) becomes:

$$\mathbb{E}_{\boldsymbol{g}} \mathrm{dist}^2(\boldsymbol{g}, t\partial \|\cdot\|_{1,2,\boldsymbol{w}}(\boldsymbol{x})) = ks + \sum_{b \in \mathcal{B}} (t w_b)^2 + \tfrac{\sum_{b \in \overline{\mathcal{B}}} \phi(t w_b)}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})}. \tag{43}$$

By normalizing to the number of blocks q and incorporating block prior information using $\boldsymbol{w} = \boldsymbol{D\omega} \in \mathbb{R}^q$ we reach

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{g}}\text{dist}^2(\boldsymbol{g}, t\partial\|\cdot\|_{1,2,\boldsymbol{w}}(\boldsymbol{x})) = \\
&ks + \sum_{i=1}^{L} |\mathcal{P}_i \cap \mathcal{B}| t^2\omega_i^2 + |\mathcal{P}_i \cap \overline{\mathcal{B}}| \tfrac{\phi(t\omega_i)}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})} \\
&= q\Big(k\sigma + \sum_{i=1}^{L} \rho_i\big(\alpha_i t^2\omega_i^2 + \tfrac{(1-\alpha_i)\phi(t\omega_i)}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})}\big)\Big) \\
&= q\Big(\sum_{i=1}^{L} \rho_i\big(\alpha_i(t^2\omega_i^2 + k) + (1-\alpha_i)\tfrac{1}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})}\phi(t\omega_i)\big)\Big),
\end{aligned} \tag{44}$$

where in the last line above, we benefited the fact that $\sigma = \sum_{i=1}^{L} \rho_i\alpha_i$.

C. Proof of Proposition 3

Define $\mathcal{C} := \partial\|\cdot\|_{1,2}(\boldsymbol{x})$ and use Lemma 1 and 2 to obtain:

$$\inf_{\boldsymbol{\omega}\in\mathbb{R}_+^L} \widehat{m}_{q,s,\boldsymbol{D\omega}} = \inf_{\boldsymbol{\omega}\in\mathbb{R}_+^L} \inf_{t\in\mathbb{R}_+} \Psi_{t,\boldsymbol{D\omega}}(\sigma, \boldsymbol{\rho}, \boldsymbol{\alpha}) = \inf_{\boldsymbol{\nu}\in\mathbb{R}_+^L} J_b(\boldsymbol{\nu}),$$

where $\Psi_{t,\boldsymbol{D\omega}}(\sigma, \boldsymbol{\rho}, \boldsymbol{\alpha})$ is defined in (21). Also, we used a change of variable $\boldsymbol{\nu} = t\boldsymbol{\omega}$ to convert multivariate optimization problem to a single variable optimization problem. Thus, the function $J_b(\boldsymbol{\nu})$ is obtained via the following equation:

$$J_b(\boldsymbol{\nu}) = \sum_{i=1}^{L} \rho_i\big(\alpha_i(\nu(i)^2 + 1) + \tfrac{(1-\alpha_i)\phi(\nu(i))}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})}\big). \tag{45}$$

By considering Lemma 2 and $\boldsymbol{D}_t := [\mathbf{1}_{\mathcal{V}_1}, ..., \mathbf{1}_{\mathcal{V}_q}]_{n\times q}[\mathbf{1}_{\mathcal{P}_1}, ..., \mathbf{1}_{\mathcal{P}_L}]_{q\times L}$, the function $J_b(\boldsymbol{\nu})$ is continuous and strictly convex and thus the unique minimizer can be obtained using $\nabla J_b(\boldsymbol{\nu}) = \mathbf{0} \in \mathbb{R}^L$ which leads to

$$2\alpha_i\nu^*(i) + \tfrac{2(1-\alpha_i)\phi'(\nu^*(i))}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})} = 0 \ : \ i = 1, ..., L. \tag{46}$$

D. Proof of Proposition 2

By the error bound (16) and (39), with $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_{1,2,\boldsymbol{w}}$ and $\boldsymbol{w} = \sum_{i=1}^{L} \omega_i \mathbf{1}_{\mathcal{P}_i} \in \mathbb{R}^q$, the numerator of (16) is given by

$$2\sup_{\boldsymbol{s}\in\partial\|\cdot\|_{1,2,\boldsymbol{w}}(\boldsymbol{x})} \|\boldsymbol{s}\|_2 \leqslant 2\sqrt{\sum_{i=1}^{q} w_b^2} = 2\sqrt{\sum_{i=1}^{L} |\mathcal{P}_i|\omega_i^2} = 2\sqrt{\sum_{i=1}^{L} q\rho_i\omega_i^2}.$$

Also, for the denominator we have:

$$\tfrac{\|\boldsymbol{x}\|_{1,2,\boldsymbol{w}}}{\|\boldsymbol{x}\|_2} \leqslant \sqrt{\sum_{i\in\mathcal{B}} w_i^2} = \sqrt{\sum_{i=1}^{L} |\mathcal{P}_i \cap \mathcal{B}|\omega_i^2} = \sqrt{\sum_{i=1}^{L} q\alpha_i\rho_i\omega_i^2}, \tag{47}$$

where the first inequality in (47) follows from Cauchy–Schwartz inequality. The error bound (16) for $\|\cdot\|_{1,2,\boldsymbol{w}}$ depends only on $\mathcal{D}(\|\cdot\|_{1,2,\boldsymbol{w}}, \boldsymbol{x})$. Moreover, $\mathcal{D}(\|\cdot\|_{1,2,\boldsymbol{w}}, \boldsymbol{x})$ only requires that $\|\boldsymbol{x}_{\mathcal{V}_b}\|_2 = w_b \ : \ \forall b \in \mathcal{B}$. So, a vector

$$\boldsymbol{z} = \left\{ \begin{array}{ll} \|\boldsymbol{z}_{\mathcal{V}_b}\|_2 = w_b, & b \in \mathcal{B} \\ 0, & b \in \overline{\mathcal{B}} \end{array} \right\} \in \mathbb{R}^n$$

can be chosen to satisfy equality in (47). Therefore, the error of obtaining the upper-bound of $m_{q,s,\boldsymbol{w}}$, i.e. $\widehat{m}_{q,s,\boldsymbol{w}}$ is

$$\tfrac{2\sqrt{\sum_{i=1}^{L} q\rho_i\omega_i^2}}{q\sqrt{\sum_{i=1}^{L} q\alpha_i\rho_i\omega_i^2}} \leqslant \tfrac{2}{q}\sqrt{\tfrac{1}{\min\limits_{i\in[q]} \frac{|\mathcal{P}_i\cap\mathcal{B}|}{|\mathcal{P}_i|}}} \leqslant \tfrac{2}{\sqrt{qL}}, \tag{48}$$

in which the last inequality follows from the facts that $|\mathcal{P}_i \cap \mathcal{B}| \geqslant 1$, $|\mathcal{P}_i| \leqslant \frac{q}{L}$ for at least one $i \in [q]$ and thus $\min\limits_{i\in[q]} \frac{|\mathcal{P}_i\cap\mathcal{B}|}{|\mathcal{P}_i|} \geqslant \frac{L}{q}$. Further, the error of $\widehat{m}_{q,s,\boldsymbol{w}}$ from $m_{q,s,\boldsymbol{w}}$ is at most $\frac{2}{\sqrt{qL}}$. For the case of $\ell_{1,2}$, by using the facts $\omega_i = 1 \ \forall i$ and $s = q\sum_{i=1}^{L} \rho_i\alpha_i$, it is straightforward to verify that $\frac{2\sqrt{\sum_{i=1}^{L} q\rho_i\omega_i^2}}{q\sqrt{\sum_{i=1}^{L} q\alpha_i\rho_i\omega_i^2}} = \frac{2}{\sqrt{qs}}$.

E. Proof of Lemma 2

Convexity. Let $\boldsymbol{\nu}$, $\tilde{\boldsymbol{\nu}} \in \mathbb{R}_+^L$ and $\theta \in [0, 1]$ with $\boldsymbol{v} = \boldsymbol{D}_t\boldsymbol{\nu}$ and $\tilde{\boldsymbol{v}} = \boldsymbol{D}_t\tilde{\boldsymbol{\nu}}$. Then we have:

$$\begin{aligned}
&\forall\epsilon, \tilde{\epsilon} > 0 \ \exists\boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathcal{C} \ \text{ such that} \\
&\|\boldsymbol{g} - \boldsymbol{v} \odot \boldsymbol{z}\|_2 \leqslant \text{dist}(\boldsymbol{g}, \boldsymbol{v} \odot \mathcal{C}) + \epsilon, \\
&\|\boldsymbol{g} - \tilde{\boldsymbol{v}} \odot \tilde{\boldsymbol{z}}\|_2 \leqslant \text{dist}(\boldsymbol{g}, \tilde{\boldsymbol{v}} \odot \mathcal{C}) + \tilde{\epsilon}.
\end{aligned} \tag{49}$$

Since otherwise we have:

$$\begin{aligned}
&\forall\boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathcal{C} : \ \|\boldsymbol{g} - \boldsymbol{v} \odot \boldsymbol{z}\|_2 > \text{dist}(\boldsymbol{g}, \boldsymbol{v} \odot \mathcal{C}) + \epsilon, \\
&\|\boldsymbol{g} - \tilde{\boldsymbol{v}} \odot \tilde{\boldsymbol{z}}\|_2 > \text{dist}(\boldsymbol{g}, \tilde{\boldsymbol{v}} \odot \mathcal{C}) + \tilde{\epsilon}.
\end{aligned} \tag{50}$$

By taking the infimum over $\boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathcal{C}$, we reach a contradiction. We proceed to show convexity of $\text{dist}(\boldsymbol{g}, (\boldsymbol{D}_t\boldsymbol{\nu}) \odot \mathcal{C})$ by writing

$$\begin{aligned}
&\text{dist}(\boldsymbol{g}, (\theta\boldsymbol{v} + (1-\theta)\tilde{\boldsymbol{v}}) \odot \mathcal{C}) = \\
&\inf_{\boldsymbol{z}\in\mathcal{C}} \|\boldsymbol{g} - (\theta\boldsymbol{v} + (1-\theta)\tilde{\boldsymbol{v}}) \odot \boldsymbol{z}\|_2 \\
&\leqslant \inf_{\boldsymbol{z}_1\in\mathcal{C}, \boldsymbol{z}_2\in\mathcal{C}} \|\boldsymbol{g} - \theta\boldsymbol{v} \odot \boldsymbol{z}_1 - (1-\theta)\boldsymbol{v} \odot \boldsymbol{z}_2\|_2 \leqslant \\
&\theta\|\boldsymbol{g} - \boldsymbol{v} \odot \boldsymbol{z}_1\|_2 + (1-\theta)\|\boldsymbol{g} - \tilde{\boldsymbol{v}} \odot \boldsymbol{z}_2\|_2 \leqslant \\
&\theta\text{dist}(\boldsymbol{g}, \boldsymbol{v} \odot \mathcal{C}) + (1-\theta)\text{dist}(\boldsymbol{g}, \tilde{\boldsymbol{v}} \odot \mathcal{C}) + \epsilon + \tilde{\epsilon}.
\end{aligned} \tag{51}$$

Since this holds for any ϵ and $\tilde{\epsilon}$, $\text{dist}(\boldsymbol{g}, (\boldsymbol{D}_t\boldsymbol{\nu}) \odot \mathcal{C})$ is a convex function. As the square of a non-negative convex function is convex, $J_{\boldsymbol{g}}(\boldsymbol{\nu})$ is a convex function. At last, the function $J(\boldsymbol{\nu})$ is the average of convex functions, hence is convex. In (51), the first inequality comes from the fact that $\forall\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{C} \ \ \exists\boldsymbol{z} \in \mathcal{C}$:

$$\begin{aligned}
&\theta\boldsymbol{v} \odot \boldsymbol{z}_1 + (1-\theta)\tilde{\boldsymbol{v}} \odot \boldsymbol{z}_2 = \\
&\left\{ \boldsymbol{y}_{n\times 1} : \begin{array}{ll} \boldsymbol{y}_{\mathcal{V}_b} = \big(\theta\boldsymbol{v}_{\mathcal{V}_b} + (1-\theta)\tilde{\boldsymbol{v}}_{\mathcal{V}_b}\big) \odot \frac{\boldsymbol{x}_{\mathcal{V}_b}}{\|\boldsymbol{x}_{\mathcal{V}_b}\|_2}, & b \in \mathcal{B} \\ \|\boldsymbol{y}_{\mathcal{V}_b}\|_2 \leqslant \theta\|\boldsymbol{v}\|_\infty\|\boldsymbol{z}_{1\mathcal{V}_b}\|_2 & \\ \quad +(1-\theta)\|\tilde{\boldsymbol{v}}\|_\infty\|\boldsymbol{z}_{2\mathcal{V}_b}\|_2, & b \in \overline{\mathcal{B}} \end{array} \right\} \\
&\in \left\{ \boldsymbol{y}_{n\times 1} : \begin{array}{ll} \boldsymbol{y}_{\mathcal{V}_b} = \big(\theta\boldsymbol{v}_{\mathcal{V}_b} + (1-\theta)\tilde{\boldsymbol{v}}_{\mathcal{V}_b}\big) \odot \frac{\boldsymbol{x}_{\mathcal{V}_b}}{\|\boldsymbol{x}_{\mathcal{V}_b}\|_2}, & b \in \mathcal{B} \\ \|\boldsymbol{y}_{\mathcal{V}_b}\|_2 \leqslant & \\ \quad \big(\theta\|\boldsymbol{v}\|_\infty + (1-\theta)\|\tilde{\boldsymbol{v}}\|_\infty\big)\|\boldsymbol{z}_{\mathcal{V}_b}\|_2, & b \in \overline{\mathcal{B}} \end{array} \right\} \\
&= (\theta\boldsymbol{v} + (1-\theta)\tilde{\boldsymbol{v}}) \odot \boldsymbol{z}.
\end{aligned} \tag{52}$$

To verify (52), we argue by contradiction:

$$\forall \boldsymbol{z} \in \mathcal{C}\ \exists d \in \overline{\mathcal{B}}:\ (\theta\|\boldsymbol{v}\|_\infty + (1-\theta)\|\tilde{\boldsymbol{v}}\|_\infty)\|\boldsymbol{z}_{\mathcal{V}_d}\|_2 < \theta\|\boldsymbol{v}\|_\infty$$
$$\|\boldsymbol{z}_{1\mathcal{V}_d}\|_2 + (1-\theta)\|\tilde{\boldsymbol{v}}\|_\infty\|\boldsymbol{z}_{2\mathcal{V}_d}\|_2 \leqslant \theta\|\boldsymbol{v}\|_\infty + (1-\theta)\|\tilde{\boldsymbol{v}}\|_\infty. \tag{53}$$

Then, by taking $\boldsymbol{z}_{\mathcal{V}_d} = \boldsymbol{e}_i \in \mathbb{R}^k$ for some $i \in [k]$, we reach a contradiction. In the second inequality in (51), we used triangle inequality of norms. The third inequality uses (49).

Strict convexity. We show strict convexity by contradiction. If $J(\boldsymbol{\nu})$ was not strictly convex, there would be vectors $\boldsymbol{\nu}, \tilde{\boldsymbol{\nu}} \in \mathbb{R}_+^L$ with $\boldsymbol{v} = \boldsymbol{D}_t\boldsymbol{\nu}, \tilde{\boldsymbol{v}} = \boldsymbol{D}_t\tilde{\boldsymbol{\nu}}$ and $\theta \in (0,1)$ such that

$$\mathbb{E}[J_{\boldsymbol{g}}(\theta\boldsymbol{\nu} + (1-\theta)\tilde{\boldsymbol{\nu}})] = \mathbb{E}[\theta J_{\boldsymbol{g}}(\boldsymbol{\nu}) + (1-\theta)J_{\boldsymbol{g}}(\tilde{\boldsymbol{\nu}})]. \tag{54}$$

For each $\boldsymbol{g}$ in (54) the left-hand side is smaller than or equal to the right-hand side. Therefore, in (54), $J_{\boldsymbol{g}}(\theta\boldsymbol{\nu} + (1-\theta)\tilde{\boldsymbol{\nu}})$ and $\theta J_{\boldsymbol{g}}(\boldsymbol{\nu}) + (1-\theta)J_{\boldsymbol{g}}(\tilde{\boldsymbol{\nu}})$ are almost surely equal (except at a set of measure zero) with respect to Gaussian measure. Moreover, we have

$$J_{\boldsymbol{0}}(\theta\boldsymbol{\nu} + (1-\theta)\tilde{\boldsymbol{\nu}}) = \operatorname{dist}^2(\boldsymbol{0}, \big(\theta\boldsymbol{v} + (1-\theta)\tilde{\boldsymbol{v}}\big) \odot \mathcal{C}) \leqslant$$
$$\inf_{\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{C}} \|\theta\boldsymbol{v} \odot \boldsymbol{z}_1 + (1-\theta)\tilde{\boldsymbol{v}} \odot \boldsymbol{z}_2\|_2^2 < \theta \inf_{\boldsymbol{z}_1 \in \mathcal{C}} \|\boldsymbol{v} \odot \boldsymbol{z}_1\|_2^2 +$$
$$(1-\theta) \inf_{\boldsymbol{z}_2 \in \mathcal{C}} \|\tilde{\boldsymbol{v}} \odot \boldsymbol{z}_2\|_2^2 = \theta J_{\boldsymbol{0}}(\boldsymbol{\nu}) + (1-\theta)J_{\boldsymbol{0}}(\tilde{\boldsymbol{\nu}}), \tag{55}$$

where the first inequality comes from (52) and the second inequality stems from the strict convexity of $\|\cdot\|_2^2$. From (52), it is easy to verify that the set $\boldsymbol{\nu} \odot \mathcal{C}$ is a convex set. The distance to a convex set e.g. $\mathcal{E}$ i.e. $\operatorname{dist}(\boldsymbol{g}, \mathcal{E})$ is a 1-lipschitz function (i.e. $|\operatorname{dist}(\boldsymbol{g}, \mathcal{E}) - \operatorname{dist}(\tilde{\boldsymbol{g}}, \mathcal{E})| \leqslant \|\boldsymbol{g} - \tilde{\boldsymbol{g}}\|_2\ :\ \forall\ \boldsymbol{g}, \tilde{\boldsymbol{g}} \in \mathbb{R}^n$) and hence continuous with respect to $\boldsymbol{g}$. Therefore, $J_{\boldsymbol{g}}(\boldsymbol{\nu})$ is continuous with respect to $\boldsymbol{g}$. So there exist an open ball around $\boldsymbol{g} = \boldsymbol{0} \in \mathbb{R}^n$ that similar to (55), we may write the following relation for some $\epsilon > 0$.

$$\exists \boldsymbol{u} \in \mathbb{B}_\epsilon^n:\ J_{\boldsymbol{u}}(\theta\boldsymbol{\nu} + (1-\theta)\tilde{\boldsymbol{\nu}}) < \theta J_{\boldsymbol{u}}(\boldsymbol{\nu}) + (1-\theta)J_{\boldsymbol{u}}(\tilde{\boldsymbol{\nu}}). \tag{56}$$

The above statement contradicts with (54) and hence we have strict convexity. Continuity along with convexity of J implies that J is convex on the whole domain $\boldsymbol{\nu} \in \mathbb{R}_+^L$.

Differentiability and continuity. The function $J_{\boldsymbol{g}}(\boldsymbol{\nu})$ is continuously differentiable and the gradient for $\boldsymbol{\nu} \in \mathbb{R}_{++}^L$ is

$$\nabla_{\boldsymbol{\nu}} J_{\boldsymbol{g}}(\boldsymbol{\nu}) =$$
$$\frac{\partial J_{\boldsymbol{g}}(\boldsymbol{\nu})}{\partial \boldsymbol{\nu}} = -2\boldsymbol{D}_t^T(\boldsymbol{D}_t\boldsymbol{\nu})^{\odot(-1)} \odot \mathcal{P}_{(\boldsymbol{D}_t\boldsymbol{\nu}) \odot \mathcal{C}}(\boldsymbol{g})$$
$$\odot\, (\boldsymbol{g} - \mathcal{P}_{\boldsymbol{D}_t\boldsymbol{\nu} \odot \mathcal{C}}(\boldsymbol{g})). \tag{57}$$

Continuity of $\frac{\partial J_{\boldsymbol{g}}(\boldsymbol{\nu})}{\partial \boldsymbol{\nu}}$ at $\boldsymbol{\nu} \in \mathbb{R}_+^L$ stems from the fact that the projection onto a convex set is continuous. For each compact set $\mathcal{I} \subseteq \mathbb{R}_+^L$ we have:

$$\mathbb{E} \sup_{\boldsymbol{\nu} \in \mathcal{I}} \|\nabla_{\boldsymbol{\nu}} J_{\boldsymbol{g}}(\boldsymbol{\nu})\|_2 \leqslant$$
$$2\|\boldsymbol{D}_t\|_{2\to 2}\sqrt{q}(\sqrt{n} + 2\sqrt{q}\big(\sup_{\boldsymbol{\nu} \in \mathcal{I}} \nu_{\max}\big)) < \infty, \tag{58}$$

where $\nu_{\max} := \max_{i \in [L]} \nu(i)$. Therefore, we have

$$\nabla_{\boldsymbol{\nu}} J(\boldsymbol{\nu}) = \big(\frac{\partial}{\partial \boldsymbol{\nu}}\big)\mathbb{E} J_{\boldsymbol{g}}(\boldsymbol{\nu}) = \mathbb{E}[\nabla_{\boldsymbol{\nu}} J_{\boldsymbol{g}}(\boldsymbol{\nu})]\ :\ \forall \boldsymbol{\nu} \in \mathbb{R}_+^L, \tag{59}$$

where in the last equality, we used the Lebesgue's dominated convergence theorem. Also, continuity of $J(\boldsymbol{\nu})$ can be concluded from the continuity of its gradient.

Attainment of the minimum. Suppose that $\boldsymbol{\nu} \geqslant \|\boldsymbol{g}\|_2 \boldsymbol{1}_{L\times 1}$. With this assumption, we may write:

$$\operatorname{dist}(\boldsymbol{g}, (\boldsymbol{D}_t\boldsymbol{\nu}) \odot \mathcal{C}) = \inf_{\boldsymbol{z} \in \mathcal{C}} \|\boldsymbol{g} - \boldsymbol{v} \odot \boldsymbol{z}\|_2 \geqslant$$
$$\inf_{\boldsymbol{z} \in \mathcal{C}} (\|\boldsymbol{v} \odot \boldsymbol{z}\|_2 - \|\boldsymbol{g}\|_2) \geqslant \nu_{\min} - \|\boldsymbol{g}\|_2 \geqslant 0, \tag{60}$$

where in (60), $\nu_{\min} := \min_{i \in [L]} \nu(i)$. By squaring (60), we reach

$$J_{\boldsymbol{g}}(\boldsymbol{\nu}) \geqslant (\nu_{\min} - \|\boldsymbol{g}\|_2)^2\ \ :\ \forall \nu > \|\boldsymbol{g}\|_2 \boldsymbol{1}_{L\times 1} \tag{61}$$

Using $\mathbb{E}\|\boldsymbol{g}\|_2 \geqslant \frac{n}{\sqrt{n+1}}$ [24, Proposition 8.1] and Marcov's inequality we obtain:

$$\mathbb{P}(\|\boldsymbol{g}\|_2 \leqslant \sqrt{n}) \geqslant 1 - \sqrt{\frac{n}{n+1}}.$$

Then we get:

$$J(\boldsymbol{\nu}) \geqslant \mathbb{E}[J_{\boldsymbol{g}}(\boldsymbol{\nu}) | \|\boldsymbol{g}\|_2 \leqslant \sqrt{n}]\mathbb{P}(\|\boldsymbol{g}\|_2 \leqslant \sqrt{n})$$
$$\geqslant (1 - \sqrt{\frac{n}{n+1}})\mathbb{E}\big[(\nu_{\min} - \|\boldsymbol{g}\|_2)^2 | \|\boldsymbol{g}\|_2 \leqslant \sqrt{n}\big]$$
$$\geqslant (1 - \sqrt{\frac{n}{n+1}})(\nu_{\min} - \sqrt{n})^2, \tag{62}$$

where the first inequality stems from total probability theorem, the second inequality follows from (61). From (62), we find out that $J(\boldsymbol{\nu}) > J(\boldsymbol{0})$ when $\nu > (2^{\frac{1}{4}} + 1)\sqrt{n}\boldsymbol{1}_{L\times 1}$. Therefore, the unique minimizer of the function J must occur in the interval $[\boldsymbol{0}, (2^{\frac{1}{4}} + 1)\sqrt{n}\boldsymbol{1}_{L\times 1}]$.

F. Proof of Theorem 3

Using optimal weights, the upper-bound for the normalized number of measurements required for $\mathsf{P}^0_{1,2,\boldsymbol{D}\boldsymbol{\omega}*}$ to succeed is:

$$\widehat{m}_{q,s,\boldsymbol{w}*} = \inf_{\omega \in \mathbb{R}_+^L} \widehat{m}_{q,s,\boldsymbol{D}\boldsymbol{\omega}} =$$
$$\sum_{i=1}^{L} [\inf_{\nu_i \in \mathbb{R}_+} \underbrace{\Big(\frac{\|\boldsymbol{x}_{\mathcal{P}_i}\|_{0,2}}{q}(\nu_i^2 + k) + (1 - \frac{\|\boldsymbol{x}_{\mathcal{P}_i}\|_{0,2}}{q})\phi_B(\nu_i)\Big)}_{\Psi_{\nu_i, \|\boldsymbol{x}_{\mathcal{P}_i}\|_{0,2}}(\frac{\|\boldsymbol{x}_{\mathcal{P}_i}\|_{0,2}}{q})}]$$
$$= \sum_{i=1}^{L} \widehat{m}_{q, \|\boldsymbol{x}_{\mathcal{P}_i}\|_{0,2}}. \tag{63}$$

The expression in the bracket is the upper-bound for normalized number of measurements required for successful recovery of $\boldsymbol{x}_{\mathcal{P}_i} \in \mathbb{R}^n$ using $\mathsf{P}^0_{1,2}$ i.e. $\widehat{m}_{q, \|\boldsymbol{x}_{\mathcal{P}_i}\|_{0,2}}$. Thus, regarding the error bounds obtained in Proposition 2, the relation between $m_{q,s,\boldsymbol{D}\boldsymbol{\omega}*}$ and $m_{q, \|\boldsymbol{x}_{\mathcal{P}_i}\|_{0,2}}$ is given by (32).

G. Proof of Proposition 4

We have that $\lim_{\alpha \to \alpha'} \frac{|\omega - \omega'|}{|\alpha - \alpha'|} = |\frac{\partial \omega}{\partial \alpha}|$. By differentiating (31), we reach

$$\frac{\partial \omega}{\partial \alpha} = -\frac{\omega + \frac{1}{2^{k/2-1}\Gamma(k/2)} \int_\omega^\infty (u - \omega) u^{k-1} e^{-\frac{u^2}{2}}\, du}{\alpha + \frac{1-\alpha}{2^{k/2-1}\Gamma(k/2)} \int_\omega^\infty u^{k-1} e^{-\frac{u^2}{2}}\, du}. \tag{64}$$

Using (31), the above equation reduces to:

$$\frac{\partial \omega}{\partial \alpha} = -\frac{\omega^2 2^{k/2-1}\Gamma(k/2)}{(1-\alpha)^2 \int_\omega^\infty u^k e^{-\frac{u^2}{2}}\, du}. \tag{65}$$

By obtaining α from (31) and replacing in (65), we reach:

$$\frac{\partial \omega}{\partial \alpha} = -\frac{\left(\int_\omega^\infty u^k e^{-\frac{u^2}{2}}\, du - \omega \int_\omega^\infty u^{k-1} e^{-\frac{u^2}{2}} + 2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})\omega\right)^2}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2}) \int_\omega^\infty u^k e^{-\frac{u^2}{2}}\, du} := f(\omega). \tag{66}$$

After some simplification, $f(\omega)$ reduces to

$$f(\omega) = \frac{\left(\sqrt{2}\omega\left(\Gamma(\frac{k}{2}) - \gamma(\frac{k}{2}, \frac{\omega^2}{2})\right) + 2\gamma(\frac{k}{2}, \frac{\omega^2}{2})\right)^2}{2\sqrt{2}\Gamma(\frac{k}{2})\gamma(\frac{k}{2}, \frac{\omega^2}{2})}. \tag{67}$$

$f(\omega)$ implicitly depends on the accuracy α since ω is related to α by (31). $f(\omega)$ can be further simplified to a function $c(k, \alpha)$ that only depends on k and α. This is accomplished by obtaining ω corresponding to any $\alpha \in [0, 1]$ by (31) and replacing the result into (67).

REFERENCES

[1] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *arXiv preprint arXiv:0804.0041*, 2008.

[2] M. M. Hyder and K. Mahata, "Direction-of-arrival estimation using a mixed $\ell_{0,2}$ norm approximation," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4646–4655, 2010.

[3] T. Euler and T. Baden, "Computational neuroscience: Species-specific motion detectors," *Nature*, 2016.

[4] J. Zhu, D. Baron, and F. Krzakala, "Performance limits for noisy multimeasurement vector problems," *IEEE Transactions on Signal Processing*, vol. 65, no. 9, pp. 2444–2454, 2017.

[5] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Communications on pure and applied mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[6] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Information and Inference: A Journal of the IMA*, vol. 3, no. 3, pp. 224–294, 2014.

[7] K. V. Mishra, M. Cho, A. Kruger, and W. Xu, "Spectral super-resolution with prior knowledge," *IEEE transactions on signal processing*, vol. 63, no. 20, pp. 5342–5357, 2015.

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

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

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

[11] D. Needell, R. Saab, and T. Woolf, "Weighted-minimization for sparse recovery under arbitrary prior information," *Information and Inference: A Journal of the IMA*, vol. 6, no. 3, pp. 284–309, 2017.

[12] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[13] W. Xu, *Compressive sensing for sparse approximations: constructions, algorithms, and analysis*. PhD thesis, California Institute of Technology, 2010.

[14] A. Bereyhi and R. R. Müller, "Maximum-a-posteriori signal recovery with prior information: Applications to compressive sensing," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4494–4498, IEEE, 2018.

[15] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[16] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[17] A. Flinth, "Optimal choice of weights for sparse recovery with prior information," *IEEE Transactions on Information Theory*, vol. 62, no. 7, pp. 4276–4284, 2016.

[18] S. D. Babacan, S. Nakajima, and M. N. Do, "Bayesian group-sparse modeling and variational inference," *IEEE transactions on signal processing*, vol. 62, no. 11, pp. 2906–2921, 2014.

[19] S. Som, L. C. Potter, and P. Schniter, "On approximate message passing for reconstruction of non-uniformly sparse signals," in *Proceedings of the IEEE 2010 National Aerospace & Electronics Conference*, pp. 223–229, IEEE, 2010.

[20] R. Ahmad and P. Schniter, "Iteratively reweighted ℓ_1 approaches to sparse composite regularization," *IEEE transactions on computational imaging*, vol. 1, no. 4, pp. 220–235, 2015.

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

[22] J. A. Tropp, "Convex recovery of a structured signal from independent random linear measurements," in *Sampling Theory, a Renaissance*, pp. 67–101, Springer, 2015.

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

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

Sajad Daei received the B.Sc. degree in electronic engineering from Guilan University, Guilan, Iran, in 2011, and the M.Sc. degree in communication engineering from Sharif University of Technology, Tehran, Iran, in 2013. He is currently pursuing his Ph.D. at Iran University of Science & Technology. His main research interests include convex optimization, compressed sensing and super resolution.

Farzan Haddadi was born in 1979. He received his B.Sc., M.Sc., and Ph.D. degrees in communication systems in 2001, 2003, and 2010, respectively, from Sharif University of Technology, Tehran, Iran. He joined Iran University of Science & Technology faculty in 2011. His main research interests are array signal processing, statistical signal processing, subspace tracking, and compressed sensing.

Arash Amini received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering (communications and signal processing) and the B.Sc. degree in petroleum engineering (reservoir) from the Sharif University of Technology, Tehran, Iran, in 2005, 2007, 2011, and 2005, respectively. He was a Researcher with the École Polytechnique fédérale de Lausanne, Lausanne, Switzerland, from 2011 to 2013, working on statistical approaches toward modeling sparsity in continuous-domain. He joined Sharif University of Technology as an assistant professor in 2013, where he is now an associate professor since 2018. He has served as an associate editor of IEEE Signal Processing Letters from 2014 to 2018. His research interests include various topics in statistical signal processing, specially compressed sensing.