

Convex block-sparse linear regression with expanders – provably

Anastasios Kyrillidis
UT Austin

Bubacarr Bah
UT Austin

Rouzbeh Hasheminezhad
Sharif University of Technology

Quoc Tran-Dinh
University of North Carolina

Luca Baldassarre
EPFL

Volkan Cevher
EPFL

April 5, 2016

Abstract

Sparse matrices are favorable objects in machine learning and optimization. When such matrices are used, in place of dense ones, the overall complexity requirements in optimization can be significantly reduced in practice, both in terms of space and run-time. Prompted by this observation, we study a convex optimization scheme for block-sparse recovery from linear measurements. To obtain linear sketches, we use expander matrices, *i.e.*, sparse matrices containing only few non-zeros per column. Hitherto, to the best of our knowledge, such algorithmic solutions have been only studied from a non-convex perspective. Our aim here is to theoretically characterize the performance of convex approaches under such setting.

Our key novelty is the expression of the recovery error in terms of the model-based norm, while assuring that solution lives in the model. To achieve this, we show that sparse model-based matrices satisfy a *group* version of the null-space property. Our experimental findings on synthetic and real applications support our claims for faster recovery in the convex setting – as opposed to using dense sensing matrices, while showing a competitive recovery performance.

1 Introduction

Consider the *sparse recovery problem* in the linear setting: given a measurement matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ ($n < p$), we seek a sparse vector $\beta^* \in \mathbb{R}^p$ that satisfies a set of measurements $\mathbf{y} = \mathbf{X}\beta^*$ or finds a solution in the feasible set $\{\beta : \|\mathbf{y} - \mathbf{X}\beta\|_2 \leq \|\xi\|_2\}$, when noise $\xi \in \mathbb{R}^n$ is present. This problem is closely related to compressive sensing [11, 9] and the subset selection problem [27], as well as to graph sketching [1] and data streaming [25].

Typically, the analysis included in such task focuses on investigating conditions on \mathbf{X} and the recovery algorithm Δ to obtain estimate $\hat{\beta} = \Delta(\mathbf{X}\beta^* + \xi)$ that yields a guarantee:

$$\|\beta^* - \hat{\beta}\|_{q_1} \leq C_1 \|\beta^* - \beta_s^*\|_{q_2} + C_2 \|\xi\|_{q_1}. \quad (1)$$

Here, β_s^* denotes the best s -sparse approximation to β^* for constants $C_1, C_2 > 0$ and $1 \leq q_2 \leq q_1 \leq 2$ [7, 5]. This type of guarantee is known as the ℓ_{q_1}/ℓ_{q_2} error guarantee.

Without further assumptions, there is a range of recovery algorithms that achieve the above, both from a non-convex [24, 21, 23] and a convex perspective [10, 5]. *In this work, we focus on the latter case.* Within this context, we highlight ℓ_1 -minimization method, also known as *Basis pursuit* (BP) [10], which is one of the prevalent schemes for compressive sensing:

$$\min_{\beta} \|\beta\|_1 \quad \text{subject to: } \mathbf{y} = \mathbf{X}\beta. \quad (2)$$

Next, we discuss common configurations for such problem settings, as in (2), and how a priori knowledge can be incorporated in optimization. We conclude this section with our contributions.

Measurement matrices in sparse recovery: Conventional wisdom states that the best compression performance, *i.e.*, using the least number of measurements to guarantee (1), is achieved by random, independent and identically distributed *dense* sub-gaussian matrices [8]. Such matrices are known to satisfy the ℓ_q -norm restricted isometry property (RIP- q) with high-probability. That is, for some $\delta_s \in (0, 1)$ and $\forall s$ -sparse $\beta \in \mathbb{R}^p$ [6, 5]:

$$(1 - \delta_s) \|\beta\|_q^q \leq \|\mathbf{X}\beta\|_q^q \leq (1 + \delta_s) \|\beta\|_q^q. \quad (3)$$

In this case, we can tractably and provably approximate sparse signals from $n = \mathcal{O}(s \log(p/s))$ sketches for generic s -sparse signals. *Unfortunately, dense matrices require high time- and storage-complexity when applied in algorithmic solutions.*¹

¹There exist *structured* dense matrices that achieve better time and storage complexity, as compared to random Gaussian matrices; see [26] for discussion on subsampled Fourier matrices for sparse recovery. However, sparse design matrices have been proved to be more advantageous over such structured dense matrices [5].

On the other hand, in several areas, such as data stream computing and combinatorial group testing [25], it is vital to use a *sparse* sensing matrix \mathbf{X} , containing only very few non-zero elements per column [18, 5]. In this work, we consider random sparse matrices \mathbf{X} that are adjacency matrices of expander graphs; a precise definition is given below.

Definition 1.1. Let a graph (p, n, \mathcal{E}) be a left-regular bipartite graph with p left (variable) nodes, n right (check) nodes, a set of edges \mathcal{E} and left degree d . If, any $\mathcal{S} \subseteq \{1, \dots, p\}$ with $|\mathcal{S}| = s \ll p$ and for some $\epsilon_s \in (0, 1/2)$, we have that the set of neighbours of \mathcal{S} , $\Gamma(\mathcal{S})$, satisfy $|\Gamma(\mathcal{S})| \geq (1 - \epsilon_s)d|\mathcal{S}|$, then the graph is called a (s, d, ϵ_s) -lossless expander graph. The adjacency matrix of an expander graph is denoted as $\mathbf{X} \in \{0, 1\}^{n \times p}$ and is called as expander matrix.

In the above, the attribute $|\Gamma(\mathcal{S})| \geq (1 - \epsilon_s)d|\mathcal{S}|$ is known as the *expansion* property, where $\epsilon_s \ll 1$ [20]. As [5] states, expander matrices satisfy the RIP-1 condition if $\forall s$ -sparse vector $\boldsymbol{\beta} \in \mathbb{R}^p$, we have:

$$(1 - 2\epsilon_s)d\|\boldsymbol{\beta}\|_1 \leq \|\mathbf{X}\boldsymbol{\beta}\|_1 \leq d\|\boldsymbol{\beta}\|_1. \quad (4)$$

Since the recovery algorithms typically use \mathbf{X} and its adjoint \mathbf{X}^T as subroutines over vectors, *sparse matrices have low computational complexity, even in the convex case* (2), without any loss in sample complexity $n = \mathcal{O}(s \log(p/s))$; see Section 7 for some illustrative results on this matter. Along with computational benefits, [5] provides recovery guarantees of the form (1) for BP recovery where $q_1 = q_2 = 1$.

Model-based sparse recovery: Nevertheless, sparsity is merely a first-order description of $\boldsymbol{\beta}^*$ and in many applications we have considerably more information *a priori*. In this work, we consider the *k-sparse block model* [29, 3, 22]:

Definition 1.2. We denote a block-sparse structure by $\mathcal{M} := \{\mathcal{G}_1, \dots, \mathcal{G}_M\}$ where $\mathcal{G}_i \subseteq [p]$, $|\mathcal{G}_i| = g$ for $i = 1, \dots, M$, $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset$, $i \neq j$, and M is the total number of groups. Given a group structure \mathcal{M} , the *k-sparse block model* is defined as the collection of sets $\mathcal{M}_k := \{\mathcal{G}_{i_1}, \dots, \mathcal{G}_{i_k}\}$ of k groups from \mathcal{M} .

We consider block-sparse models such that $\cup_{\mathcal{G} \in \mathcal{M}} \mathcal{G} \equiv \{1, \dots, p\}$ and $g = p/M$ where M is the number of groups. The idea behind group models is the identification of group of variates that should be either selected or discarded (*i.e.*, set to zero) together. Such settings naturally appear in applications such as gene expression data [30] and neuroimaging [13].

As an indicator of what can be achieved, Model-based Compressive Sensing [4], leverages such models with *dense sensing matrices* to provably reduce the number of measurements for stable recovery from $\mathcal{O}(kg \log(p/(kg)))$ to $\mathcal{O}(kg + k \log(M/k))$ using *non-convex* schemes²; see [14] for a more recent discussion and extensions on this matter. Along this research direction, [17] and [2] propose non-convex algorithms for equally-sized and variable-sized block-sparse signal recovery, respectively, using sparse matrices. To show reduction in sampling complexity, \mathbf{X} is assumed to satisfy the *model RIP-1* [17]:³

$$(1 - 2\epsilon_{\mathcal{M}_k})d\|\boldsymbol{\beta}\|_1 \leq \|\mathbf{X}\boldsymbol{\beta}\|_1 \leq d\|\boldsymbol{\beta}\|_1, \quad (5)$$

$\forall k$ -block-sparse vector $\boldsymbol{\beta} \in \mathbb{R}^p$, where $\epsilon_{\mathcal{M}_k}$ represents the *model-based expansion* constant.

1.1 Contributions

Restricting error guarantees to only variations of standard ℓ_{q_1}/ℓ_{q_2} -norms might be inconvenient in some applications. Here, we broaden the results in [16] towards having non-standard ℓ_q distances in approximation guarantees. To the best of our knowledge, this work is the first attempt for provable convex recovery with approximate guarantees in the appropriate norm and using sparse matrices. Similar results – but for a different model – for the case of dense (Gaussian) sensing matrices are presented in [12].

In particular, in Section 4, we provide provable error guarantees for the convex criterion:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_{2,1} \quad \text{subject to: } \mathbf{y} = \mathbf{X}\boldsymbol{\beta}, \quad (6)$$

where $\|\cdot\|_{2,1}$ is the $\ell_{2,1}$ -norm; see next Section for details. Our key novelty is to provide $\ell_{2,1}/\ell_{2,1}$ approximation guarantees using sparse matrices for sensing; see Section 5. In practice, we show the merits of using sparse matrices in convex approaches for (6), both on synthetic and real data cases; see Section 7.

²Most of the non-convex approaches known heretofore consider a (block) sparse constrained optimization criterion, where one is minimizing a data fidelity term (*e.g.*, a least-squares function) over cardinality constraints.

³While probabilistic constructions of expander matrices satisfying model RIP-1 coincide with those of regular RIP-1 expander matrices, the model-based assumption on signals $\boldsymbol{\beta}$ in (5) results into constructions of expanders \mathbf{X} with less number of rows and guaranteed signal recovery.

2 Preliminaries

Scalars are mostly denoted by plain letters (*e.g.* k, p), vectors by lowercase boldface letters (*e.g.*, \mathbf{x}), matrices by uppercase boldface letters (*e.g.*, \mathbf{A}) and sets by uppercase calligraphic letters (*e.g.*, \mathcal{S}), with the exception of $[p]$ which denotes the index set $\{1, \dots, p\}$. Given a set \mathcal{S} in item space \mathcal{I} such that $\mathcal{S} \subseteq \mathcal{I}$, we denote its complement by $\mathcal{I} \setminus \mathcal{S}$; we also use the notation \mathcal{S}^c , when the item space is clearly described by context. For vector $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\beta}_{\mathcal{S}}$ denotes the restriction of $\boldsymbol{\beta}$ onto \mathcal{S} , *i.e.*, $(\boldsymbol{\beta}_{\mathcal{S}})_i = \beta_i$ if $i \in \mathcal{S}$ and 0 otherwise. We use $|\mathcal{S}|$ to denote the cardinality of a set \mathcal{S} . The ℓ_p norm of a vector $\mathbf{x} \in \mathbb{R}^p$ is defined as $\|\mathbf{x}\|_p := (\sum_{i=1}^p |x_i|^p)^{1/p}$.

The $\ell_{2,1}$ -norm: The first convex relaxation for block-sparse approximation is due to [29], who propose group sparsity-inducing $\ell_{2,1}$ convex norm:

$$\|\boldsymbol{\beta}\|_{2,1} := \sum_{\mathcal{G} \in \mathcal{M}} w_{\mathcal{G}} \|\boldsymbol{\beta}_{\mathcal{G}}\|_2, \quad w_{\mathcal{G}} > 0.$$

The $\ell_{2,1}$ -norm construction follows group supports according to a *predefined* model \mathcal{M} and promotes sparsity in the group level. We assume all the $w_i = 1$, but may be otherwise for a more general setting [19].

Given a vector $\mathbf{w} \in \mathbb{R}^p$, we define \mathcal{M}_k^* as the *best k -block sparse index set* according to:

$$\mathcal{M}_k^* \in \operatorname{argmin}_{\mathcal{M}_k \subseteq \mathcal{M}} \|\mathbf{w} - \mathbf{w}_{\mathcal{M}_k}\|_{2,1}$$

3 Problem statement

To properly set up the problem and our results, consider the following question:

QUESTION: Let \mathcal{M} be a predefined and known a priori block sparse model of M groups of size g and $k > 0$ be a user-defined group-sparsity level. Consider $\mathbf{X} \in \{0, 1\}^{n \times p}$ ($n < p$) be a known $(k \cdot g, d, \epsilon_{\mathcal{M}_k})$ expander sparse matrix, satisfying model-RIP-1 in (5) for some degree $d > 0$ and $\epsilon_{\mathcal{M}_k} \in (0, 1/2)$. Assume an unknown vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is observed through $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^*$. Using criterion (6) to obtain a solution $\hat{\boldsymbol{\beta}}$ such that $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$, can we achieve a constant approximation of the form:

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{2,1} \leq C_1 \cdot \|\boldsymbol{\beta}^* - \boldsymbol{\beta}_{\mathcal{M}_k^*}^*\|_{2,1}, \quad C_1 > 0,$$

where \mathcal{M}_k^* contains the groups of the best k -block sparse approximation of $\boldsymbol{\beta}^*$?

4 Main result

In this section, we answer affirmatively to the above question with the following result:

Theorem 4.1. Let $\mathbf{X} \in \{0, 1\}^{n \times p}$ be a $(k \cdot g, d, \epsilon_{\mathcal{M}_k})$ expander, satisfying model-RIP-1 (5) for block sparse model \mathcal{M} , as in Definition 1.2. Consider two vectors $\boldsymbol{\beta}^*, \hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ such that $\mathbf{X}\boldsymbol{\beta}^* = \mathbf{X}\hat{\boldsymbol{\beta}}$, where $\boldsymbol{\beta}^*$ is unknown and $\hat{\boldsymbol{\beta}}$ is the solution to (6). Then, $\|\hat{\boldsymbol{\beta}}\|_{2,1} \leq \|\boldsymbol{\beta}^*\|_{2,1}$ and:

$$\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_{2,1} \leq \frac{2}{1 - \frac{4\epsilon_{\mathcal{M}_k} \cdot g}{1 - 2\epsilon_{\mathcal{M}_k}}} \cdot \|\boldsymbol{\beta}^* - \boldsymbol{\beta}_{\mathcal{M}_k^*}^*\|_{2,1}.$$

In other words, Theorem 4.1 shows that, given a proper sparse matrix \mathbf{X} satisfying the model-RIP-1 (5) for structure \mathcal{M} , the $\ell_{2,1}$ -convex program in (6) provides a good block-sparse solution $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$. Next, we also state a corollary to the theorem above for the more realistic noisy case.

Corollary 4.2. Assume the setting in Theorem 4.1 and let two vectors $\boldsymbol{\beta}^*, \hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ such that $\|\mathbf{X}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})\|_1 = \gamma \geq 0$, where $\boldsymbol{\beta}^*$ is unknown and $\hat{\boldsymbol{\beta}}$ is the solution to (6). Then, $\|\hat{\boldsymbol{\beta}}\|_{2,1} \leq \|\boldsymbol{\beta}^*\|_{2,1}$ and:

$$\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_{2,1} \leq \frac{2}{1 - \frac{4\epsilon_{\mathcal{M}_k} \cdot g}{1 - 2\epsilon_{\mathcal{M}_k}}} \cdot \|\boldsymbol{\beta}^* - \boldsymbol{\beta}_{\mathcal{M}_k^*}^*\|_{2,1} + \frac{\gamma}{1 - \frac{4\epsilon_{\mathcal{M}_k} \cdot g}{1 - 2\epsilon_{\mathcal{M}_k}}}.$$

5 Proof of Theorem 4.1

Let us consider two solutions to the set of linear equations $\mathbf{X}\boldsymbol{\beta}^* = \mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$, where $\boldsymbol{\beta}^*, \hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ and let $\mathbf{X} \in \{0, 1\}^{n \times p}$ be a given expander satisfying (5). We define $\mathbf{z} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in \ker(\mathbf{X})$ since $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\boldsymbol{\beta}^* \implies \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \mathbf{0}$.

Using the group model formulation, each vector $\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*$ can be decomposed as: $\widehat{\boldsymbol{\beta}} = \sum_{\mathcal{G} \in \mathcal{M}} \widehat{\boldsymbol{\beta}}_{\mathcal{G}}$ and $\boldsymbol{\beta}^* = \sum_{\mathcal{G} \in \mathcal{M}} \boldsymbol{\beta}_{\mathcal{G}}^*$, where $\text{supp}(\widehat{\boldsymbol{\beta}}_{\mathcal{G}}) = \mathcal{G}$, $\text{supp}(\boldsymbol{\beta}_{\mathcal{G}}^*) = \mathcal{G}$, $\forall \mathcal{G} \in \mathcal{M}$; moreover, assume that $\widehat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*$. Thus, given the above decompositions, we also have:

$$\mathbf{z} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = \sum_{\mathcal{G} \in \mathcal{M}} (\widehat{\boldsymbol{\beta}}_{\mathcal{G}} - \boldsymbol{\beta}_{\mathcal{G}}^*) =: \sum_{\mathcal{G} \in \mathcal{M}} \mathbf{z}_{\mathcal{G}}, \quad (7)$$

where $\text{supp}(\mathbf{z}_{\mathcal{G}}) = \mathcal{G}$. By assumption, we know that $\|\boldsymbol{\beta}^*\|_{2,1} \geq \|\widehat{\boldsymbol{\beta}}\|_{2,1}$. Using the definition of the $\|\cdot\|_{2,1}$, we have:

$$\sum_{\mathcal{G} \in \mathcal{M}} \|\boldsymbol{\beta}_{\mathcal{G}}^*\|_2 \geq \sum_{\mathcal{G} \in \mathcal{M}} \|\widehat{\boldsymbol{\beta}}_{\mathcal{G}}\|_2 = \sum_{\mathcal{G} \in \mathcal{M}} \|\mathbf{z}_{\mathcal{G}} + \boldsymbol{\beta}_{\mathcal{G}}^*\|_2 \quad (8)$$

Denote the set of k groups in the best k -block sparse approximation of $\boldsymbol{\beta}^*$ as \mathcal{M}_k^* . Then,

$$\begin{aligned} \sum_{\mathcal{G} \in \mathcal{M}} \|\boldsymbol{\beta}_{\mathcal{G}}^*\|_2 &\geq \sum_{\mathcal{G} \in \mathcal{M}_k^*} \|\mathbf{z}_{\mathcal{G}} + \boldsymbol{\beta}_{\mathcal{G}}^*\|_2 + \sum_{\mathcal{G} \in \mathcal{M} \setminus \mathcal{M}_k^*} \|\mathbf{z}_{\mathcal{G}} + \boldsymbol{\beta}_{\mathcal{G}}^*\|_2 \\ &\geq \sum_{\mathcal{G} \in \mathcal{M}_k^*} \|\boldsymbol{\beta}_{\mathcal{G}}^*\|_2 - \sum_{\mathcal{G} \in \mathcal{M}_k^*} \|\mathbf{z}_{\mathcal{G}}\|_2 \\ &\quad + \sum_{\mathcal{G} \in \mathcal{M} \setminus \mathcal{M}_k^*} \|\mathbf{z}_{\mathcal{G}}\|_2 - \sum_{\mathcal{G} \in \mathcal{M} \setminus \mathcal{M}_k^*} \|\boldsymbol{\beta}_{\mathcal{G}}^*\|_2 \\ &= \sum_{\mathcal{G} \in \mathcal{M}} \|\boldsymbol{\beta}_{\mathcal{G}}^*\|_2 - 2 \sum_{\mathcal{G} \in \mathcal{M} \setminus \mathcal{M}_k^*} \|\boldsymbol{\beta}_{\mathcal{G}}^*\|_2 \\ &\quad + \sum_{\mathcal{G} \in \mathcal{M}} \|\mathbf{z}_{\mathcal{G}}\|_2 - 2 \sum_{\mathcal{G} \in \mathcal{M}_k^*} \|\mathbf{z}_{\mathcal{G}}\|_2 \end{aligned} \quad (9)$$

where the first equality is due to the block-sparse model and the first inequality is due to the triangle inequality. To proceed, we require the following Lemma; the proof is based on [5].

Lemma 5.1. *Let $\mathbf{z} \in \mathbb{R}^p$ such that $\mathbf{X}\mathbf{z} = \mathbf{0}$ and $\mathbf{z} = \sum_{\mathcal{G} \in \mathcal{M}} \mathbf{z}_{\mathcal{G}}$, where $\text{supp}(\mathbf{z}_{\mathcal{G}}) \subseteq \mathcal{G}, \forall \mathcal{G} \in \mathcal{M}$. Then,*

$$\sum_{\mathcal{G} \in \mathcal{M}_k^*} \|\mathbf{z}_{\mathcal{G}}\|_2 \leq \frac{2\epsilon_{\mathcal{M}_k} \cdot g}{1 - 2\epsilon_{\mathcal{M}_k}} \cdot \sum_{\mathcal{G} \in \mathcal{M}} \|\mathbf{z}_{\mathcal{G}}\|_2, \quad (10)$$

where $\epsilon_{\mathcal{M}_k} \in (0, 1/2)$ denotes the model-based expansion parameter of expander matrix $\mathbf{X} \in \{0, 1\}^{n \times p}$.

Proof. Following [5], we assume the following decomposition of p indices, according to \mathcal{M} : Due to the non-overlapping nature of \mathcal{M} , we split coordinates in p into k -block sparse index sets $\mathcal{M}_k^0 \equiv \mathcal{M}_k^*, \mathcal{M}_k^1, \mathcal{M}_k^2, \dots, \mathcal{M}_k^t$ such that (i) each $\mathcal{M}_k^l, \forall l$, has k groups (except probably \mathcal{M}_k^t), (ii) each group $\mathcal{G} \in \mathcal{M}_k^l$ has g indices and (iii) there is ordering on groups such that:

$$\|\mathbf{z}_{\mathcal{G}}\|_2 \geq \|\mathbf{z}_{\mathcal{G}'}\|_2, \quad \forall \mathcal{G} \in \mathcal{M}_k^l, \forall \mathcal{G}' \in \mathcal{M}_k^q \text{ s.t. } l \leq q.$$

Since $\mathbf{z} \in \ker(\mathbf{X})$, we have $0 = \|\mathbf{X}\mathbf{z}\|_1$. Moreover, we denote as $\Gamma(\mathcal{M}_k^*)$ the set of indices of the rows of \mathbf{X} that correspond to the neighbours of left-nodes in \mathcal{M}_k^* ; see Figure 1. Thus, without loss of generality, we reorder the rows of \mathbf{X} such that the top $|\cup_{\mathcal{G} \in \mathcal{M}_k^*} \mathcal{G}|$ rows are indexed by $\Gamma(\mathcal{M}_k^*)$ —the union of the set of neighbors of all $\mathcal{G} \in \mathcal{M}_k^*$. Given the above partition, we denote the submatrix of \mathbf{X} composed of these rows as \mathbf{X}_{Γ} , such that

$$\mathbf{X} = \begin{pmatrix} \text{---} & \mathbf{X}_{\Gamma} & \text{---} \\ \text{---} & & \text{---} \\ \text{---} & \mathbf{X}_{\Gamma^c} & \text{---} \end{pmatrix}.$$

Thus, we have:

$$\begin{aligned} 0 &= \|\mathbf{X}\mathbf{z}\|_1 = \|\mathbf{X}_{\Gamma}\mathbf{z}\|_1 = \|\mathbf{X}_{\Gamma} \cdot \sum_{\mathcal{G} \in \mathcal{M}} \mathbf{z}_{\mathcal{G}}\|_1 \\ &= \|\mathbf{X}_{\Gamma} \cdot \left(\sum_{\mathcal{G} \in \mathcal{M}_k^*} \mathbf{z}_{\mathcal{G}} + \sum_{\mathcal{G} \in \mathcal{M} \setminus \mathcal{M}_k^*} \mathbf{z}_{\mathcal{G}} \right)\|_1 \\ &= \|\mathbf{X}_{\Gamma} \cdot \left(\sum_{\mathcal{G} \in \mathcal{M}_k^0} \mathbf{z}_{\mathcal{G}} + \sum_{\mathcal{G} \in \mathcal{M} \setminus \mathcal{M}_k^0} \mathbf{z}_{\mathcal{G}} \right)\|_1 \\ &\geq \|\mathbf{X}_{\Gamma} \cdot \sum_{\mathcal{G} \in \mathcal{M}_k^0} \mathbf{z}_{\mathcal{G}}\|_1 - \|\mathbf{X}_{\Gamma} \cdot \sum_{\mathcal{G} \in \mathcal{M} \setminus \mathcal{M}_k^0} \mathbf{z}_{\mathcal{G}}\|_1 \end{aligned}$$

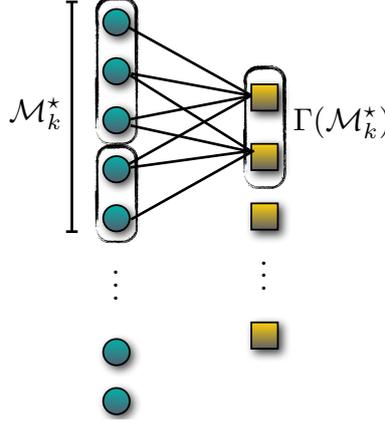


Figure 1: Representation of neighbours of left-nodes, indexed by the k -group sparse set \mathcal{M}_k^* .

By using the model-RIP-1 property (5) of \mathbf{X} on the input vector $\sum_{\mathcal{G} \in \mathcal{M}_k^0} \mathbf{z}_{\mathcal{G}}$ we have:

$$\begin{aligned}
\|\mathbf{X}_{\Gamma} \cdot \sum_{\mathcal{G} \in \mathcal{M}_k^0} \mathbf{z}_{\mathcal{G}}\|_1 &= \|\mathbf{X} \cdot \sum_{\mathcal{G} \in \mathcal{M}_k^0} \mathbf{z}_{\mathcal{G}}\|_1 \\
&\geq (1 - 2\epsilon_{\mathcal{M}_k}) \cdot d \cdot \left\| \sum_{\mathcal{G} \in \mathcal{M}_k^0} \mathbf{z}_{\mathcal{G}} \right\|_1 \\
&= (1 - 2\epsilon_{\mathcal{M}_k}) \cdot d \cdot \sum_{\mathcal{G} \in \mathcal{M}_k^0} \|\mathbf{z}_{\mathcal{G}}\|_1 \\
&\geq (1 - 2\epsilon_{\mathcal{M}_k}) \cdot d \cdot \sum_{\mathcal{G} \in \mathcal{M}_k^0} \|\mathbf{z}_{\mathcal{G}}\|_2
\end{aligned}$$

where the first equality is due to $\mathbf{X}_{\Gamma^c} \cdot \sum_{\mathcal{G} \in \mathcal{M}_k^0} \mathbf{z}_{\mathcal{G}} = \mathbf{0}$, the second equality due to the non-overlapping groups and the last inequality since $\|\boldsymbol{\beta}\|_1 \geq \|\boldsymbol{\beta}\|_2, \forall \boldsymbol{\beta} \in \mathbb{R}^p$.

Therefore, we have:

$$\begin{aligned}
0 &\geq (1 - 2\epsilon_{\mathcal{M}_k}) \cdot d \cdot \sum_{\mathcal{G} \in \mathcal{M}_k^0} \|\mathbf{z}_{\mathcal{G}}\|_2 - \|\mathbf{X}_{\Gamma} \sum_{\mathcal{G} \in \mathcal{M} \setminus \mathcal{M}_k^0} \mathbf{z}_{\mathcal{G}}\|_1 \\
&\geq (1 - 2\epsilon_{\mathcal{M}_k}) \cdot d \cdot \sum_{\mathcal{G} \in \mathcal{M}_k^0} \|\mathbf{z}_{\mathcal{G}}\|_2 - \sum_{\mathcal{G} \in \mathcal{M} \setminus \mathcal{M}_k^0} \|\mathbf{X}_{\Gamma} \mathbf{z}_{\mathcal{G}}\|_1 \\
&= (1 - 2\epsilon_{\mathcal{M}_k}) \cdot d \cdot \sum_{\mathcal{G} \in \mathcal{M}_k^0} \|\mathbf{z}_{\mathcal{G}}\|_2 - \sum_{l \geq 1} \sum_{\mathcal{G} \in \mathcal{M}_k^l} \sum_{\substack{(i,j) \in \mathcal{E} \\ i \in \mathcal{G} \\ j \in \Gamma}} |(\mathbf{z}_{\mathcal{G}})_i| \\
&\geq (1 - 2\epsilon_{\mathcal{M}_k}) \cdot d \cdot \sum_{\mathcal{G} \in \mathcal{M}_k^0} \|\mathbf{z}_{\mathcal{G}}\|_2 - \sum_{l \geq 1} |\# \text{ edges } (i, j) : i \in \mathcal{G}, j \in \Gamma, \mathcal{G} \in \mathcal{M}_k^l| \cdot \max_{\mathcal{G} \in \mathcal{M}_k^l} \max_{i \in \mathcal{G}} |(\mathbf{z}_{\mathcal{G}})_i|
\end{aligned}$$

The quantity $|\# \text{ edges } (i, j) : i \in \mathcal{G}, j \in \Gamma, \mathcal{G} \in \mathcal{M}_k^l|$ further satisfies:

$$\begin{aligned}
|\# \text{ edges } (i, j) : i \in \mathcal{G}, j \in \Gamma, \mathcal{G} \in \mathcal{M}_k^l| &= |\Gamma(\mathcal{M}_k^0) \cap \Gamma(\mathcal{M}_k^l)| \\
&\stackrel{(i)}{=} |\Gamma(\mathcal{M}_k^0)| + |\Gamma(\mathcal{M}_k^l)| - |\Gamma(\mathcal{M}_k^0) \cup \Gamma(\mathcal{M}_k^l)| \\
&\leq d \cdot |\mathcal{M}_k^0| + d \cdot |\mathcal{M}_k^l| - |\Gamma(\mathcal{M}_k^0) \cup \Gamma(\mathcal{M}_k^l)| \\
&\leq 2d \cdot k \cdot g - |\Gamma(\mathcal{M}_k^0) \cup \Gamma(\mathcal{M}_k^l)| \\
&\stackrel{(ii)}{\leq} 2d \cdot k \cdot g - d(1 - \epsilon_{\mathcal{M}_k}) |\mathcal{M}_k^0 \cup \mathcal{M}_k^l| \\
&\stackrel{(iii)}{\leq} 2d \cdot k \cdot \epsilon_{\mathcal{M}_k} \cdot g
\end{aligned}$$

where (i) is due to the inclusion-exclusion principle, (ii) is due to the expansion property and, (iii) is due to $|\mathcal{G}| = g, \forall \mathcal{G} \in \mathcal{M}$.

Thus, the above inequality becomes:

$$0 \geq (1 - 2\epsilon_{\mathcal{M}_k}) \cdot d \cdot \sum_{\mathcal{G} \in \mathcal{M}_k^0} \|\mathbf{z}_{\mathcal{G}}\|_2 - 2d \cdot k \cdot \epsilon_{\mathcal{M}_k} \cdot g \sum_{l \geq 1} \max_{\mathcal{G} \in \mathcal{M}_k^l} \max_{i \in \mathcal{G}} |(\mathbf{z}_{\mathcal{G}})_i|$$

Let \mathcal{G}_{\max}^l denote the group that contains the maximizing index i_{\max} such that:

$$i_{\max} \in \underset{i \in \mathcal{G} \text{ s.t. } \mathcal{G} \in \mathcal{M}_k^l}{\operatorname{argmax}} |(\mathbf{z}_{\mathcal{G}})_i|.$$

Then:

$$0 \geq (1 - 2\epsilon_{\mathcal{M}_k}) \cdot d \cdot \sum_{\mathcal{G} \in \mathcal{M}_k^0} \|\mathbf{z}_{\mathcal{G}}\|_2 - 2d \cdot k \cdot \epsilon_{\mathcal{M}_k} \cdot g \sum_{l \geq 1} \|\mathbf{z}_{\mathcal{G}_{\max}^l}\|_{\infty}$$

However, due to the ordering of ℓ_2 -norm of groups:

$$\|\mathbf{z}_{\mathcal{G}_{\max}^l}\|_{\infty} \leq \|\mathbf{z}_{\mathcal{G}_{\max}^l}\|_2 \leq \min_{\mathcal{G} \in \mathcal{M}_k^{l-1}} \|\mathbf{z}_{\mathcal{G}}\|_2 \leq \frac{1}{k} \sum_{\mathcal{G} \in \mathcal{M}_k^{l-1}} \|\mathbf{z}_{\mathcal{G}}\|_2,$$

we further have:

$$\begin{aligned} 0 &\geq (1 - 2\epsilon_{\mathcal{M}_k}) \cdot d \cdot \sum_{\mathcal{G} \in \mathcal{M}_k^0} \|\mathbf{z}_{\mathcal{G}}\|_2 - 2d \cdot k \cdot \epsilon_{\mathcal{M}_k} \cdot g \sum_{l \geq 1} \frac{1}{k} \sum_{\mathcal{G} \in \mathcal{M}_k^{l-1}} \|\mathbf{z}_{\mathcal{G}}\|_2 \\ &\geq (1 - 2\epsilon_{\mathcal{M}_k}) \cdot d \cdot \sum_{\mathcal{G} \in \mathcal{M}_k^0} \|\mathbf{z}_{\mathcal{G}}\|_2 - 2d \cdot \epsilon_{\mathcal{M}_k} \cdot g \sum_{\mathcal{G} \in \mathcal{M}} \|\mathbf{z}_{\mathcal{G}}\|_2 \end{aligned}$$

which leads to:

$$\sum_{\mathcal{G} \in \mathcal{M}_k^0} \|\mathbf{z}_{\mathcal{G}}\|_2 \leq \frac{2\epsilon_{\mathcal{M}_k} \cdot g}{1 - 2\epsilon_{\mathcal{M}_k}} \cdot \sum_{\mathcal{G} \in \mathcal{M}} \|\mathbf{z}_{\mathcal{G}}\|_2$$

□

Using (10) in (9), we further have:

$$\begin{aligned} 0 &\geq -2 \sum_{\mathcal{G} \in \mathcal{M} \setminus \mathcal{M}_k^0} \|\beta_{\mathcal{G}}^*\|_2 + \sum_{\mathcal{G} \in \mathcal{M}} \|\mathbf{z}_{\mathcal{G}}\|_2 - \frac{4\epsilon_{\mathcal{M}_k} \cdot g}{1 - 2\epsilon_{\mathcal{M}_k}} \cdot \sum_{\mathcal{G} \in \mathcal{M}} \|\mathbf{z}_{\mathcal{G}}\|_2 \implies \\ \sum_{\mathcal{G} \in \mathcal{M}} \|\mathbf{z}_{\mathcal{G}}\|_2 &\leq \frac{2}{1 - \frac{4\epsilon_{\mathcal{M}_k} \cdot g}{1 - 2\epsilon_{\mathcal{M}_k}}} \sum_{\mathcal{G} \in \mathcal{M} \setminus \mathcal{M}_k^0} \|\beta_{\mathcal{G}}^*\|_2, \end{aligned}$$

which is the desired result.

6 Comparison to state-of-the-art

To justify our theoretical results, here we compare with the only known results on *convex sparse recovery using expander matrices* of [5]. We note that, in [5], no apriori knowledge is assumed, beyond plain sparsity.

We start our discussion with the following remark.

Remark 6.1. *In the extreme case of $g = 1$, Lemma 5.1 is analogous to Lemma 16 of [5]. To see this, observe that $\|\mathbf{z}_{\mathcal{G}}\|_2 = |(\mathbf{z})_i|$ for $g = 1$, where $|\mathcal{G}| = 1$ and $\mathcal{G} \equiv i \in \mathcal{M}_k^0$ or $\in \mathcal{M}$.*

We highlight that, as g grows, feasible values of $\epsilon_{\mathcal{M}_k} \rightarrow 0$, *i.e.*, we require more rows to construct an expander matrix \mathbf{X} with such expansion property. For simplicity, in the discussion below we use $\epsilon \equiv \epsilon_{\mathcal{M}_k}$ interchangeably, where the type of ϵ used is apparent from context.

In the case where we are *oblivious to any, a-priori known, structured sparsity model*, [5] prove the following error guarantees for the vanilla BP formulation (2), using expander matrices:

$$\|\beta^* - \hat{\beta}\|_1 \leq \frac{2}{1 - 4 \cdot \frac{\epsilon}{1 - 2\epsilon}} \cdot \|\beta^* - \beta_{\mathcal{S}}^*\|_1, \quad (11)$$

where $\mathcal{S} \subseteq [p]$ such that $|\mathcal{S}| = |\cup_{\mathcal{G} \in \mathcal{M}_k^*} \cup_{i \in \mathcal{G}} i|$; *i.e.*, we are looking for a solution of the same sparsity as the union of groups in the block-sparse case. In order to compare (11) with our result, a naive transformation of (11) into $\ell_{2,1}$ terms leads to:

$$\|\beta^* - \hat{\beta}\|_{2,1} \leq \frac{2\sqrt{g}}{1 - 4 \cdot \frac{\epsilon}{1 - 2\epsilon}} \cdot \|\beta^* - \beta_{\mathcal{M}_k^*}^*\|_{2,1}. \quad (12)$$

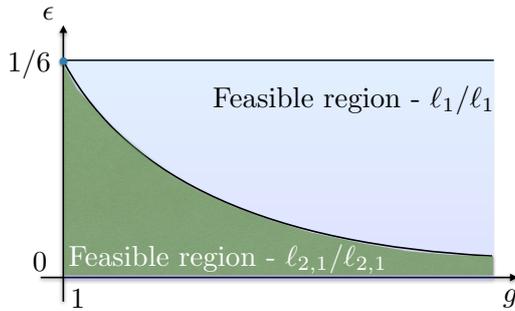


Figure 2: Feasibility regions for our approach and [5] as a function of ϵ and g variables.

To define the conditions under which (12) is valid, we require:

$$1 - 4 \cdot \frac{\epsilon}{1 - 2\epsilon} > 0 \Rightarrow \epsilon < 1/6,$$

i.e., ϵ is independent of g , while in our case, we have:

$$1 - 4 \cdot \frac{\epsilon \cdot g}{1 - 2\epsilon} > 0 \Rightarrow \epsilon < \frac{1}{2(1 + 2g)} \quad \left(\stackrel{g=1}{=} 1/6 \right);$$

i.e., our analysis provides weaker bounds with respect to the range of ϵ values such that the error guarantees is meaningful; see also Figure 2.

However, as already mentioned, (12) is *oblivious* to any model \mathcal{M} : *the solution $\hat{\beta}$ does not necessarily belong to \mathcal{M}* . This fact provides degrees of freedom to obtain better approximation constants. Nevertheless, this does not guarantee $\hat{\beta} \in \mathcal{M}$, considering only simple sparsity. Section 7 includes examples that highlight the superiority of $\ell_{2,1}$ -norm in optimization.

7 Experiments

We start our discussion (i) with a comparison between ℓ_1 - and $\ell_{2,1}$ -norm convex approaches, when the ground truth is known to be block sparse, and (ii) with a comparison between dense sub-Gaussian and sparse sensing matrices in (6), both w.r.t. sampling complexity and computational complexity requirements. We conclude with the task of recovering 2D images from compressed measurements, using block sparsity.

Solver. To solve both ℓ_1 - and $\ell_{2,1}$ -norm instances (2) and (6), we use the primal-dual convex optimization framework in [28], that solves (2) and (6) – among other objectives, by using ideas from the alternating direction methods of multipliers. Using the same framework for solution, we obtain more accurate and credible comparison results between different problem settings.

ℓ_1 - vs. $\ell_{2,1}$ -norm. In this experiment, we verify that *a priori structure knowledge helps in recovery*. To show this in the convex domain, consider the following artificial example. Let $\mathbf{y} = \mathbf{X}\beta^*$ be the set of observations, where $\beta^* \in \mathbb{R}^{10^3}$ is a k -block sparse vector, for $k = 8$. Here, we assume a block-sparse model \mathcal{M} with $M = 100$ non-overlapping groups. Observe that $g = p/M = 10$.

Both for ℓ_1 - and $\ell_{2,1}$ -norm cases, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is designed as an expander, with degree $d = \lceil 22 \cdot \log(M)/g \rceil = 11$, complying with our theory. Further, we make the convention $\mathbf{X} := 1/d \cdot \mathbf{X}$.

We consider two cases: (i) β^* is generated from a Gaussian distribution and (ii) β^* is generated as a binary signal $\in \{\pm 1\}$. In both cases, β^* is normalized such that $\|\beta^*\|_2 = 1$. For all cases, we used the same algorithmic framework [28] and performed 10 Monte Carlo realizations. Figure 3 shows the average probability for *successful recovery* as a function of the total number of measurements observed; we declare a success when $\|\hat{\beta} - \beta^*\|_2 \leq 10^{-5}$. It is apparent that knowing the model a priori can guarantee recovery with less number of observations.

Sub-Gaussian vs. sparse sensing matrices in $\ell_{2,1}$ -norm recovery. Let us focus now on the $\ell_{2,1}$ -norm case. For this setting, we perform two experiments. First, we consider a similar setting as previously; the only difference lies in the way we generate the sensing matrix \mathbf{X} . We consider two cases: (i) $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, 1/n \cdot \mathbf{I})$ and (ii) \mathbf{X} is a sparse expander matrix, again with degree $d = \lceil 22 \cdot \log(M)/g \rceil = 11$. Here, we only use the $\ell_{2,1}$ -norm solver. Figure 4 depicts some results. We observe that expander matrices perform worse – as a function of

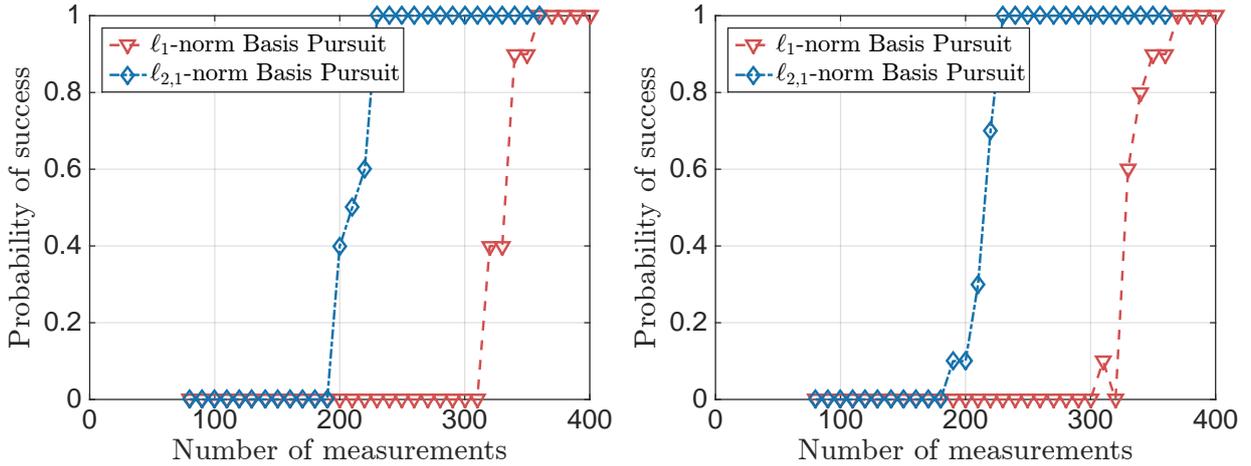


Figure 3: Average probability of successful recovery, as a function of the total number of observations, over 10 Monte Carlo iterations. *Top panel:* β^* is generated from a normal distribution. *Bottom panel:* β^* is generated from a Bernoulli distribution. $\ell_{2,1}$ -norm solver requires much less number of measurements for successful recovery, compared to ℓ_1 -norm solver.

the total number of measurements required for successful recovery. Nevertheless, using sparse matrices is still considerably competitive to dense Gaussian matrices.

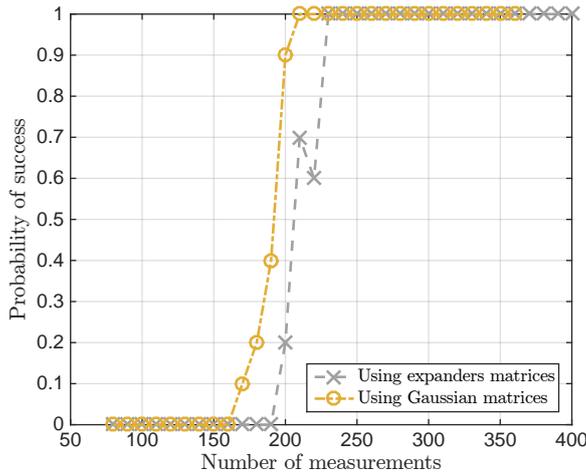


Figure 4: Average probability of successful recovery, as a function of the total number of observations, over 10 Monte Carlo iterations. Expander matrices perform worse but competitive to dense Gaussian matrices.

Now, to grasp the full picture in this setting, we scale our experiments to higher dimensions. Table 1 summons up the results. All experiments were repeated for 10 independent realizations and the table contains the median scores. For $p = \{10^4, 2 \cdot 10^4, 5 \cdot 10^4, 10^5\}$, the total number of non-overlapping groups in \mathcal{M} was $M = \{10^3, 2 \cdot 10^3, 5 \cdot 10^3, 5 \cdot 10^3\}$, respectively. The column cardinality is selected as $d = \lceil 22 \cdot \log(M)/g \rceil$. For each p , the block sparsity ranges as $k \in \{3, \dots, 6\}$. In all cases, $n = \lceil 0.4 \cdot p \rceil$.

One can observe that using sparse matrices in this setting results into *faster convergence* – as in total time required for stopping criterion to be met. Meanwhile, the solution quality is at least at the same levels, compared to that when dense Gaussian matrices are used.

Finally, we highlight that, for $p = 10^5$, since M does not increase, the number of non-zeros d per column decreases (*i.e.*, $d = 7$ while $d > 11$ in all other cases). This results into a small deterioration of the recovery quality; though, still comparable to the convex counterpart.⁴ Meanwhile, the time required for convergence in the expander case remains at the same levels as when $p = 5 \cdot 10^4$; however, the same does not apply for the

⁴Observe that in most configurations, expander matrices find a solution closer to β^* , compared to the dense setting, except for the case of $p = 10^5$, where we decrease the number of zeros per column.

Model		$\ \widehat{\beta} - \beta^*\ _2$		Time (sec)	
p	$k \cdot g$	Gaus.	Exp.	Gaus.	Exp.
10000	300	8.6e-07	3.3e-06	24.3	5.8
	400	8.2e-06	3.4e-06	27.5	6.2
	500	8.6e-06	3.2e-06	27.8	6.0
	600	8.6e-06	3.4e-06	31.2	7.6
20000	300	8.1e-07	3.4e-06	95.5	18.5
	400	8.1e-06	3.3e-06	79.4	16.4
	500	8.5e-06	3.4e-06	83.9	15.8
	600	8.5e-06	3.5e-06	91.3	18.9
50000	300	8.2e-06	3.3e-06	419.3	49.6
	400	8.1e-06	3.4e-06	432.8	45.8
	500	8.5e-06	3.6e-06	436.0	52.9
	600	8.4e-06	3.5e-06	435.4	51.1
100000	600	8.1e-06	9.4e-06	1585.5	55.1
	800	8.1e-06	9.5e-06	1598.2	54.5
	1000	8.4e-06	9.4e-06	1600.6	56.2
	1200	8.1e-06	9.3e-06	1648.0	55.6

Table 1: Summary of comparison results for reconstruction and efficiency. Median values are reported. As a stopping criterion, we used $\|\beta_{i+1} - \beta_i\|_2 / \|\beta_{i+1}\|_2 \leq 10^{-6}$, where β_i is the estimate at the i -th iteration. In all cases, $n = \lceil 0.4 \cdot p \rceil$.

dense counterpart. This constitutes the use of expander matrices appealing in real applications.

Block sparsity in image processing. For this experiment, we use the real background-subtracted image dataset, presented in [15]. Out of 205 frames, we randomly selected 100 frames to process. Each frame is rescaled to be of size $2^8 \times 2^8$ pixels. Each pixel takes values in $[0, 1]$. We observe $\mathbf{y} = \mathbf{X}\beta^*$ where $\beta^* \in [0, 1]^p$, $p = 2^{16}$, is the ground-truth vectorized frame and, \mathbf{X} is either sparse or dense sensing matrix, designed as in the previous experiments.

For the purpose of this experiment, we set up an upper wall time of 10^4 seconds (*i.e.*, 2.8 hours) to process 100 frames for each solver. This translates into 100 seconds per frame.

Due to the nature of the dataset, we can safely assume that nonzeros are clustered together. Thus, we assume group models \mathcal{M} where groups are constituted of consecutive column pixels and the p indices are divided in consecutive groups of equal size $g = \{4, 8, 16\}$. No other parameters are required – this is an advantage over non-convex approaches, where a sparsity level is usually required to be known a priori. All experiments are repeated 10 independent times for different \mathbf{X} ’s.

Figure 5 shows some representative results. Left panel illustrates the recovery performance for different settings – ℓ_1 - vs. $\ell_{2,1}$ -norm and Gaussian vs. sparse matrices \mathbf{X} ; results for other configurations are presented in the appendix. The first row considers a “simple” image with a small number of non-zeros; the other two rows show two less sparse cases. While for the “simple” case, solving $\ell_{2,1}$ -norm minimization with Gaussian matrices lead to better recovery results – within the time constraints, the same does not apply for the more “complex” cases. Overall, we observe that, *given such time restrictions per frame*, by using expander matrices one can achieve a better solution in terms of PSNR *faster*. This is shown in more detail in Figure 5 (right panel); see also the appendix for more results.

8 Conclusions

Sparse matrices are favorable objects in machine learning and optimization. When such matrices can be applied in favor of dense ones, the computational requirements can be significantly reduced in practice, both in terms of space and runtime complexity. In this work, we both show theoretically and experimentally that such selection is advantageous for the case of group-based basis pursuit recovery from linear measurements. As future work, one can consider other sparsity models, as well as different objective criteria.

References

- [1] K. Ahn, S. Guha, and A. McGregor. Graph sketches: Sparsification, spanners, and subgraphs. In *Proceedings of the 31st symposium on Principles of Database Systems*, pages 5–14. ACM, 2012.
- [2] B. Bah, L. Baldassarre, and V. Cevher. Model-based sketching and recovery with expanders. In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, pages 1529–1543. ACM–SIAM, 2014.

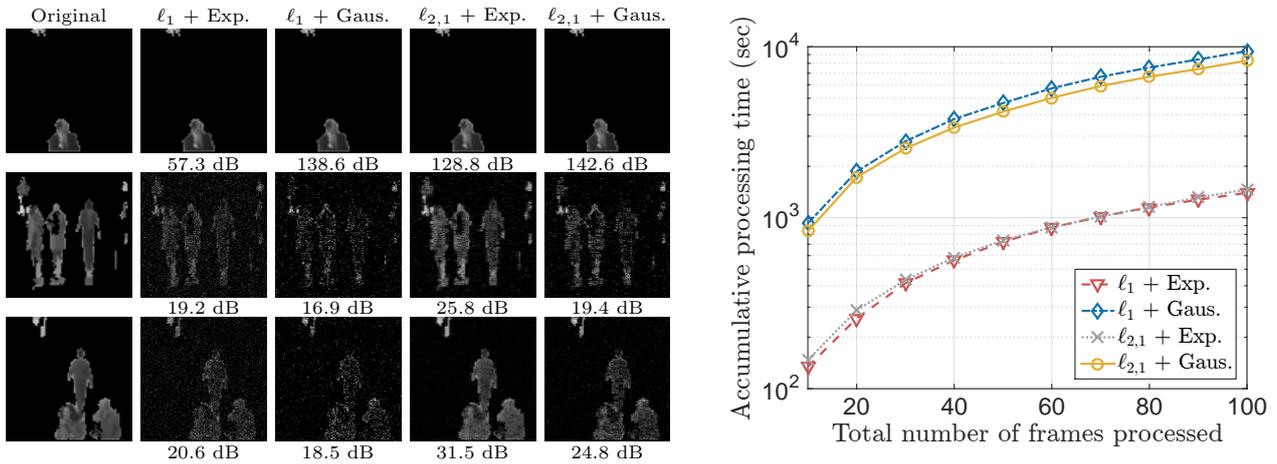


Figure 5: *Left panel*: representative examples of subtracted frame recovery from compressed measurements. Here, $n = \lceil 0.3 \cdot p \rceil$ measurements are observed for $p = 2^{16}$. Block sparse model \mathcal{M} contains groups of consecutive indices where $g = 4$. *Right panel*: Accumulative computational time required to process 100 frames. Overall, using Gaussian matrices in the $\ell_{2,1}$ -norm case, DECOPT required almost 2.8 hours (upper bound), as compared to 0.55 hours when \mathbf{X} is a sparse expander matrix. Thus, while Gaussian matrices is known to lead to better recovery results if no time restrictions apply, sparse sensing matrices constitute an appealing choice in practice.

- [3] Luca Baldassarre, Nirav Bhan, Volkan Cevher, Anastasios Kyriklidis, and Siddhartha Satpathi. Group-sparse model selection: Hardness and relaxations. *arXiv preprint arXiv:1303.3207*, 2013.
- [4] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *Information Theory, IEEE Transactions on*, 56(4):1982–2001, 2010.
- [5] R. Berinde, A. Gilbert, P. Indyk, H. Karloff, and M. Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. In *Communication, Control, and Computing, 2008 Allerton Conference on*, pages 798–805. IEEE, 2008.
- [6] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.
- [7] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.
- [8] E. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.
- [9] E. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, March 2008.
- [10] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998.
- [11] D. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.
- [12] Y. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *Information Theory, IEEE Transactions on*, 55(11):5302–5316, 2009.
- [13] A. Gramfort and M. Kowalski. Improving M/EEG source localization with an inter-condition sparse prior. In *ISBI'09. IEEE International Symposium on*, pages 141–144. IEEE, 2009.
- [14] Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt. A nearly-linear time framework for graph-structured sparsity. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 928–937, 2015.
- [15] Junzhou Huang, Tong Zhang, and Dimitris Metaxas. Learning with structured sparsity. *The Journal of Machine Learning Research*, 12:3371–3412, 2011.
- [16] P. Indyk and E. Price. K-median clustering, model-based compressive sensing, and sparse recovery for earth mover distance. In *Proceedings of ACM symposium on Theory of Computing*, pages 627–636. ACM, 2011.
- [17] P. Indyk and I. Razenshteyn. On model-based RIP-1 matrices. In *Automata, Languages, and Programming*, pages 564–575. Springer, 2013.
- [18] P. Indyk and M. Ruzic. Near-optimal sparse recovery in the ℓ_1 -norm. In *Foundations of Computer Science, 2008. IEEE 49th Annual IEEE Symposium on*, pages 199–207. IEEE, 2008.
- [19] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440. ACM, 2009.
- [20] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank. Efficient and robust compressed sensing using optimized expander graphs. *Information Theory, IEEE Transactions on*, 55(9):4299–4308, 2009.

- [21] A. Kyrillidis and V. Cevher. Recipes on hard thresholding methods. In *4th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, 2011.
- [22] Anastasios Kyrillidis, Luca Baldassarre, Marwa El Halabi, Quoc Tran-Dinh, and Volkan Cevher. Structured sparsity: Discrete and convex approaches. In *Compressed Sensing and its Applications*, pages 341–387. Springer, 2015.
- [23] Anastasios Kyrillidis and Volkan Cevher. Combinatorial selection and least absolute shrinkage via the CLASH algorithm. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, pages 2216–2220. IEEE, 2012.
- [24] D. Needell and J. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.
- [25] J. Nelson, H. Nguyen, and D. Woodruff. On deterministic sketching and streaming for sparse recovery and norm estimation. pages 627–638, 2012.
- [26] Eric Price. *Sparse recovery and Fourier sampling*. PhD thesis, Massachusetts Institute of Technology, 2013.
- [27] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [28] Quoc Tran-Dinh and Volkan Cevher. Constrained convex minimization via model-based excessive gap. In *Advances in Neural Information Processing Systems*, pages 721–729, 2014.
- [29] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.
- [30] H. Zhou, M. Sehl, J. Sinsheimer, and K. Lange. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, 26(19):2375, 2010.

9 Appendix

Here, we report further results on the 2D image recovery problem. We remind that, for the purpose of this experiment, we set up an upper wall time of 10^4 seconds (*i.e.*, 2.8 hours) to process 100 frames for each solver. This translates into 100 seconds per frame.

9.1 Varying group size g

For this case, we focus on a single frame. Due to its higher number of non-zeros, we have selected the frame shown in Figure 6. For this case, we consider a roughly sufficient number of measurements is acquired where $n = \lceil 0.3 \cdot p \rceil$. By varying the group size g , we obtain the results in Figure 6.

9.2 Varying number of measurements

Here, let $g = 4$ as this group selection performs better, as shown in the previous subsection. Here, we consider n take values from $n \in \lceil \{0.25, 0.3, 0.35, 0.4\} \cdot p \rceil$. The results, are shown in Figure 7.

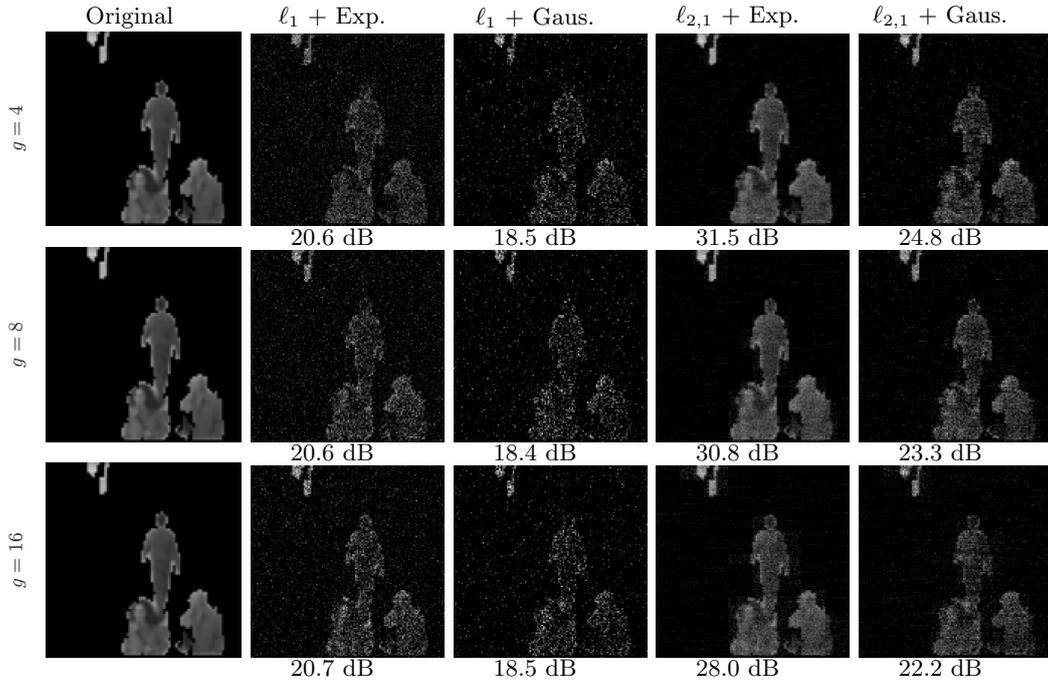


Figure 6: Results from real data. Representative examples of subtracted frame recovery from compressed measurements. Here, $n = \lceil 0.3 \cdot p \rceil$ measurements are observed for $p = 2^{16}$. From top to bottom, each line corresponds to block sparse model \mathcal{M} with groups of consecutive indices, where $g = 4$, $g = 8$, and $g = 16$, respectively. One can observe that one obtains worse recovery as the group size increases; thus a model with groups $g = 4$ is a good choice for this case.

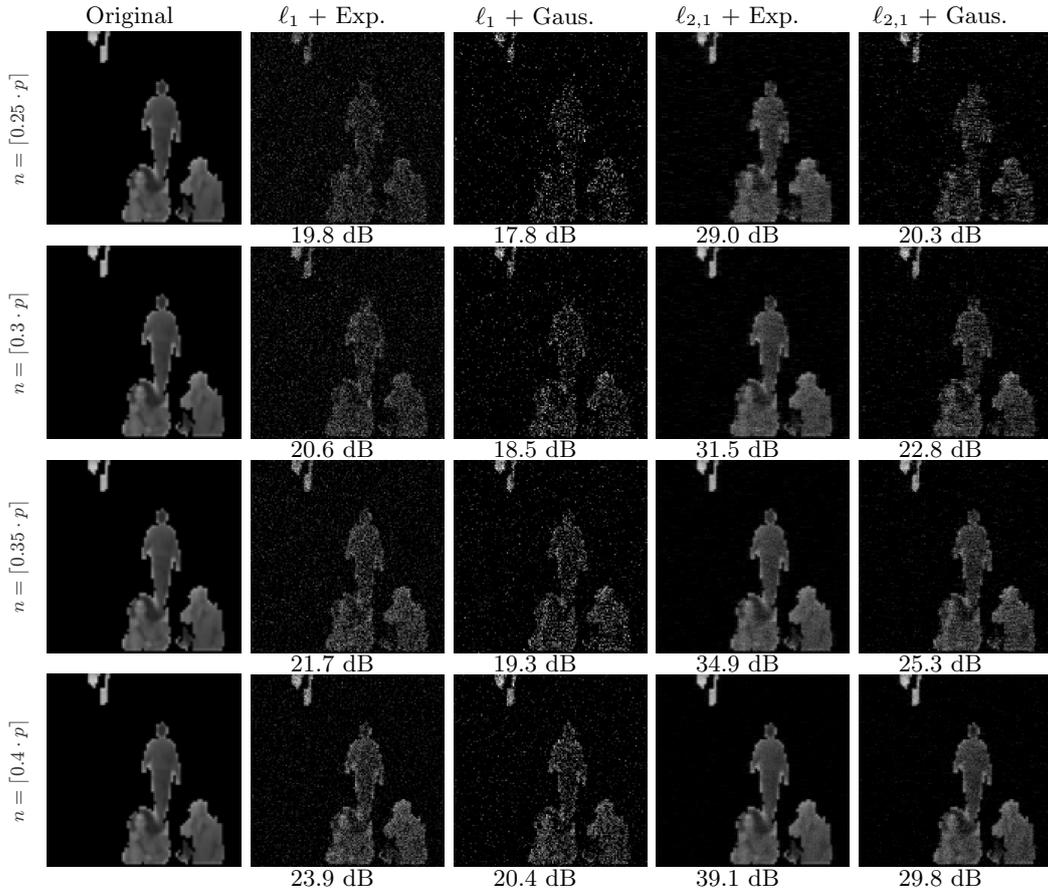


Figure 7: Results from real data. Representative examples of subtracted frame recovery from compressed measurements. Here, we consider a block sparse model fixed, with $g = 4$ block size per group. From top to bottom, the number of measurements range from $\lceil 0.25 \cdot p \rceil$ to $\lceil 0.4 \cdot p \rceil$, for $p = 2^{16}$. One can observe that one obtains better recovery as the number of measurements increases.