Submitted to the Annals of Applied Statistics

A GENERAL FRAMEWORK FOR HIGH-DIMENSIONAL ESTIMATION IN THE PRESENCE OF INCOHERENCE

By Yuxin Chen[*,†] and Sujay Sanghavi[*,‡]

Stanford University[†] and University of Texas at Austin[‡]

High-dimensional statistical inference requires the recovery/estimation of a large number of parameters from a possibly much smaller number of samples. A growing body of recent work has established that this is possible provided (a) the signal possesses an appropriate low-dimensional structure, and (b) the sampling is "incoherent", i.e. does not suppress this structure. Popular structural assumptions include sparsity, block-sparsity, low-rank etc., and popular recovery approaches include regularization via convex penalties, alternating projections, greedy approaches etc. However, in the existing literature, analysis for each combination of structure and method has proceeded on a case by case basis.

In this work we provide a unified framework that broadly characterizes when incoherence will enable consistent estimation in the high-dimensional setting. Specifically, we provide general definitions for structure and incoherence, and then establish that incoherence guarantees success in recovery (exactly in the noiseless case, and approximately in noisy case) for two broad classes of methods: *(a)* appropriate convex regularization, and *(b)* a new algorithm – Generalized Projections – that we propose. We identify several existing results that are recovered as special cases of each of our results. Our work builds on the recent framework for convex regularizers by Negahban *et.al.*; in particular one of our results is a characterization, in the presence of incoherence, of a crucial constant they define but do not evaluate in general. Finally, we also extend our framework to the case of multiple superimposed structures, where we define a new inter-structure notions of incoherence – Restricted Orthogonality Property.

1. Introduction. Recent advances have shown that, in several instances, it is possible to tractably recover a high-dimensional signal/object even when the number of measurements of the signal is far smaller than the ambient degrees of freedom, provided (a) the signal possesses an appropriate low-dimensional structure, and (b) the measurements are "incoherent", i.e. they do not suppress this structure.

[*]This work has been supported in part by NSF CAREER Grant 0954059 and NSF Grant 0964391.

Keywords and phrases: high dimensional estimation, incoherence

Indeed there is now an ever-expanding number of low-dimensional structures for which exact / approximate recovery has been established in the presence of conditions like (a) and (b) above. A specific (but incomplete) list includes (i) *sparsity:* methods like LASSO [1–4] for sparse regression, or compressed sensing [5–8] for exact or approximate recovery; (ii) *low-rank:* methods like principal component analysis (PCA) [9, 10], matrix completion [11–13] and recovery [14, 15], (iii) *block-sparsity:* methods like ℓ_1/ℓ_q regularization in multi-task learning / multivariate regression [16–18].

Simultaneously, there has also been an increasing set of methodological / algorithmic approaches for tractable recovery. Important broad approaches include regularization and convex optimization [19, 9, 20], iterative projection-based algorithms [21–23], greedy methods [24, 25] etc. Several results establish the success of these methods in performing structure recovery, again with assumptions like (a) and (b) above.

Analytical performance guarantees in all of the above have primarily been established on a *case by case basis*. In this paper, we develop a general framework and analysis for when incoherence enables the recovery of high-dimensional signals with low-dimensional structure; we recover several of the existing results above as special cases, often up to exact constants. In particular, we first develop a general notion of low-dimensional structures, extending the development in [26, 19]. We consider a general measure of incoherence, that we call Generalized Restricted Isometry Property (G-RIP), for these structures. We then establish our main results:

- *Recovery via convex optimization:* A recently popular approach to low-dimensional structure recovery is to minimize (in the noiseless case) or regularize with (in noisy case) a norm that penalizes deviations from the presumed structure. Theorem 1 and Corollary 1 in Section 3 show that, in our very general setting, G-RIP ensures (a) exact recovery in the noiseless case, and (b) approximate recovery in the noisy case. This result generalizes existing results [6] in Compressed Sensing, [14] in low-rank matrix recovery, [20] in block-sparse vectors, etc. We also capture the underlying connection between G-RIP and restricted strong convexity defined in [19]. A lower bound on the deviation parameter $\gamma(\mathcal{L})$ (which is crucial to their optimality conditions) has been derived in Theorem 2, which is not provided [19] in general settings.
- *Recovery via alternating projection:* Another approach to low-dimensional structure recovery, seen to be faster in practice and often with better provable guarantees, are methods that alternate between gradient descent and projection onto low-dimensional structures. Examples include GraDeS [22] for compressed sensing, and Singular Value Projec-

tion (SVP) [23] for low-rank matrix recovery. In Section 4 we propose the Generalized Projections algorithm, which recovers GraDeS and SVP as special cases. Theorem 3 shows the convergence to optimality of Generalized Projections; this recovers, exactly including preconstants, the performance guarantees previously shown for GraDeS and SVP.

- Multi-structure case: Verifying that RIP, or incoherence, holds for a particular structure and measurement is an often laborious process that happens on a case-by-case basis. This begs the natural question: if we have already established it for single structures (e.g. sparse matrices, low-rank matrices etc.), what *additional* property do we need when we have signals that can be separated into a mixture of single structures (e.g. a matrix where one part is sparse and the other low-rank). We establish a new property – *Restricted Orthogonality* – which we show is the only additional (and often easily verifiable) property we need, over and above individual RIPs, for recovery.

Some of our results are related to the recent work by Negahban *et. al.* [19] on convex regularizers, but our work contains several differences: *(a)* our framework for low-dimensional signals extends the one pioneered in [19]; in particular, we need to further develop the an appropriate formal notion of *hierarchy* of structures, as well as impose additional assumptions for a single structure; *(b)* [19] only provides results on convex relaxation, while we also characterize the success of lower-complexity alternating projection-based methods, and *(c)* we evaluate a critical constant – related to "restricted strong convexity" – that is defined in [19] but is not evaluated (or even shown to be non-zero) in general. In particular, [19] establishes general error bounds on the recovered signal, in terms of this constant. These bounds are only meaningful if we have knowledge of how the constant scales, or indeed if it is even non-zero. However, neither of these properties is established in general; it is only evaluated for special cases. We establish a general lower bound on this constant; doing so requires significant technical analysis and framework extension. Indeed our work builds on their work by establishing general conditions under which their bounds are meaningful.

2. Problem Setup and Key Definitions. We are given a few exact linear measurements of the form $b = \mathcal{A}(S) + w$ where w is some bounded noise obeying $\|w\|_F \leq \sigma$, and the task is to recover the high dimensional source signal S^*, given b and the map $\mathcal{A}$. We introduce several definitions, including the key one of Generalized Restricted Isometry Constant, that set up our general framework.

2.1. *Low-dimensional Structures.* The source signal S is assumed to be an $n_1 \times n_2$ matrix; thus $\mathbb{R}^{n_1 \times n_2}$ is the ambient high-dimensional space. Inspired by recent work [26, 19, 20], our general model of low-dimensional structure is to assume that $S \in \mathcal{V}$, where $\mathcal{V}$ is a *union of subspaces* of $\mathbb{R}^{n_1 \times n_2}$. As a concrete example, consider the low-dimensional structure represented by the set of all k-sparse matrices $\mathcal{V} = \{M : \|M\|_0 \leq k\}$. For each index set Ω containing at most k elements, we can associate a subspace $\{M : \mathrm{supp}(M) = \Omega\}$. Then the set of all k-sparse matrices is a union of these subspaces. Other examples include the set of all matrices with rank at most r, matrices with at most l non-zero rows, etc. It should be noted that this setup also incorporates multi-source multi-structure case (i.e. $\mathcal{V} = \{(S_1, \cdots, S_l) : S_1 \in \mathcal{V}_1, \cdots, S_l \in \mathcal{V}_l\}$), since we can consider the l distinct sources as a large matrix $[S_1, \cdots, S_l]$.

Denoting by $\oplus$ the Minkowski sum operator such that $\mathcal{V}_1 \oplus \mathcal{V}_2 = \{X_1 + X_2 \mid X_1 \in \mathcal{V}_1, X_2 \in \mathcal{V}_2\}$, we define the j^{th} Minkowski sum $\mathcal{V}^{(j)}$ of $\mathcal{V}$ as

$$\begin{aligned} \mathcal{V}^{(j)} &\triangleq \underbrace{\mathcal{V} \oplus \mathcal{V} \oplus \cdots \oplus \mathcal{V}}_{j^{\text{th}} \text{ iterate}} \\ &= \left\{ \sum_{k=1}^{j} X_k : \quad X_k \in \mathcal{V} \ \ (1 \leq k \leq j) \right\} \end{aligned} \tag{2.1}$$

$\mathcal{V}^{(j)}$ is thus another subspace union, which captures the *hierarchy* of specific low-dimensional space. For example, if $\mathcal{V}$ is the set of all rank-r matrices, then $\mathcal{V}^{(3)}$ is the one consisting of all matrices of rank at most $3r$, which is again a subspace union.

Additionally, the *projection* of any matrix X onto the subspace union $\mathcal{V}$ is defined as $\mathcal{P}_{\mathcal{V}}(X) = \arg\min_Y \{\|Y - X\|_F : Y \in \mathcal{V}\}$, with $\mathcal{P}_{\mathcal{V}^\perp}(X) = X - \mathcal{P}_{\mathcal{V}}(X)$ denoting its complement. Here $\|X\|_F := \sqrt{\sum_{i,j} X_{ij}^2}$ denotes the Frobenius norm.

Multi-Source Multi-Structure Case: We also consider the problem where there are several sources $S_1^*, \cdots, S_l^*$, each of which has its *own* low-dimensional structure (for example, one of them could be sparse, the other low-rank, and yet another block-sparse). We are interested in recovery given the *superimposed* observation $b = \mathcal{A}_1(S_1^*) + \cdots + \mathcal{A}_l(S_l^*) + w$ and the maps $\mathcal{A}_i (1 \leq i \leq l)$. Such a problem naturally arises in several applications [9, 27]. If we define S^* as $[S_1^*, \cdots, S_l^*]$, this problem can alternatively be formulated as $b = \mathcal{A}(S^*) + w$ using a generalized linear operator $\mathcal{A}$, which is a specialization of the general setting.

2.2. *Key Properties of the Linear Operators.* We want to obtain recovery from insufficient measurements using low-dimensional structure – clearly this will not always be possible. We now introduce the following properties on the maps $\mathcal{A}$ that we will restrict attention to, which is first defined as "$\mathcal{A}$-restricted isometry" in [26].

Definition 1 (**Generalized Restricted Isometry Property (G-RIP)**). For a subspace union $\mathcal{V}$, the restricted isometry constant $\delta_{\mathcal{V}}(\mathcal{A})$ of the operator $\mathcal{A}$ restricted to $\mathcal{V}$ is the smallest quantity such that

$$(1-\delta_{\mathcal{V}}(\mathcal{A}))\|X\|_F \leq \|\mathcal{A}(X)\|_F \leq (1+\delta_{\mathcal{V}}(\mathcal{A}))\|X\|_F \tag{2.2}$$

holds for all matrices $X \in \mathcal{V}$.

G-RIP is a natural generalization of similar properties for specific models like sparse vectors [6] and low-rank matrices [14], where it has been shown to enable recovery via tractable algorithms. Encouragingly, G-RIP will become the key property to ensure exact recovery for this general setting.

2.3. *Other Notions.* Before proceeding to our main technical parts, we provide a brief summary of some notations used throughout this paper. The inner product of two matrices are defined as $\langle X, Y\rangle = \operatorname{trace}(X'Y)$. The adjoint $\mathcal{A}^*$ of a linear operator $\mathcal{A}$ is the linear operator that satisfies $\langle \mathcal{A}(X), Y\rangle = \langle X, \mathcal{A}^*(Y)\rangle$ for all X, Y. The following norms are also used: $\|X\|_* = \sum_i \sigma_i(X)$ denotes the nuclear norm, $\|X\|_\infty = |\max_{i,j}(X_{ij})|$ the entrywise max norm, and $\|X\|_1 = \sum_{i,j}|X_{ij}|$ the entrywise ℓ_1 norm. Also, if we denote by x^i the ith row of X, then the $\ell_{p,q}$ norm of X is defined as $\|X\|_{p,q} = \left(\sum_i \|x^i\|_p^q\right)^{1/q}$. Additionally, $\|\mathcal{A}\| = \max_{X\neq 0}\frac{\|\mathcal{A}(X)\|_F}{\|X\|_F}$ denotes the operator norm.

3. Recovery via Convex Optimization. In order to enable exact/approximate recovery via convex optimization, a natural approach is to minimize a convex loss function, typically a norm $\|\cdot\|_N$, as follows.

$$\text{(CVX)} \qquad \begin{aligned} \underset{S}{\text{minimize}} \quad & \|S\|_N \\ \text{subject to} \quad & \|\mathcal{A}(S) - b\|_F \leq \sigma \end{aligned} \tag{3.1}$$

The norm $\|\cdot\|_N$ is designed to encourage solutions that have presumed low-dimensional structure; the hope is that the resulting optimum can approximate the true underlying S^*. Indeed the central issue studied in several of the

TABLE 1
Summary of Notation

$\mathcal{V}$	the subspace union
$\mathcal{CV}$	the subspace collection $\left\{(V_A, V_B) : V_A \in \mathcal{V}, V_A \subseteq V_B^{\perp}\right\}$
$\overline{\mathcal{V}}$	the augmented subspace union $\left\{V_B^{\perp} : (V_A, V_B) \in \mathcal{CV}\right\}$
$\mathcal{V}^{(i)}$	the i^{th} Minkowski sum of $\mathcal{V}$
$\mathcal{CV}^{(i)}$	the j^{th} hierarchy of $\mathcal{CV}$, i.e. $\mathcal{CV}^{(j)} = \left\{\left(\bigcup_{1\le i\le j} V_{A_i}, \bigcap_{1\le i\le j} V_{B_i}\right) : (V_{A_i}, V_{B_i}) \in \mathcal{CV}\right\}$
$\mathcal{A}$	the linear operator
$\mathcal{P}_{\mathcal{V}}$,$\mathcal{P}_{\mathcal{V}^{\perp}}$	the projection operator onto $\mathcal{V}$ and its complement $(\mathcal{P}_{\mathcal{V}^{\perp}}(X) = X - \mathcal{P}_{\mathcal{V}}(X))$
$\delta_{\mathcal{V}}$	the restricted isometry constant of $\mathcal{A}$ when restricted to subspace union $\mathcal{V}$
$\gamma(\mathcal{L})$	restricted strict convexity constant
$\theta_{\mathcal{V}_i,\mathcal{V}_j}$	the restricted orthogonality constant of $\mathcal{A}_i$ and $\mathcal{A}_j$ when restricted to $\mathcal{V}_i$ and $\mathcal{V}_j$.
$\mathcal{C}_{\mathcal{V}}(\epsilon)$	the covering number of subspace union $\mathcal{V}$ at resolution ϵ

existing works regarding special structures are sufficient conditions (usually presented as "incoherence" properties) on $\mathcal{A}$ under which this method succeeds in (exactly) recovering S^*. In the general setting, we are aimed at investigating what types of convex loss functions will enable exact/approximate solution, in the presence of incoherence on $\mathcal{A}$. An encouraging message is that if both the convex surrogates and the hierarchy of the subspace unions regarding the presumed structure exhibit some intrinsic properties, the convex optimization program will solve the problem under rather weak conditions.

3.1. *Key Definitions and Properties .* We now introduce several useful notions and properties, some of which are borrowed from [19] by Negahban *et. al.*

1) Subspace Collection and Augmented Subspace Union. We first make definitions on subspace collection and augmented subspace union, which are crucial for identifying desired convex loss function with respect to specific signal structure.

A family of subspace pairs $\mathcal{CV} = \{(V_A, V_B)\}$ is said to be a *proper subspace collection* with respect to the subspace union $\mathcal{V}$ if the following two conditions are satisfied: (a) for any subspace pair $(V_A, V_B) \in \mathcal{CV}$, we have $V_A \subseteq \mathcal{V}$ and $V_A \subseteq V_B^{\perp}$; (b) for any subspace $V_A \subseteq \mathcal{V}$, there exists a subspace pair $(V_A, V_B) \in \mathcal{CV}$.

Additionally, we define the *augmented subspace union* $\overline{\mathcal{V}}$ of $\mathcal{V}$ with respect

to $\mathcal{CV}$ as $\bigcup_{(V_A,V_B)\in\mathcal{CV}} V_B^{\perp}$. Clearly, $\overline{\mathcal{V}}$ is a subspace union that satisfies $\mathcal{V}\subseteq\overline{\mathcal{V}}$, since $V_A\subseteq V_B^{\perp}$ holds for each subspace pair $(V_A,V_B)\in\mathcal{CV}$. If the subspace union $\mathcal{V}$ is a low-dimensional manifold as well, V_B can usually be chosen such that $\overline{\mathcal{V}}$ is the union of all tangent spaces.

Given $\mathcal{V}$ and $\mathcal{CV}$, we can construct

$$\mathcal{CV}^{(j)}=\left\{\left(\bigcup_{1\le i\le j}V_{A_i},\bigcap_{1\le i\le j}V_{B_i}\right):(V_{A_i},V_{B_i})\in\mathcal{CV}\right\},$$

which captures the *hierarchy* of the subspace collection.

2) Decomposability. The following is a key property on the convex loss function that encourages the presumed structure.

Definition 2 (**[19, Decomposability]**). The norm $\|\cdot\|$ is *decomposable* associated with a proper subspace collection $\mathcal{CV}$ if for any subspace pair $(V_A,V_B)\in\mathcal{CV}$, the following holds

$$\|X+Y\|=\|X\|+\|Y\| \qquad \forall\ X\in V_A \text{ and } \forall\ Y\in V_B. \tag{3.2}$$

Also, given a proper subspace collection $\mathcal{CV}$ associated with a subspace union $\mathcal{V}$, the norm $\|\cdot\|$ is said to be *decomposable* with respect to $\mathcal{V}$ if for any $X\in\mathbb{R}^{n_1\times n_2}$, there exists a subspace pair $(V_A,V_B)\in\mathcal{CV}$ such that $\mathcal{P}_{\mathcal{V}}(X)\in V_A$ and $X-\mathcal{P}_{\mathcal{V}}(X)\in V_B$ hold.

3) Examples. For a number of low-dimensional structures, there exist convex loss functions that exhibit the above decomposability property. Some examples are listed below to illustrate the generality and utility of the above notions.

- *k-Sparse Matrices.* The subspace union $\mathcal{V}$ is given as $\mathcal{V}=\{X:|\text{support }(X)|\le k\}$, and one possible choice of $\mathcal{CV}$ is chosen such that for each subspace pair $(V_A,V_B)\in\mathcal{CV}$, there exists an index set Ω with $|\Omega|\le k$ such that $V_A=\{X\mid\mathcal{P}_{\Omega^{\perp}}(X)=0\}$ and $V_B=\{X\mid\mathcal{P}_{\Omega}(X)=0\}$. Obviously, the augmented subspace union is $\overline{\mathcal{V}}=\mathcal{V}$. In addition, since $\mathcal{P}_{\mathcal{V}}(X)$ and $\mathcal{P}_{\mathcal{V}^{\perp}}(X)$ have disjoint support, it can be verified that the ℓ_1 norm is decomposable with respect to both $\mathcal{V}$ and $\mathcal{CV}$.
- *k-Row-Sparse Matrices.* The subspace union is $\mathcal{V}=\left\{X:\|X\|_{2,0}\le k\right\}$, and one possible $\mathcal{CV}$ is such that $\forall\,(V_A,V_B)\in\mathcal{CV}$, there exists a row index set Ω_r with $|\Omega_r|\le k$ such that $V_A=\left\{X\mid\mathcal{P}_{\Omega_r^{\perp}}(X)=0\right\}$ and $V_B=\{X\mid\mathcal{P}_{\Omega_r}(X)=0\}$. The augmented subspace union obeys $\overline{\mathcal{V}}=\mathcal{V}$. Also, the $\ell_{2,1}$ norm is a candidate that obeys decomposability with respect to $\mathcal{V}$ and $\mathcal{CV}$.

- *Rank-r Matrices.* The associated subspace union can be expressed as $\mathcal{V} = \{X \mid \text{rank}\,(X) \leq r\}$. We can find a proper subspace collection $\mathcal{CV}$ such that $\forall\,(V_A, V_B) \in \mathcal{CV}$, there exist r-dimensional subspaces U and V such that

$$\begin{cases} V_A = & \{X \mid \text{col}\,(X) \subseteq U \text{ and } \text{row}\,(X) \subseteq V\} \\ V_B = & \left\{X \mid \text{col}\,(X) \subseteq U^{\perp} \text{ and } \text{row}\,(X) \subseteq V^{\perp}\right\} \end{cases}$$

  where $\text{row}(X)$ and $\text{col}(X)$ denote the row space and column space of X, respectively. The nuclear norm is decomposable with respect to $\mathcal{V}$, because the row (column) spaces of $\mathcal{P}_{\mathcal{V}}(X)$ and $\mathcal{P}_{\mathcal{V}^{\perp}}(X)$ are orthogonal. $\overline{\mathcal{V}}$ can be expressed as the union of all tangent spaces [27], i.e. $\overline{\mathcal{V}} = \left\{UX + XV^T \mid \forall r\text{-dimensional spaces } U, V\right\}$.
- *Multi-Source Case.* Suppose $S_i (1 \leq i \leq l)$ lies in subspace union $\mathcal{V}_i$ and $\|\cdot\|_{N_i}$ is decomposable with respect to both $\mathcal{V}_i$ and $\mathcal{CV}_i$. Define a general norm $\|\cdot\|_N$ as $\|S\|_N = \sum\limits_{1 \leq i \leq l} \lambda_i \|S_i\|_{N_i}$ with $\lambda_i > 0$ denoting the tradeoff parameter among structures. Then it can be verified that $\|\cdot\|_N$ is decomposable with respect to both $\mathcal{V}$ and $\mathcal{CV}$.

3.2. *Key Assumptions for Exact/Approximate Recovery.* In the subsequent analysis, we impose the following assumptions on the subspace union $\mathcal{V}$ and the associated norm $\|\cdot\|_N$, which characterize the key properties that enable exact/approximate solution.

Assumption 1. *For each subspace union $\mathcal{V}$ and its associated proper subspace collection $\mathcal{CV}$:*

*(1.a) (**Equivalence of Norms**) For any $X \in \overline{\mathcal{V}}$, we have $\|X\|_N \leq D \|X\|_F$ for some $D > 0$. For any $Y \in \mathbb{R}^{n_1 \times n_2}$, $\overline{D}_1 \|Y\|_F \leq \|Y\|_N \leq \overline{D}_2 \|Y\|_F$ for some $\overline{D}_1, \overline{D}_2 > 0$.*

*(1.b) (**Decomposability**) The norm $\|\cdot\|_N$ is decomposable with respect to both $\mathcal{V}^{(3)}$ and $\mathcal{CV}^{(3)}$.*

*(1.c) (**Successive Projectability and Decay**) For any X, $\|\mathcal{P}_{\mathcal{V}^{(3)}}(X)\|_N \geq G \|X\|_N$ holds for some constant $G > 0$. Also, $\|\mathcal{P}_{\mathcal{V}^{(3)}}(X)\|_N \geq C \|\mathcal{P}_{\mathcal{V}^{(3)}}(\mathcal{P}_{\mathcal{V}^{(3)\perp}}(X))\|_F$ holds for some constant $C > 0$.*

*(1.d) **(Closure under Projection)** For any $(V_A, V_B) \in \mathcal{CV}$, if $X \in V_B$, then $\mathcal{P}_{\mathcal{V}^{(3)}}(X) \in V_B$.*

Remark 1. Alternatively, we can find other positive integer j other than 3 such that the Assumption (1.b) – (1.d) hold with respect to $\mathcal{V}^{(j)}$ and some $\mathcal{CV}^{(j)}$.

Explanations for the above assumptions are in order, followed by a couple of specific examples.

1. Assumption (1.a) characterizes the equivalence relation between two norms, which typically exists for most well-defined norms. It should be noted that D is a parameter depending on the presumed low-dimensional structure. Also, $\overline{D}_1, \overline{D}_2 > 0$ implies that $\|\cdot\|_N$ is non-degenerate, which is necessary if we want to guarantee uniqueness of the optimal solution.
2. Assumption (1.b) is a key assumption on the convex loss function, which encourages the presumed low-dimensional structure to be the solution to the convex program.
3. Assumption (1.c) allows any non-negative matrix X to be decomposed into a sum of countable amount of matrices each following the presumed low-dimensional structure. Additionally, this assumption ensures that the projection of X will typically have higher energy than the subsequent projection of its complement. The value of G is not critical, and we only require it to be a fixed positive constant.
4. Assumption (1.d) simply states that the projection of any matrix lying in some complement subspace would still fall within that subspace, i.e. V_B is closed under projection operation $\mathcal{P}_{\mathcal{V}^{(3)}}$.

Examples. The generality of these assumptions can be better interpreted and understood through the following examples.

- *k-sparse matrices.* Consider $\mathcal{V}$ as the set of all k-sparse matrices with ℓ_1 norm as a candidate norm, where $\mathcal{V}^{(3)}$ is exactly the set of $3k$-sparse matrices. We notice that for each k-sparse $X \in \overline{\mathcal{V}} = \mathcal{V}$, $\|X\|_1 \leq \sqrt{k}\|X\|_F$ holds, i.e. $D = \sqrt{k}$. The projection $\mathcal{P}_{\mathcal{V}^{(3)}}(X)$ can be given through the $3k$ largest entries (in magnitude), which implies $\|\mathcal{P}_{\mathcal{V}^{(3)}}(X)\|_1 \geq \frac{3k}{n_1 n_2}\|X\|_1$. Setting $Y := \mathcal{P}_{\mathcal{V}^{(3)\perp}}(X)$, then $\mathcal{P}_{\mathcal{V}^{(3)}}(Y)$ just involves choosing the $3k$ largest entries of Y (i.e. $(3k+1)^{\text{th}}$ to $(6k)^{\text{th}}$ largest entries of X). We thus have $\|\mathcal{P}_{\mathcal{V}^{(3)}}(Y)\|_\infty \leq \|\mathcal{P}_{\mathcal{V}^{(3)}}(X)\|_1 / 3k$, which yields $\|\mathcal{P}_{\mathcal{V}^{(3)}}(Y)\|_F \leq \frac{1}{\sqrt{3k}}\|\mathcal{P}_{\mathcal{V}^{(3)}}(X)\|_1$. This gives us $C_i = \sqrt{3k}$. Besides, the support of $\mathcal{P}_{\mathcal{V}^{(3)}}(X)$ is always contained in the support of X, which satisfies Assumption (1.d).
- *Rank-r matrices; k-column-sparse matrices.* Similarly, for the set of all matrices of rank no larger than r, the nuclear norm gives $D = \sqrt{2r}$, $G = 3r/\min(n_1, n_2)$, and $C = \sqrt{3r}$. For the set of all k-column sparse matrices, $D = \sqrt{k}$, $G = 3k/n_2$, and $C = \sqrt{3k}$.
- *Multi-Source Case.* Suppose C_i and D_i are the parameters associated

with $\mathcal{V}_i$, and that $\sum_{1\leq i\leq l} \lambda_i \|S_i\|_{N_i}$ is used as the objective function. Adjusting the value of λ_i (e.g. $\lambda_i = C_i^{-1}$) allows us to obtain a uniform constant C and an upper bound on D as $C\max_i \frac{D_i}{C_i}$.

3.3. *Main Results.* Now, we are ready to state our main result as follows.

Theorem 1. *Let δ denote the RIP constant of the continuous operator $\mathcal{A}$ restricted to $\mathcal{V}^{(3)} \oplus \overline{\mathcal{V}}$ for notational simplicity. Under Assumption 1, if $\delta < \frac{C-D}{C+D}$ holds, then the solution $\hat{S}$ to (CVX) obeys*

$$\left\|S^* - \hat{S}\right\|_N \leq \frac{\sigma}{\frac{1-\delta}{2D} - \frac{1+\delta}{2C}} \tag{3.3}$$

Corollary 1. *Under the assumptions in Theorem 1, the solution to (CVX) is unique and exact in noise-free setting, i.e. $\sigma = 0$.*

The above results can recover known results in single-structure reconstruction problem. For example, if we take $\mathcal{V}$ as the rank-r manifold, the sufficient condition becomes $\delta < \frac{\sqrt{3}-\sqrt{2}}{\sqrt{3}+\sqrt{2}} \approx \frac{1}{10}$ with δ denoting the RIP constant for rank-$5r$ manifold, which is consistent with [14]. Theorem 1 essentially indicates that small G-RIP constants ensure approximate recovery for the general setting.

If we set $\|\mathcal{A}(X) - b\|_F^2$ as the convex penalty function, then our analytical framework implies certain restricted strong convexity (RSC) that is crucial in [19]. Formally, the RSC constant is defined as follows.

Definition 3 (**Restricted Strong Convexity**). For a convex penalty function $\mathcal{L}(X)$, the restricted strong convexity constant $\gamma_{\mathcal{V}}(\mathcal{L})$ of $\mathcal{L}$ restricted to $\mathcal{V}$ with respect to the norm $\|\cdot\|_N$ is the largest quantity such that for all matrices $X \in \mathcal{V}$:

$$\mathcal{L}(X+H) - \mathcal{L}(X) - \langle \nabla \mathcal{L}(X), H\rangle \geq \gamma(\mathcal{L}) \|H\|_N^2 \tag{3.4}$$

holds for all $H \in \{H \mid \|X\|_N \geq \|X+H\|_N\}$.

We are now ready to state our result formally as follows.

Theorem 2. *Take the penalty function as $\mathcal{L}(X) := \|\mathcal{A}(X) - b\|_F^2$. Under the assumptions made in 1, the RSC constant satisfies*

$$\gamma(\mathcal{L}) \geq \left(\frac{1-\delta}{2D} - \frac{1+\delta}{2C}\right)^2 \tag{3.5}$$

Remark 2. The parameter $\gamma(\mathcal{L})$ characterizes the deviation of the loss function from linear estimate, which is a key assumption in [19] but has not been quantified for the unified setting. For their results to be useful for recovery, we require that $\gamma(\mathcal{L}) > 0$, which is not ascertained in general setting in [19]. Our results immediately show that good G-RIP condition implies restricted strong convexity (i.e. $\delta < \frac{C-D}{C+D} \Longrightarrow \gamma(L) > 0$), which is desired for both exact and approximate recovery.

Now we would like to sketch the proofs for *noiseless* setting, which already captures all key ideas. Since the proofs of Theorem 1 and 2 are closely related, we outline their proof ideas together as follows.

Sketch of Proof of Theorem 1 and 2. Denote by $H := \hat{S} - S^*$ the difference between the solution to (CVX) and the true signal, and suppose instead that $\hat{S} \neq S^*$ (or $H \neq 0$).

1) We can observe that H belongs to the restricted set associated with the RSC constant. In the noiseless setting, another prior information is that $\mathcal{A}(H) = 0$. If we choose a convex penalty function as $\mathcal{L}(X) := \|\mathcal{A}(X) - b\|_F^2$, it can be easily verified that proving the *positiveness* of the RSC constant is equivalent to showing $\|H\|_F = 0$. Hence, our focus becomes how to bound the magnitude of H given that $\|\mathcal{A}(H)\|_F = 0$.

2) The restricted set which H lies in largely depends on the structure of the true signal S^*. This inspires us to partition H into two parts: H_0 (which lies in the same subspace as S^*) and H_c (which lies in the complement subspace). The optimality of $S^* + H$ combined with the decomposability assumption implies that $\|H_0\|_N \geq \|H_c\|_N$. That said, the "principal component" of H is approximately concentrated on the same subspace as the true signal S^*, and little mass is allocated to outside of the subspace where S^* belong.

3) In general, $\|\mathcal{A}(H)\|_F = 0$ does not necessarily imply $H = 0$ even under good RIP condition, since RIP can only ensure isometric property for a restricted set of signals. This inspires us to decompose H into a sequence of subcomponents $\{H_i, i \geq 0\}$, each exhibiting the presumed low-dimensional structure. The decomposition is well-defined due to the successive projectability assumption.

4) In addition to the fact that $H_c = \sum_{j\geq 1} H_j$, the decomposability assumption guarantees that $\|H_c\|_N = \sum_{j\geq 1} \|H_j\|_N$, which further implies that the principal component $\|H_0\|_N$ is larger than the sum of all others. RIP condition helps relate $\{\|\mathcal{A}(H_i)\|_F\}$ with $\{\|H_i\|_F\}$, which eventually establishs a connection among $\|\mathcal{A}(\sum_i H_i)\|_F$, $\{\|H_i\|_F, i \geq 1\}$ and $\|H_0\|_F$.

5) Finally, $\mathcal{A}(H) = 0$ occurs only if the principal component $\|H_0\|_F = 0$, which further leads to $H = 0$.

See Appendix A.1 and B for detailed derivation. □

4. Recovery via Generalized Projection. Convex optimization is often computationally expensive due to the large number of constraints in the high-dimensional setting. An interesting counter-intuitive property of several low-dimensional structures, however, is that even though the corresponding union of subspaces $\mathcal{V}$ is a non-convex region, we can efficiently compute the projection $\mathcal{P}_{\mathcal{V}}$ onto the $\mathcal{V}$. For example, PCA using SVD accomplishes projection onto the set of low-rank matrices, and projection onto k-sparse matrices just involves choosing the k elements of largest magnitude and setting others to 0.

To leverage this efficiency, we re-pose our recovery problem in the noise-free setting as follows:

$$\begin{array}{ll} \underset{S}{\text{minimize}} & \frac{1}{2}\left\|\mathcal{A}\left(S\right)-b\right\|_2^2 \\ \text{subject to} & S \in \mathcal{V} \end{array} \tag{4.1}$$

where $b = \mathcal{A}(S^*)$, and $\mathcal{V}$ is typically a non-convex set. Clearly the true S^* represent optima (since the objective function is 0), and G-RIP ensures the optimum will be unique. Motivated by [22, 23], our iterative projection algorithm (Algorithm 1) involves alternating gradient descent on (4.1) with projection onto low-dimensional structures.

Algorithm 1 Generalized Projections

Set $S^0 = 0$, choose step sizes η, and iteratively update until convergence

$$S^{t+1} \leftarrow \mathcal{P}_{\mathcal{V}}\left(S^t - \eta\mathcal{A}^*\left(\mathcal{A}\left(S^t\right)-b\right)\right)$$

Remark 3. The algorithm naturally allows for the simultaneous recovery of multiple structures by performing alternating projection for each structure in parallel as $\mathcal{P}_{\mathcal{V}_i}\left(S_i^t - \eta\mathcal{A}_i^*\left(\sum_j \mathcal{A}_j\left(S_j^t\right)-b\right)\right)$.

Remark 4. Recall that $\mathcal{A}^*$ represents the *adjoint* of the operator $\mathcal{A}$. The term $\mathcal{A}^*\left(\mathcal{A}\left(S\right)-b\right)$ in the update rule represents the gradient of the objective of (4.1) with respect to S.

Theorem 3. *Denote by $\delta_{(2)}$ the RIP constant of $\mathcal{A}$ restricted to $\mathcal{V}^{(2)}$, and define $\rho = \frac{2\delta_{(2)}}{1-\delta_{(2)}}$. Consider the noise-free setting. If the step size is*

set to be $\eta^t = \frac{1}{1+\delta_{(2)}}$*, and if* $\rho < 1$*, then the algorithm will converge to an output* S *that satisfies (a)* $S \in \mathcal{V}$*; (b)* $\|\mathcal{A}(S) - b\|_F^2 \leq \epsilon$ *with the amount of iterations not exceeding* $\left\lceil \frac{1}{-\log(\rho)} \log \frac{\|b\|_F^2}{\epsilon} \right\rceil$.

PROOF. The proof proceeds in the same spirit as in the specialized analysis for sparse vectors in [22]. See Appendix C for details. □

This theorem allows the following simpler claim to be derived: if $\delta_{(2)} < 1/3$, then we can observe geometric convergence rate. This theorem, while exactly recovering the known results for specific single-structure models including compressed sensing [22] and low-rank matrix recovery [23], immediately gives new results for other structures. For example, specializing it on block-sparse signals yields an alternating minimization algorithms for block-sparse structure recovery, which is expected to converge with geometric rate in the presence of incoherence.

5. G-RIP Constant for Multi-Source Case. The RIP constants of some classes of random operators restricted to specific low-dimensional structures such as low-rank matrices and sparse vectors, have been quantified in [14][6]. For multi-source setup, we notice that

$$\left\| \sum_{i=1}^{l} \mathcal{A}_i \left(S_i \right) \right\|_F^2 = \sum_{i=1}^{l} \left\| \mathcal{A}_i \left(S_i \right) \right\|_F^2 + \sum_{i \neq j} \left\langle \mathcal{A}_i \left(S_i \right), \mathcal{A}_j \left(S_j \right) \right\rangle$$

where $\mathcal{A}_i \left(\mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}^m \right)$ are linear maps. There is a series of cross-structure terms in addition to the within-structure ones, which allows us to interpret G-RIP alternatively using both within-structure RIP (RIP of $\mathcal{A}_i$ restricted to S_i) and the following cross-structure properties.

Definition 4 (**Restricted Orthogonality Property (ROP)**). For subspace unions $\mathcal{V}_1$ and $\mathcal{V}_2$, the pairwise restricted orthogonality constant $\theta_{\mathcal{V}_1,\mathcal{V}_2} \left(\mathcal{A}, \mathcal{B} \right)$ of the operators $\mathcal{A}$ and $\mathcal{B}$ restricted to $\mathcal{V}_1$ and $\mathcal{V}_2$, respectively, is the smallest quantity such that

$$\left| \left\langle \mathcal{A} \left(X \right), \mathcal{B} \left(Y \right) \right\rangle \right| \leq \theta_{\mathcal{V}_1,\mathcal{V}_2} \left(\mathcal{A}, \mathcal{B} \right) \left\| X \right\|_F \left\| Y \right\|_F \tag{5.1}$$

holds for all matrices $X \in \mathcal{V}_1$ and $Y \in \mathcal{V}_2$.

ROP measures how far apart the two distinct components of measurements can be. A small ROP constant can be interpreted as preventing the

two components from being aligned to the same direction (and hence obscured), when restricted to respective low-dimensional structures. Denote $\theta_{i,j}$ as $\theta_{\mathcal{V}_i,\mathcal{V}_j}(\mathcal{A}_i,\mathcal{A}_j)$ and δ_i as $\delta_{\mathcal{V}_i}(\mathcal{A}_i)$ for notational simplicity. It can be observed that

$$\begin{aligned}\left\|\sum_{i=1}^{l}\mathcal{A}_i(S_i)\right\|_F^2 &\le \sum_{i=1}^{l}(1+\delta_i)\|S_i\|_F^2+\sum_{i\neq j}\theta_{i,j}\|S_i\|_F^2\|S_j\|_F^2\\ &\le \sum_{i=1}^{l}(1+\delta_i+\sum_{i\neq j}\theta_{i,j})\|S_i\|_F^2\\ &\le \left(1+\max_i\left(\delta_i+\sum_{i\neq j}\theta_{i,j}\right)\right)\sum_{k=1}^{l}\|S_k\|_F^2\end{aligned}$$

Similarly, it can be verified that

$$\left\|\sum_{i=1}^{l}\mathcal{A}_i(S_i)\right\|_F^2 \ge \left(1-\max_i\left(\delta_i+\sum_{i\neq j}\theta_{i,j}\right)\right)\sum_{k=1}^{l}\|S_k\|_F^2$$

which implies that the G-RIP constant obeys $\delta \le \max_i\left(\delta_i+\sum_{i\neq j}\theta_{i,j}\right)$. If we already know the RIP constants for specific single-structure models (e.g. δ_i), identifying $\theta_{i,j}$ yields a reasonable upper bound on δ, which motivates the needs for quantifying ROP constants. This may sometimes be more convenient to derive. We show several natural instances of random operators where the pairwise ROP will hold.

5.1. *Nearly Isometric Operators and Covering Number.* A desired family of random operators is the one that is nearly isometric [14] for all matrices, which exhibits concentration behavior with reasonably light tail bounds. We demonstrate our results primarily based on this class of random operators defined below.

Definition 5 (**Nearly-Isometric Operator**). An operator $\mathcal{A}$ is *nearly-isometric* with tail parameter q if for any given X and a fixed constant $\varepsilon < 1$, we have

$$\mathbb{P}\left(\left|\|\mathcal{A}_j(X)\|_F^2-\|X\|_F^2\right|>\varepsilon\|X\|_F^2\right)\le\exp\left(-c_jq\right). \tag{5.2}$$

for some constant c_j independent of q.

Examples of nearly-isometric operators include i.i.d. Gaussian ensembles and Bernoulli ensembles. Although these operators are not uniformly isometric for all matrices, they could become "well-conditioned" when restricted

to specific low-dimensional manifold that has fairly small measure relative to the ambient high-dimensional space. If we know that the *degrees of freedom* (which we quantify via covering argument) of this low-dimensional set is small, discretization argument may help obtain a uniform control over all matrices of specific low-dimensional structure.

Before continuing, we provide the formal definition of the covering number.

Definition 6 (**Covering Number**). The covering number $\mathcal{C}_{\mathcal{V}}(\epsilon)$ of the subspace union $\mathcal{V}$ at resolution $\epsilon(\epsilon < 1)$ with respect to the augmented subspace union $\overline{\mathcal{V}}$ is defined to be the smallest number of subspaces $\overline{V}_i \subseteq \overline{\mathcal{V}}$ such that for any $X \in \mathcal{V}$, there is an i with $\left\|X - \mathcal{P}_{\overline{V}_i}(X)\right\|_F \leq \epsilon \|X\|_F$.

For example, if $\mathcal{V}$ is the set of k-sparse matrices, then $\mathcal{C}_{\mathcal{V}}(\epsilon) \leq \binom{n^2}{k} \leq \left(\frac{C_0}{\epsilon}\right)^{k \log \frac{n^2}{k}}$ for some constant C_0; if $\mathcal{V}$ is the set of all rank-r matrices, then $\mathcal{C}_{\mathcal{V}}(\epsilon) \leq (C_0'/\epsilon)^{2r(n-r)}$ for some C_0'[14]. It should be noted that $\log \mathcal{C}_{\mathcal{V}}(\epsilon)$ in some sense quantifies the order of the *degrees of freedom* of $\mathcal{V}$.

5.2. *Restricted Orthogonality Constant.* In this subsection, we shall consider two types of measurements: (a) the case with two random operators $\mathcal{A}_i$ and $\mathcal{A}_j$ independently picked; (b) the case with one random operator applied to two incoherent bases. We assume $S_i \in \mathbb{R}^{n \times n}$ for simplicity.

5.2.1. *ROP for Independent Nearly-Isometric Operators.* A somewhat counter-intuitive fact is that two independent nearly-isometric operators are insufficient to ensure a small ROP constant. For instance, take $\mathcal{V}_i$ and $\mathcal{V}_j$ to be both k-sparse manifolds, and $\mathcal{A}_i$ and $\mathcal{A}_j$ to be identity operators that clearly satisfy isometric properties. Exact splitting is obviously impossible here. The key insight is that *enough randomness* on the operators is needed, which we formalize quantitatively below.

Assumption 2. *Suppose $\mathcal{A}_i$ and $\mathcal{A}_j$ are independently picked.*
(2.a) There is a constant C_0 independent of ϵ and n such that the covering numbers of subspace unions $\mathcal{V}_i$ and $\mathcal{V}_j$ obey

$$\mathcal{C}_{\mathcal{V}_i}(\epsilon) \leq \left(\frac{C_0}{\epsilon}\right)^{r_i} \text{ and } \mathcal{C}_{\mathcal{V}_j}(\epsilon) \leq \left(\frac{C_0}{\epsilon}\right)^{r_j} \tag{5.3}$$

for some r_i and r_j depending on n.

(2.b) For any nearly-isometric operator $\mathcal{A}$ with tail parameter $p > \bar{c}_q r_j \log n$, we have $\delta_{\mathcal{V}_j}(\mathcal{A}) < \delta$ with probability at most $\exp(-c_p p)$ for some fixed $\bar{c}_q$ and c_p.

(2.c) $\mathcal{A}_j$ is nearly isometric with tail parameter m.

(2.d) For any given subspace U of dimension q_i, $\frac{m}{q_i}\mathcal{P}_U\mathcal{A}_i^*$ is nearly isometric with tail parameter q_i

(2.e) $\exists c^*, c_q^* > 0$ such that the operator norm $\|\mathcal{A}_i\| < n^{c^*}$ with probability at least $1 - \exp\left(-c_q^* q_i\right)$.

(2.f) $m > q_i > \bar{c}_q \max\{r_j, r_i\} \log n$ for some constant $\bar{c}_q$.

Some expanations are in order, followed by the formal statement of our theorem.

1. Assumption (2.a) points out that the degrees of freedom for $\mathcal{V}_i$ and $\mathcal{V}_j$ are r_i and r_j, respectively
2. Assumption (2.b) basically means that a near-isometric operator $\mathcal{A}$ will satisfy within-structure RIP for $\mathcal{V}_j$ with high probability.
3. Assumption (2.d) requires $\frac{m}{q_i}\mathcal{P}_U\mathcal{A}_i^*$ satisfies nearly-isometric property for any U of fixed dimension. Since U is arbitrarily chosen, a certain degree of randomness on $\mathcal{A}_i$ is required, which precludes the possibility of $\mathcal{A}_i$ being deterministic.
4. Assumption (2.e) basically means that $\|\mathcal{A}_i\|$ should have *bounded* operator norm, which naturally holds for most operator of interests.
5. Assumption (2.f) implies that the number of measurements is larger than the degrees of freedom.

Theorem 4. *Under Assumption 2, we have*

$$\theta_{\mathcal{V}_i,\mathcal{V}_j}(\mathcal{A}_i, \mathcal{A}_j) \leq (1+\delta)\sqrt{\frac{q_i}{m}} \tag{5.4}$$

holds with probability exceeding $1 - \exp(-c_q q_i)$ for some constant c_q.

Theorem 4 shows that if we set q_i to be larger than the degrees of freedom, and take m reasonably larger than q_i, a small ROP constant can be guaranteed with high probability. This is order-optimal in the sense that it is impossible to impose a uniform control over all matrices of the presumed low-dimensional structure if the number of measurements is smaller than the degrees of freedom.

Sketch of Proof of Theorem 4. Discretization argument is used to obtain a uniform control on the deviations. Since the covering number of each

structure is small, we can find a relatively small number of subspaces $\{V_\#\}$ to approximate the entire subspace union at a good-enough resolution. Specifically, $\forall S_i \in \mathcal{V}_i$, we will be able to find one $V_\#$ from this subspace set such that $\|S_i - \mathcal{P}_{V^\#}(S_i)\|_F$ is small. Therefore, the inner product $\langle \mathcal{A}_i(S_i), \mathcal{A}_j(S_j)\rangle$ can be split into two parts

$$\langle \mathcal{A}_i(\mathcal{P}_{V^\#}(S_i)), \mathcal{A}_j(S_j)\rangle \text{ and } \langle \mathcal{A}_i(S_i - \mathcal{P}_{V^\#}(S_i)), \mathcal{A}_j(S_j)\rangle,$$

where the second term is small since $\|S_i - \mathcal{P}_{V^\#}(S_i)\|_F$ can be arbitrarily small.

The first term can be rewritten as

$$|\langle \mathcal{A}_i(\mathcal{P}_{V^\#}(S_i)), \mathcal{A}_j(S_j)\rangle| = |\langle S_i, \mathcal{P}_{V^\#}\mathcal{A}_i^*\mathcal{A}_j(S_j)\rangle|. \tag{5.5}$$

Our assumption states that both $\mathcal{A}_j$ and $\mathcal{P}_{V^\#}\mathcal{A}_i^*$ are nearly isometric, which implies that $\mathcal{P}_{V^\#}\mathcal{A}_i^*\mathcal{A}_j$ is also nearly isometric due to the fact that $\mathcal{A}_i$ and $\mathcal{A}_j$ are independently picked. Therefore, with high probability, we can bound the inner-product reasonably well for all subspaces $\left\{V^\#\right\}$. That said, the second term can be well bounded. Combining the above facts yields the uniform control over all possible subspaces.

See Appendix D for detailed derivation. □

Example (Gaussian Ensembles). One example of the operators that satisfy the above conditions are i.i.d. Gaussian ensembles. Consider $\mathcal{V}_1$ as rank-r manifold, and $\mathcal{V}_2$ contains all k-sparse matrices. Suppose $[\mathcal{A}_1(X)]_{ij} = \left\langle A_1^{ij}, X\right\rangle$, where all entries of A_1^{ij} $(\forall i, j)$ are i.i.d. drawn from Gaussian entries $\mathcal{N}\left(0, \sqrt{\frac{1}{m}}\right)$. $\mathcal{A}_2$ is set to be *identity* operator, which is essentially nearly-isometric. Observing that $\mathcal{A}_1^*$ is a Gaussian ensemble isometric under orthogonal transformation, we can verify from that $\mathcal{P}_U\mathcal{A}_1^*$ obeys

$$\mathbb{P}\left(\frac{m}{q_i}\|\mathcal{P}_U\mathcal{A}_1^*(Y)\|_F^2 - \|Y\|_F^2 > t\|Y\|_F^2\right) \le \exp(-c_u q_i) \tag{5.6}$$

for some c_u. Hence, if $q_i \gg \max(k\log n, r(n-r)\log n)$, a small ROP constant can be obtained with high probability.

5.2.2. *ROP Constant for Sparse Representation.* Now we consider how to characterize ROP for another type of operators that may occur in sparse representation case. Here, the two operators are no longer independently chosen, but certain *incoherence* property between the two bases still allows a small ROP constant to be derived.

Consider the following two-structure setup. Suppose we want to recover signals S_1 and S_2 given measurements $b = \mathcal{A}(S) = \mathcal{A}(\Phi(S_1) + \Psi(S_2))$. Therefore, our objective is to quantify the ROP constant associated with operators $\mathcal{A}\cdot\Phi$ and $\mathcal{A}\cdot\Psi$. The relation between ROP and the *incoherence* between two bases Φ and Ψ is characterized as follows.

Theorem 5. *Denote by* $\theta_{k_1,k_2} \triangleq \theta_{k_1,k_2}(\mathcal{A}\cdot\Phi, \mathcal{A}\cdot\Psi)$ *the ROP constant associated with operators* $\mathcal{A}\cdot\Phi$ *and* $\mathcal{A}\cdot\Psi$ *when restricted to* k_1*-sparse and* k_2*-sparse manifold, respectively. Define the incoherence measure as*

$$\mu = \sup_{i,j,k,l} |\langle \Phi_{ij}, \Psi_{kl}\rangle|, \tag{5.7}$$

where $\{\Phi_{kl}\}$ *and* $\{\Psi_{kl}\}$ *are the orthonormal bases of* Φ *and* Ψ*, respectively. If* $\mathcal{A}$ *is nearly isometric with tail parameter* $m (m > c_s\mu^{-1}\log n)$ *for some constant* c_s*, then we have*

$$\theta_{k_1,k_2} \leq 2\mu\sqrt{k_1k_2} \tag{5.8}$$

with probability exceeding $1 - \exp(-\tilde{c}_A m)$ *for some positive quantity* $\tilde{c}_A$*.*

We know that $\mu = \Theta\left(\frac{1}{n}\right)$ is typical for incoherent bases, e.g. when Φ and Ψ represent bases in time domain and frequency domain, respectively. We thus require $k_1k_2 = O(n^2)$ to ensure a small ROP constant. This is consistent with the uncertainty principle derived in [28, 29]. Basically, if $k_1k_2 \geq c_1n^2$ for some constant c_1, it would be impossible to obtain a unique sparse representation, because there may exist another sparse matrix pair obeying the required condition. The key message of this theorem is that: the more incoherent two bases are, the smaller our ROP constant would be.

Sketch of Proof of Theorem 5. Instead of directly studying the properties of μ, we first look at another incoherence measure associated with $\mathcal{A}$ as follows

$$\mu_{\mathcal{A}} = \max_{i,j,k,l} |\langle \mathcal{A}(\Phi_{ij}), \mathcal{A}(\Psi_{kl})\rangle|. \tag{5.9}$$

Simply by expanding $\langle \mathcal{A}\cdot\Phi(X), \mathcal{A}\cdot\Psi(Y)\rangle$ according to the bases of Φ and Ψ and applying Cauchy-Schwartz inequality, we can bound the ROP constant θ_{k_1,k_2} with the incoherence measure $\mu_{\mathcal{A}}$, i.e. $\theta_{k_1,k_2} \leq \mu_{\mathcal{A}}\sqrt{k_1k_2}$.

Next, we need to bound $\mu_{\mathcal{A}}$ using μ. Since $\mathcal{A}$ is nearly isometric, we can obtain a uniform control of $\|\mathcal{A}(\Phi_{ij})\|_F$,$\|\mathcal{A}(\Psi_{kl})\|_F$, and $\|\mathcal{A}(\Phi_{ij} \pm \Psi_{kl})\|_F$ over all $\{i,j,k,l\}$, provided that the number of measurements satisfies $m =$

$\Omega(\log n)$. Such RIP condition allows us to establish the connection between two incoherence measure $\mu_{\mathcal{A}}$ and μ.

See detailed derivation in Appendix E. □

6. Conclusion. In this paper, we provide a unified framework for low-dimensional structure recovery from high-dimensional ambient spaces, in the presence of incoherence on the sampling operators. The main innovation in our results is in the general setting, e.g. defining the hierarchy of subspace unions as the general object that captures low-dimensional structure, and investigating the general conditions under which a candidate penalty function recovers the sought-after low-dimensional structure. Our results provide a general understanding of when incoherence recovers low-dimensional signals, which may guide the search for appropriate algorithms for new specialized structures.

Acknowledgment. Y. Chen wishes to thank Prof. Jeffrey G. Andrews for his helpful comments and consistent support.

APPENDIX A: PROOF OF THEOREMS

A.1. Proof of Theorem 1. Suppose instead that there exists H such that $\mathcal{A}(S^* + H) = \mathcal{A}(S^*)$ and

$$\|S^*\|_N \geq \|S^* + H\|_N . \tag{A.1}$$

Suppose that S^* lies in the subspace $V_A \subseteq \mathcal{V}$, and $(V_A, V_B) \in \mathcal{CV}$ is the associated subspace pair in the subspace collection. Denote by $\mathcal{P}_{V_B}$ and $\mathcal{P}_{V_B^\perp}$ the orthogonal projector onto the subspace V_B and its orthogonal complement $V_B^\perp$, respectively. Set $H_c = \mathcal{P}_{V_B}(H)$ and $H_0 = \mathcal{P}_{V_B^\perp}(H)$. Obviously, $H = \mathcal{P}_{V_B}(H) + \mathcal{P}_{V_B^\perp}(H) = H_c + H_0$. We can thus observe that

$$\begin{aligned}
\|S^* + H\|_N &= \|S^* + H_c + H_0\|_N \\
&\geq \|S^* + H_c\|_N - \|H_0\|_N && \text{(A.2)} \\
&\geq \|S^*\|_N + \|H_c\|_N - \|H_0\|_N && \text{(A.3)} \\
&\geq \|S^* + H\|_N + \|H_c\|_N - \|H_0\|_N , && \text{(A.4)}
\end{aligned}$$

where (A.2) follows from the triangle inequality of norms, (A.3) follows from the decomposability assumption, and (A.4) is a consequence of (A.1). Hence, manipulation yields

$$\|H_0\|_N \geq \|H_c\|_N . \tag{A.5}$$

Next, we will successively project H_c onto subspace union $\mathcal{V}^{(3)}$ so as to decompose H_c into a couple of low-dimensional components. Set $\pi_0 := H_c$, then such sequential projection can be formally expressed as

$$\begin{cases} H_t &= \mathcal{P}_{\mathcal{V}^{(3)}}(\pi_{t-1}) \\ \pi_t &= \pi_{t-1} - H_t = \mathcal{P}_{\mathcal{V}^{(3)\perp}}(\pi_{t-1}) \end{cases} \quad (t \geq 1) \tag{A.6}$$

Hence, H_c is partitioned into a sum of matrices $\{H_1, H_2, \cdots\}$. In order to ensure that H_c can be well represented by a sum of these low-dimensional components, we need to prove that $\sum_{i\geq 1} H_i$ converges to H_c.

Assumption (1.c) guarantees that

$$\|H_t\|_N = \|\mathcal{P}_{\mathcal{V}^{(3)}}(\pi_{t-1})\|_N \geq G \|\pi_{t-1}\|_N \tag{A.7}$$

for some constant G. In addition, we know from Assumption (b) that $\|\cdot\|_N$ is decomposable with respect to $\mathcal{V}^{(3)}$ and $\mathcal{C}\mathcal{V}^{(3)}$. This indicates that there exists a subspace pair $\left(\tilde{V}_A, \tilde{V}_B\right) \in \mathcal{C}\mathcal{V}^{(3)}$ such that

$$\begin{cases} H_t = \mathcal{P}_{\mathcal{V}^{(3)}}(\pi_{t-1}) \in \tilde{V}_A \\ \pi_t = \pi_{t-1} - \mathcal{P}_{\mathcal{V}^{(3)}}(\pi_{t-1}) \in \tilde{V}_B \end{cases} . \tag{A.8}$$

By definition of decomposability, we can obtain

$$\|\pi_t\|_N = \|\pi_{t-1}\|_N - \|H_t\|_N \leq (1-G)\|\pi_{t-1}\|_N . \tag{A.9}$$

This indicates that the magnitude of π_t is decaying rapidly with geometric rate. Therefore, for any constant $\epsilon > 0$, there exists an integer $t_\epsilon = \Theta\left(\log \frac{1}{\epsilon}\right)$ such that $\|\pi_t\|_N \leq \epsilon \|\pi_0\|_N$ for all $t \geq t_\epsilon$. We know that

$$\begin{aligned} \|\pi_t\|_N &= \|\pi_{t-1}\|_N - \|H_t\|_N = \|\pi_{t-2}\|_N - \|H_{t-1}\|_N - \|H_t\|_N \\ &= \cdots \quad = \|\pi_0\|_N - \sum_{i=1}^{t} \|H_i\|_N \end{aligned}$$

for all $t \geq t_\epsilon$. Since $\pi_0 = H_c$, we can derive

$$\sum_{i=1}^{t} \|H_i\|_N = \|H_c\|_N - \|\pi_t\|_N \geq (1-\varepsilon)\|H_c\|_N \tag{A.10}$$

Also, $\sum_{i=1}^{t} \|H_i\|_N = \|\pi_0\|_N - \|\pi_t\|_N \leq \|\pi_0\|_N$. In addition, we can obtain

$$\begin{aligned} \left\| H_c - \sum_{i=1}^{t} H_i \right\|_F &\leq \frac{1}{\overline{D}_1} \left\| H_c - \sum_{i=1}^{t} H_i \right\|_N = \frac{1}{\overline{D}_1} \|\pi_t\|_N \\ &\leq \epsilon \frac{1}{\overline{D}_1} \|H_c\|_N \leq \epsilon \left(\frac{\overline{D}_2}{\overline{D}_1} \|H_c\|_F \right) \end{aligned} \tag{A.11}$$

for arbitrary positive ϵ. Combining the above facts yields

$$\begin{cases} \lim_{t\to\infty} \sum_{i=1}^{t} \|H_i\|_N & = \|H_c\|_N \\ \lim_{t\to\infty} \sum_{i=1}^{t} H_i & = H_c \end{cases}, \tag{A.12}$$

which indicates that the above decomposition is well-defined.

We also observe from Assumption (1.c) that

$$\begin{aligned} \|H_{j+1}\|_F &= \|\mathcal{P}_{\mathcal{V}^{(3)}}(\pi_j)\|_F = \|\mathcal{P}_{\mathcal{V}^{(3)}}(\mathcal{P}_{\mathcal{V}^{(3)\perp}}(\pi_{j-1}))\|_F \\ &\le \frac{1}{C} \|\mathcal{P}_{\mathcal{V}^{(3)}}(\pi_{j-1})\|_N \\ &= \frac{1}{C} \|H_j\|_N . \end{aligned}$$

Combining this with the decomposition of H_c yields

$$\lim_{t\to\infty} \sum_{j=2}^{t} \|H_j\|_F \le \lim_{t\to\infty} \frac{1}{C} \sum_{j=1}^{t} \|H_j\|_N = \frac{1}{C} \|H_c\|_N .$$

Recall from (A.5) that $\|H_c\|_N \le \|H_0\|_N$, then we have

$$C \lim_{t\to\infty} \sum_{j=2}^{t} \|H_j\|_F \le \|H_c\|_N \le \|H_0\|_N \le D \|H_0\|_F ,$$

where the last inequality arises from Assumption (1.a) that characterizes the equivalence relation of norms.

Since $H_c = \mathcal{P}_{V_B}(H) \in V_B$ and $H_0 \in V_B^{\perp}$, we know from Assumption (1.d) that $H_1 = \mathcal{P}_{\mathcal{V}^{(3)}}(H_c) \in V_B$, which further implies that

$$\langle H_0, H_1 \rangle = 0, \tag{A.13}$$

and hence we can derive

$$\|H_0 + H_1\|_F = \sqrt{\|H_0\|_F^2 + \|H_1\|_F^2} \ge \|H_0\|_F . \tag{A.14}$$

Since $H_0 \in V_B^{\perp}$ lies in the augmented subspace union $\overline{\mathcal{V}}$ and $H_1 = \mathcal{P}_{\mathcal{V}^{(3)}}(H_c) \in \mathcal{V}^{(3)}$, it can be seen that $H_0 + H_1$ lies in the subspace union $\mathcal{V}^{(3)} \oplus \overline{\mathcal{V}}$. Since

$\mathcal{A}$ is continuous, we have

$$\begin{aligned}
\|\mathcal{A}(H)\|_F &= \left\| \mathcal{A}\left(H_0 + \lim_{t\to\infty} \sum_{j=1}^{t} H_j \right) \right\|_F \\
&\geq \|\mathcal{A}(H_0 + H_1)\|_F - \lim_{t\to\infty} \sum_{j=2}^{t} \|\mathcal{A}(H_j)\|_F \\
&\geq (1-\delta)\|H_0 + H_1)\|_F - (1+\delta) \lim_{t\to\infty} \sum_{j=2}^{t} \|H_j\|_F \\
&\geq (1-\delta)\|H_0 + H_1)\|_F - (1+\delta)\frac{D}{C}\|H_0\|_F \\
&\geq \left(1 - \delta - (1+\delta)\frac{D}{C}\right)\|H_0\|_F \qquad \text{(A.15)}
\end{aligned}$$

where the last inequality (A.15) is a consequence of (A.14).

Since $H_0 \in V_B^{\perp} \subseteq \overline{\mathcal{V}}$, we can see from Assumption (1.a) that

$$\begin{aligned}
\|H_0\|_F &\geq \frac{1}{D}\|H_0\|_N \geq \frac{1}{2D}\left(\|H_0\|_N + \|H_c\|_N\right) \\
&\geq \frac{1}{2D}\|H\|_N
\end{aligned}$$

From the convex constraint we know that $\|\mathcal{A}(S^* + H) - b\|_F = \|\mathcal{A}(H)\|_F \leq \sigma$, therefore we have

$$\begin{aligned}
\sigma \geq \|\mathcal{A}(H)\|_F &\geq \left(1 - \delta - (1+\delta)\frac{D}{C}\right)\|H_0\|_F \\
&\geq \left(\frac{1-\delta}{2D} - \frac{1+\delta}{2C}\right)\|H\|_N . \qquad \text{(A.16)}
\end{aligned}$$

If $\frac{1-\delta}{1+\delta} > \frac{D}{C}$ holds, then we can derive

$$\|H\|_N \leq \frac{\sigma}{\frac{1-\delta}{2D} - \frac{1+\delta}{2C}} \qquad \text{(A.17)}$$

which completes the proof.

A.2. Proof of Theorem 2 . Let us take the loss function as $\mathcal{L}(S) = \|b - \mathcal{A}(S)\|_F^2$. Then we have

$$\begin{aligned}
&\mathcal{L}(S^* + H) - \mathcal{L}(S^*) - \langle H, \nabla\mathcal{L}(S^*)\rangle \\
=\ &\|b - \mathcal{A}(S^* + H)\|_F^2 - \|b - \mathcal{A}(S^*)\|_F^2 \\
&\quad - \langle H, 2\mathcal{A}^*(\mathcal{A}(S^*) - b)\rangle \\
=\ &\|\mathcal{A}(H)\|_F^2
\end{aligned}$$

We can immediately see from (A.16) that

$$\|\mathcal{A}(H)\|_F^2 \geq \left(\frac{1-\delta}{2D} - \frac{1+\delta}{2C}\right)^2 \|H\|_N^2 \tag{A.18}$$

which implies

$$\gamma(L) \geq \left(\frac{1-\delta}{2D} - \frac{1+\delta}{2C}\right)^2 \tag{A.19}$$

with error norm $d(\cdot)$ being $\|\cdot\|_N$. If the assumption in the theorem is satisfied, we can obtain $\gamma(L) > 0$.

A.3. Proof of Theorem 3. The proof idea for our generalized algorithm is exactly the same as the specialized analysis in [22], which we provide here for the completeness of the paper. We define $\hat{S}^t := S^t - S^*$ to be the estimation error in round t. Hence, $\psi(S^t) = \frac{1}{2}\|\mathcal{A}(S^t) - b\|_F^2 = \frac{1}{2}\left\|\mathcal{A}(\hat{S}^t)\right\|_F^2$.

Since $S^{t+1} = \arg\max_Y \left\{Y - S^t + \eta^t \mathcal{A}^*\mathcal{A}\left(\hat{S}^t\right) : Y \in \mathcal{V}\right\}$ and $S^* \in \mathcal{V}$, we have

$$\begin{aligned}
&\left\|\hat{S}^{t+1} - \hat{S}^t + \frac{1}{1+\delta_{(2)}}\mathcal{A}^*\mathcal{A}\left(\hat{S}^t\right)\right\|_F \\
&= \left\|S^{t+1} - S^t + \eta^t \mathcal{A}^*\mathcal{A}\left(\hat{S}^t\right)\right\|_F \\
&\leq \left\|S^* - S^t + \eta^t \mathcal{A}^*\mathcal{A}\left(\hat{S}^t\right)\right\|_F \\
&= \left\|-S^t + \eta^t \mathcal{A}^*\mathcal{A}\left(\hat{S}^t\right)\right\|_F
\end{aligned} \tag{A.20}$$

Now, the approximation error can be rearranged as follows

$$\begin{aligned}
&\psi\left(S^{t+1}\right) - \psi\left(S^t\right) \\
&= \frac{1}{2}\left\|\mathcal{A}\left(\hat{S}^{t+1}\right)\right\|_F^2 - \frac{1}{2}\left\|\mathcal{A}\left(\hat{S}^t\right)\right\|_F^2 \\
&= \frac{1}{2}\left\|\mathcal{A}\left(\hat{S}^{t+1} - \hat{S}^t\right)\right\|_F^2 + \left\langle \hat{S}^{t+1}, \mathcal{A}^*\mathcal{A}\left(\hat{S}^t\right)\right\rangle - \left\|\mathcal{A}\left(\hat{S}^t\right)\right\|_F^2 \\
&\leq \frac{(1+\delta_{(2)})}{2}\left\|\hat{S}^{t+1} - \hat{S}^t\right\|_F^2 + \left\langle \hat{S}^{t+1} - \hat{S}^t, \mathcal{A}^*\mathcal{A}\left(\hat{S}^t\right)\right\rangle \\
&\quad + \left\langle \hat{S}^t, \mathcal{A}^*\mathcal{A}\left(\hat{S}^t\right)\right\rangle - \left\|\mathcal{A}\left(\hat{S}^t\right)\right\|_F^2
\end{aligned}$$

By taking advantage of (C.1) and the fact $\left\langle \hat{S}^t, \mathcal{A}^*\mathcal{A}\left(\hat{S}^t\right)\right\rangle = \left\|\mathcal{A}\left(\hat{S}^t\right)\right\|_F^2$, we can obtain

$$\begin{aligned}
&\psi\left(S^{t+1}\right) - \psi\left(S^t\right) \\
\leq\ & \frac{(1+\delta_{(2)})}{2}\left\|-\hat{S}^t + \eta^t\mathcal{A}^*\mathcal{A}\left(\hat{S}^t\right)\right\|_F^2 - \frac{\left(\eta^t\right)^2}{2}\left\|\mathcal{A}\left(\hat{S}^t\right)\right\|_F^2 \\
=\ & \frac{(1+\delta_{(2)})}{2}\left\|\hat{S}^t\right\|_F^2 - \left\|\mathcal{A}\left(\hat{S}^t\right)\right\|_F^2 \\
\leq\ & \left(\frac{(1+\delta_{(2)})}{2(1-\delta_{(2)})} - 1\right)\left\|\mathcal{A}\left(\hat{S}^t\right)\right\|_F^2
\end{aligned}$$

Since $\left\|\mathcal{A}\left(\hat{S}^t\right)\right\|_F^2 = \psi(S^t)$, simple manipulation yields

$$\psi\left(S^{t+1}\right) \leq \frac{2\delta_{(2)}}{1-\delta_{(2)}}\psi\left(S^t\right) = \rho\psi\left(S^t\right) \tag{A.21}$$

Setting the initial guess as 0 will result in an upper bound on the number of iterations.

A.4. Proof of Theorem 4. For any matrix $S_i \in \mathcal{V}_i$, we can always find one subspace $V_\#$ from $\mathcal{C}_i(\epsilon)$ properly chosen candidates such that $\left\|S_i - \mathcal{P}_{V_\#}(S_i)\right\|_F < \epsilon\left\|S_i\right\|_F$. Also, for any subspace $V_\#$, we have

$$\begin{aligned}
&\left|\left\langle \mathcal{A}_i\left(\mathcal{P}_{V_\#}S_i\right), \mathcal{A}_j\left(S_j\right)\right\rangle\right| \\
=\ & \left|\left\langle \mathcal{P}_{V_\#}S_i, \mathcal{A}_i^*\mathcal{A}_j\left(S_j\right)\right\rangle\right| \\
\leq\ & \left\|\mathcal{P}_{V_\#}S_i\right\|_F\left\|\mathcal{P}_{V_\#}\mathcal{A}_i^*\mathcal{A}_j\left(S_j\right)\right\|_F
\end{aligned}$$

Let us expand each subspace $V_\#$ to some q_i-dimensional subspace $\overline{V_\#}$ such that $V_\# \subseteq \overline{V_\#}$. Our near isometry assumption implies that

$$\begin{aligned}
&\mathbb{P}\left(\left\|\mathcal{P}_{\overline{V_\#}}\mathcal{A}_i^*\mathcal{A}_j\left(S_j\right)\right\|_F^2 \leq (1+\varepsilon)^2\frac{q_i}{m}\left\|S_j\right\|_F^2\right) \\
\geq\ & \mathbb{P}\left(\left\|\mathcal{A}_j\left(S_j\right)\right\|_F^2 \leq (1+\varepsilon)\left\|S_j\right\|_F^2\right) \times \\
& \mathbb{P}\left\{\left\|\mathcal{P}_{\overline{V_\#}}\mathcal{A}_i^*\mathcal{A}_j\left(S_j\right)\right\|_F^2 \leq (1+\varepsilon)\frac{q_i}{m}\left\|\mathcal{A}_j\left(S_j\right)\right\|_F^2\right| \\
& \qquad \left\|\mathcal{A}_j\left(S_j\right)\right\|_F^2 \leq (1+\varepsilon)\left\|S_j\right\|_F^2\Big\} \\
\geq\ & \left(1-\exp\left(-c_jm\right)\right)\left(1-\exp\left(-c_iq_i\right)\right) \\
\geq\ & 1-\exp\left(-c_{ij}q_i\right)
\end{aligned}$$

for some constant c_{ij}. This implies that $\sqrt{\frac{m}{q_i}}\mathcal{P}_{\overline{V_\#}}\mathcal{A}_i^*\mathcal{A}_j$ is nearly isometric with tail parameter q_i. We can thus see that if $q_i \geq \bar{c}_q r_j \log n$, then for all matrices $S_j \in \mathcal{V}_j$,

$$\left\|\mathcal{P}_{\overline{V_\#}}\mathcal{A}_i^*\mathcal{A}_j\left(S_j\right)\right\|_F \leq (1+\delta)\sqrt{\frac{q_i}{m}}\left\|S_j\right\|_F, \tag{A.22}$$

with probability exceeding $1-\exp\left(-c_p q_i\right)$. Take $\epsilon < \sqrt{\frac{\delta^2 q_i}{n^{c^*} m}}$ and $q_i > \tilde{c}_q r_i \log n$ for some constant $\tilde{c}_q$. The union bound implies that (D.1) holds for all $V_\#$s with probability exceeding $1-\exp\left(-c_q q_i\right)$.

We can thus conclude that with high probability, the operators $\mathcal{A}_i$ and $\mathcal{A}_j$ satisfy: for any $S_i \in \mathcal{V}_i$ in subspace $V \subseteq \mathcal{V}_i$ and any $S_j \in \mathcal{V}_j$,

$$\begin{aligned}
& \left|\left\langle \mathcal{A}_i\left(S_i\right), \mathcal{A}_j\left(S_j\right)\right\rangle\right| \\
\leq & \left|\left\langle \mathcal{A}_i\left(S_i - \mathcal{P}_{V_\#} S_i\right), \mathcal{A}_j\left(S_j\right)\right\rangle\right| + \left|\left\langle \mathcal{P}_{V_\#} S_i, \mathcal{P}_{V_\#}\mathcal{A}_i^*\mathcal{A}_j\left(S_j\right)\right\rangle\right| \\
\leq & \epsilon\left\|\mathcal{A}_i\right\|\left\|S_i\right\|_F\left\|\mathcal{A}_j\left(S_j\right)\right\|_F + \left\|S_i\right\|_F\left\|\mathcal{P}_{\overline{V_\#}}\mathcal{A}_i^*\mathcal{A}_j\left(S_j\right)\right\|_F \\
\leq & \epsilon(1+\delta)\sqrt{n^{c^*}}\left\|S_i\right\|_F\left\|S_j\right\|_F + (1+\delta)\sqrt{\frac{q_i}{m}}\left\|S_i\right\|_F\left\|S_j\right\|_F \\
\leq & (1+\delta)^2\sqrt{\frac{q_i}{m}}\left\|S_i\right\|_F\left\|S_j\right\|_F
\end{aligned}$$

That says, with probability exceeding $1-\exp\left(-c_q q_i\right)$, we can obtain $\theta_{i,j} \leq (1+\delta)^2\sqrt{\frac{q_i}{m}}$.

A.5. Proof of Theorem 5. For any k_1-sparse matrix X and k_2-sparse matrix Y, they can be decomposed as

$$\begin{cases} \Phi(X) & = \sum_{(i,j)\in \mathrm{supp}(X)} X_{ij}\Phi_{ij} \\ \Psi(Y) & = \sum_{(k,l)\in \mathrm{supp}(Y)} Y_{kl}\Psi_{kl} \end{cases}, \tag{A.23}$$

where $\Phi_{ij} = \mathbf{e}_i \mathbf{e}_j^T$ and $\{\Psi_{kl}\}$ are the orthonormal bases of Ψ. Therefore,

$$\begin{aligned}
&|\langle \mathcal{A} \cdot \Phi(X), \mathcal{A} \cdot \Psi(Y) \rangle| \\
=& \left| \left\langle \sum_{(i,j)\in \mathrm{supp}(X)} \mathcal{A}(X_{ij}\Phi_{ij}), \sum_{(k,l)\in \mathrm{supp}(Y)} \mathcal{A}(Y_{kl}\Psi_{kl}) \right\rangle \right| \\
=& \left| \sum_{(i,j)\in \mathrm{supp}(X),(k,l)\in \mathrm{supp}(Y)} X_{ij} Y_{kl} \langle \mathcal{A}(\Phi_{ij}), \mathcal{A}(\Psi_{kl}) \rangle \right| \\
\leq& \mu_{\mathcal{A}} \sum_{(i,j)\in \mathrm{supp}(X),(k,l)\in \mathrm{supp}(Y)} |X_{ij} \cdot Y_{kl}| \\
\leq& \mu_{\mathcal{A}} \left(\sum_{(i,j)\in \mathrm{supp}(X)} |X_{ij}| \right) \cdot \left(\sum_{(k,l)\in \mathrm{supp}(X)} |Y_{kl}| \right) \\
\leq& \mu_{\mathcal{A}} \sqrt{k_1 k_2} \|X\|_F \|Y\|_F
\end{aligned}$$

where $\mu_{\mathcal{A}} = \max_{i,j,k,l} |\langle \mathcal{A}(\Phi_{ij}), \mathcal{A}(\Psi_{kl}) \rangle|$. Additionally, it can be seen that if $m > \bar{c} \log n$ for some constant $\bar{c}$, union bound implies that

$$\begin{cases}
\forall (i,j): \ (1-\delta)\|\Phi_{ij}\|_F \leq \|\mathcal{A}(\Phi_{ij})\|_F \leq (1+\delta)\|\Phi_{ij}\|_F \\
\forall (k,l): \ (1-\delta)\|\Psi_{kl}\|_F \leq \|\mathcal{A}(\Psi_{kl})\|_F \leq (1+\delta)\|\Psi_{kl}\|_F \\
\forall (i,j,k,l): \ (1-\delta)\|\Phi_{ij} \pm \Psi_{kl}\|_F \leq \|\mathcal{A}(\Phi_{ij} \pm \Psi_{kl})\|_F \\
\qquad\qquad\qquad\qquad \leq (1+\delta)\|\Phi_{ij} \pm \Psi_{kl}\|_F
\end{cases}$$

holds with probability at least $1 - 3n^4 \exp(-c_A m) \geq 1 - \exp(-\tilde{c}_A m)$ for some constant $\tilde{c}_A$. This implies that

$$\begin{aligned}
&\langle \mathcal{A}(\Phi_{ij}), \mathcal{A}(\Psi_{kl}) \rangle \\
=& \frac{\|\mathcal{A}(\Phi_{ij} + \Psi_{kl})\|_F^2 - \|\mathcal{A}(\Phi_{ij} - \Psi_{kl})\|_F^2}{4} \\
\leq& \frac{(1+\delta)\|\Phi_{ij} + \Psi_{kl}\|_F^2 - (1-\delta)\|\Phi_{ij} - \Psi_{kl}\|_F^2}{4} \\
=& \frac{2\delta\left(\|\Phi_{ij}\|_F^2 + \|\Psi_{kl}\|_F^2\right) + 4\langle \Phi_{ij}, \Psi_{kl} \rangle}{4} \\
\leq& \delta + \mu
\end{aligned}$$

Similarly

$$\begin{aligned} & \langle \mathcal{A}(\Phi_{ij}), \mathcal{A}(\Psi_{kl}) \rangle \\ \geq & \frac{(1-\delta)\|\Phi_{ij}+\Psi_{kl}\|_F^2 - (1+\delta)\|\Phi_{ij}-\Psi_{kl}\|_F^2}{4} \\ \geq & -\delta-\mu \end{aligned}$$

We can thus see that

$$\sup_{i,j,k,l} |\langle \mathcal{A}(\Phi_{ij}), \mathcal{A}(\Psi_{kl}) \rangle| \leq \delta + \mu, \tag{A.24}$$

which immediately yields

$$\theta_{k_1,k_2} \leq (\delta+\mu)\sqrt{k_1 k_2}. \tag{A.25}$$

The near isometric property of $\mathcal{A}$ implies that if $m > c_s \mu^{-1} \log n$ for some constant c_s, then $\delta < \mu$.

REFERENCES

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[2] M. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using constrained quadratic programming (Lasso)," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2183 –2202, May 2009.

[3] S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[4] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, Nov. 2006.

[5] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[6] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[7] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289 –1306, April 2006.

[8] E. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406 –5425, December 2006.

[9] E. J. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis," *submitted to Journal of ACM*, 2009. [Online]. Available: http://arxiv.org/abs/0912.3599

[10] H. Xu, C. Caramanis, and S. Mannor, "Principal component analysis with contaminated data: The high dimensional case," *The 23rd Annual Conference on Learning Theory (COLT)*, June 2010.

[11] E. J. Candes and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, April 2009.

[12] E. Candes and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053 –2080, May 2010.

[13] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980 –2998, June 2010.

[14] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[15] E. Candes and Y. Plan, "Accurate low-rank matrix recovery from a small number of linear measurements," *47th Annual Allerton Conference on Communication, Control, and Computing*, pp. 1223 –1230, September 2009.

[16] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[17] G. Obozinski, M. J. Wainwright, and M. I. Jordan, "Support union recovery in high-dimensional multivariate regression," *to appear in Annals of Statistics*, 2010.

[18] S. Negahban and M. J. Wainwright., "Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞-regularization." *UC Berkeley Technical Report 774*, May 2009. [Online]. Available: http://arxiv.org/abs/0905.0642

[19] S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," *NIPS 2009*, December 2009.

[20] Y. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302 –5316, Nov. 2009.

[21] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265 – 274, 2009.

[22] R. Garg and R. Khandekar, "Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property," *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 337–344, 2009.

[23] R. Meka, P. Jain, and I. S. Dhillon, "Guaranteed rank minimization via singular value projection," *preprint*, 2009. [Online]. Available: http://arxiv.org/abs/0909.5457

[24] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655 –4666, Dec. 2007.

[25] Y. Bresler and K. Lee, "Efficient and guaranteed rank minimization by atomic decomposition," *IEEE International Symposium on Information Theory*, pp. 314 –318, June 2009.

[26] T. Blumensath and M. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1872 –1882, April 2009.

[27] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky, "Rank-sparsity incoherence for matrix decomposition," *under revision, SIAM Journal on Optimization*, December 2009. [Online]. Available: http://arxiv.org/abs/0906.2220

[28] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2845 –2862, November 2001.

[29] M. Elad and A. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Transactions on Information Theory*, vol. 48, no. 9, pp. 2558 – 2567, Sep 2002.

Department of Electrical Engineering
Stanford University
Stanford, California 94305
U.S.A.
E-mail: yxchen@stanford.edu

Department of Electrical Engineering
The University of Texas at Austin
Austin, Texas 78712
U.S.A.
E-mail: sanghavi@mail.utexas.edu