Analysis and Applications, Vol. 18, No. 1 (2020) 25–48

DOI: 10.1142/S0219530519410094





# Jointly low-rank and bisparse recovery: Questions and partial answers

Simon Foucart*
*Texas A&M University, USA*
*foucart@tamu.edu*

Rémi Gribonval
*Univ Rennes, Inria, CNRS, IRISA, France*
*Univ Lyon, Inria, CNRS, ENS de Lyon*
*UCB Lyon 1, LIP UMR 5668, France*

Laurent Jacques
*ICTEAM/INMA, UCLouvain, Belgium*

Holger Rauhut
*RWTH Aachen University*
*Chair for Mathematics of Information Processing*
*Pontdriesch 10, 52056 Aachen, Germany*

Received 26 July 2019
Accepted 29 October 2019
Published 27 November 2019

We investigate the problem of recovering jointly $r$-rank and $s$-bisparse matrices from as few linear measurements as possible, considering arbitrary measurements as well as rank-one measurements. In both cases, we show that $m \asymp rs \ln(en/s)$ measurements make the recovery possible in theory, meaning via a nonpractical algorithm. In case of arbitrary measurements, we investigate the possibility of achieving practical recovery via an iterative-hard-thresholding algorithm when $m \asymp rs^\gamma \ln(en/s)$ for some exponent $\gamma > 0$. We show that this is feasible for $\gamma = 2$, and that the proposed analysis cannot cover the case $\gamma \leq 1$. The precise value of the optimal exponent $\gamma \in [1, 2]$ is the object of a question, raised but unresolved in this paper, about head projections for the jointly low-rank and bisparse structure. Some related questions are partially answered in passing. For rank-one measurements, we suggest on arcane grounds an iterative-hard-thresholding algorithm modified to exploit the nonstandard restricted isometry property obeyed by this type of measurements.

*Keywords*: Compressive sensing; simultaneity of structures; rank-one measurements; sample complexity; restricted isometry properties; iterative thresholding algorithms; head and tail projections.

Mathematics Subject Classification 2010: 15A29, 15B99, 60B20, 65F10

*Corresponding author.

## 1. Introduction

This whole paper is concerned with the inquiry below.

**Main Question.** *What is the minimal number of linear measurements needed to recover jointly $r$-rank and $s$-bisparse symmetric $n \times n$ matrices via an efficient algorithm*?

This minimal number of measurements will be called sample complexity. We will show that it is of the order $rs\ln(en/s)$. Nevertheless, we do not consider the question fully resolved because of the lack of efficient algorithms for arbitrary measurements and of the limitation of an efficient algorithm to factorized measurements, and thus to the only applications that could support such a structured sensing. Settling the question by providing an efficient algorithm applicable to any type of measurements is therefore still open. Before diving into our investigations, let us start by clarifying a few points.

- What are 'jointly $r$-rank and $s$-bisparse symmetric $n \times n$ matrices'?

  In this paper, we consider exclusively matrices $\mathbf{X} \in \mathbb{R}^{n\times n}$ that are symmetric, i.e. $\mathbf{X}^\top = \mathbf{X}$. The set of $r$-rank (symmetric) matrices will be denoted as

$$\Sigma^{[r]} := \{\mathbf{X} \in \mathbb{R}^{n\times n} : \mathbf{X}^\top = \mathbf{X},\ \mathrm{rank}(\mathbf{X}) \le r\} \tag{1}$$

  and the set of $s$-bisparse (symmetric) matrices will be denoted as

$$\Sigma_{(s)} := \{\mathbf{X} \in \mathbb{R}^{n\times n} : \mathbf{X}^\top = \mathbf{X}, \mathbf{X}_{\overline{S\times S}} = \mathbf{0} \text{ for some } S \subseteq [\![1:n]\!] \text{ with } |S| = s\}, \tag{2}$$

  where $\mathbf{M}_\Omega = \mathbf{0}$ for $\mathbf{M} \in \mathbb{R}^{n\times n}$ and $\Omega \subseteq [\![1:n]\!] \times [\![1:n]\!]$ means that all entries of $\mathbf{M}$ indexed by $\Omega$ are zeros, and $\overline{\Omega}$ stands for the complement of $\Omega$.

  Hence, the jointly $r$-rank and $s$-bisparse (symmetric) matrices we are interested in are elements of

$$\Sigma_{(s)}^{[r]} := \Sigma^{[r]} \cap \Sigma_{(s)}. \tag{3}$$

  We will often use the fact that $\Sigma_{(s)}^{[r]} + \Sigma_{(s)}^{[r]} \subseteq \Sigma_{(2s)}^{[2r]}$.

  Note that, as described below, $\Sigma_{(s)}^{[1]}$ is for instance the set associated with the lifting of sparse signals to rank-one matrices when one is interested in their recovery from phaseless (complex) measurements [16], while for $r > 1$, any matrix of $\Sigma_{(s)}^{[r]}$ describes a quadratic function of both few variables and few quadratic terms whose sampling and recovery — an important problem in, e.g., approximation theory and high-dimensional statistics — are related to the Main Question [8, 6].
- What are the 'linear measurements' considered?

  They can be of the arbitrary type

$$y_i = \langle \mathbf{X}, \mathbf{A}_i \rangle_F = \mathrm{tr}(\mathbf{A}_i^\top \mathbf{X}), \quad i \in [\![1:m]\!], \tag{4}$$

or of the specific (rank-one) type

$$y_i = \langle \mathbf{X}\mathbf{a}_i, \mathbf{a}_i \rangle = \mathrm{tr}(\mathbf{a}_i \mathbf{a}_i^\top \mathbf{X}), \quad i \in [\![1 : m]\!]. \tag{5}$$

Generically, we write $\mathbf{y} = \boldsymbol{\mathcal{A}}(\mathbf{X})$, where $\boldsymbol{\mathcal{A}} : \mathbb{R}^{n \times n} \to \mathbb{R}^m$ is a linear map.

- What is meant by 'recover'?

  More than just finding a map $\Delta : \mathbb{R}^m \to \mathbb{R}^{n \times n}$ such that $\Delta(\boldsymbol{\mathcal{A}}(\mathbf{X})) = \mathbf{X}$ for all $\mathbf{X} \in \Sigma_{(s)}^{[r]}$. Indeed, we require the recovery procedure to be stable and robust, in the sense that we want

$$\|\mathbf{X} - \Delta(\boldsymbol{\mathcal{A}}(\mathbf{X}) + \mathbf{e})\| \le C \min_{\mathbf{Z} \in \Sigma_{(s)}^{[r]}} \|\mathbf{X} - \mathbf{Z}\| + D\|\mathbf{e}\| \tag{6}$$

to hold for all $\mathbf{X} \in \mathbb{R}^{n \times n}$ and all $\mathbf{e} \in \mathbb{R}^m$. We give ourselves some freedom on the choice of the three norms appearing in (6). We also require the recovery procedure to be implementable by a practical algorithm, that is, an efficient algorithm whose run-time is at most polynomial in $n$ and $m$ (ideally, a polynomial of low degree, of course).

In our study of the Main Question, we faced the following puzzle.

**Question 1.** Given a positive constant $c \le 1$, for which value of $s'$, depending on $s$, can one find a practical algorithm that constructs, for each symmetric matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, an index set $S'$ of size $s'$ such that

$$\|\mathbf{M}_{S' \times S'}\|_F^2 \ge c \max_{|S| = s} \|\mathbf{M}_{S \times S}\|_F^2 ? \tag{7}$$

In reality, the relevant question for our goal is broader. It involves the projection $P^{[r]}$ onto $\Sigma^{[r]}$.

**Question 2.** Given a positive constant $c \le 1$, for which value of $s'$, depending on $s$, can one find a practical algorithm that constructs, for each symmetric matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, an index set $S'$ of size $s'$ such that, with $r'$ proportional to $r$,

$$\|P^{[r']}(\mathbf{M}_{S' \times S'})\|_F^2 \ge c \max_{|S| = s} \|P^{[r]}(\mathbf{M}_{S \times S})\|_F^2 ? \tag{8}$$

If $s'$ could be chosen proportional to $s$ in Question 2, then the Main Question could be answered with $m \asymp rs \ln(en/s)$ measurements satisfying the so-called restricted isometry property (see below). This is shown in Sec. 4.

We come up with partial answers to the above questions: in Proposition 8 we show that for $c = 1$ the answer to Question 1 is positive with $s' = s^2$, but that it is negative for any $c > 0$ when $s' = O(s)$. Combined with the results of Sec. 4 this establishes that the answer to the Main Question is positive with $m \asymp rs^\gamma \ln(en/s)$ and $\gamma = 2$, using a practical variant of iterative hard thresholding, and that the proposed analysis cannot cover the case $\gamma \le 1$.

In principle, we are more interested in the measurements of type (5). Indeed, in the particular case $r = 1$, the measurements taken on a matrix of the type

$\mathbf{X} = \mathbf{x}\mathbf{x}^\top \in \Sigma_{(s)}^{[1]}$ with an $s$-sparse $\mathbf{x} \in \mathbb{R}^n$ would read

$$y_i = |\langle \mathbf{a}_i, \mathbf{x} \rangle|^2, \quad i \in [\![1 : m]\!]. \tag{9}$$

This is exactly the framework of sparse phaseless recovery (except that everything should be written in the complex setting). In this case, the sample complexity is known [16] to be of the order $m \asymp s \ln(en/s)$, although it is unclear if this can be achieved with independent Gaussian vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$.

**Remark 1.** Similar problems as studied here appear in the context of low-rank tensor recovery where one would like to project onto the intersection of two or more low-rank structures defined by different matricizations. It is NP-hard to compute exact projections and efficiently computable approximate projections are not yet good enough to show low-rank tensor recovery results for corresponding iterative hard thresholding guarantees [23]. They are also considered in the context of sparse PCA from inaccurate and incomplete measurements where the problem of recovering a low-rank matrix with sparse (or compressible) right-singular vectors is analyzed [7]. In this work, a multi-penalty approach called A-T-LAS$_{1,2}$ provably reaches local convergence from a reliable, computable initialization. Other locally convergent methods applied to the recovery of row-sparse (or column-sparse) and low-rank matrices are the sparse power factorization (SPF) and its subspace-concatenated variant (SCSPF), see [19]. While the latter work assumes a high peak-to-average power ratio on the singular vectors of the observed matrix, [13] recently enlarged the class of recoverable matrices by relaxing this constraint.

## 2. Theoretical Sample Complexity

Restricted isometry properties have been central in all sorts of structured recovery problems. It is no surprise that another instance of a restricted isometry property plays a key role here, too. The proof sketch is deferred to the appendix.

**Theorem 2.** *Suppose $\mathbf{A}_1, \ldots, \mathbf{A}_m$ are independent random matrices with independent $\mathcal{N}(0, 1/m)$ entries. Given $\delta > 0$, there exist two values $C, c > 0$ (only depending on $\delta$), such that, with failure probability at most $2\exp(-cm)$,*

$$(1-\delta)\|\mathbf{Z}\|_F^2 \le \|\boldsymbol{\mathcal{A}}(\mathbf{Z})\|_2^2 \le (1+\delta)\|\mathbf{Z}\|_F^2 \quad \text{for all } \mathbf{Z} \in \Sigma_{(s)}^{[r]} \tag{10}$$

*provided $m \ge Crs\ln(en/s)$.*

For the rest of this section, we place ourselves in the situation where the measurement map $\boldsymbol{\mathcal{A}}$ satisfies the restricted isometry property (10), which can occur as soon as $m$ is of the order $rs\ln(en/s)$. We can then propose several robust algorithms that recover $\mathbf{X} \in \Sigma_{(s)}^{[r]}$ from $\mathbf{y} = \boldsymbol{\mathcal{A}}(\mathbf{X}) + \mathbf{e}$. The first obvious candidate is

$$\Delta(\mathbf{y}) = \underset{\mathbf{Z} \in \Sigma_{(s)}^{[r]}}{\operatorname{argmin}} \|\mathbf{y} - \boldsymbol{\mathcal{A}}(\mathbf{Z})\|_2. \tag{11}$$

We immediately see that $\|\mathbf{y} - \boldsymbol{\mathcal{A}}(\Delta(\mathbf{y}))\|_2 \leq \|\mathbf{y} - \boldsymbol{\mathcal{A}}(\mathbf{X})\|_2 = \|\mathbf{e}\|_2$, from where it follows that

$$\|\boldsymbol{\mathcal{A}}(\mathbf{X}) - \boldsymbol{\mathcal{A}}(\Delta(\mathbf{y}))\|_2 \leq \|\mathbf{y} - \boldsymbol{\mathcal{A}}(\Delta(\mathbf{y}))\|_2 + \|\mathbf{e}\|_2 \leq 2\|\mathbf{e}\|_2, \tag{12}$$

and we finally derive that

$$\|\mathbf{X} - \Delta(\boldsymbol{\mathcal{A}}(\mathbf{X}) + \mathbf{e})\|_F \leq \frac{1}{\sqrt{1-\delta}}\|\boldsymbol{\mathcal{A}}(\mathbf{X}) - \boldsymbol{\mathcal{A}}(\Delta(\boldsymbol{\mathcal{A}}(\mathbf{X}) + \mathbf{e}))\|_2 \leq \frac{2}{\sqrt{1-\delta}}\|\mathbf{e}\|_2. \tag{13}$$

However, this scheme is not really an appropriate candidate, since producing $\Delta(\mathbf{y})$ is NP-hard in general (see below).

After a decade or so of $\ell_1$-norm and nuclear norm minimizations, the next obvious candidate stands out as

$$\Delta(\mathbf{y}) = \underset{\mathbf{Z} \in \mathbb{R}^{n \times n}}{\operatorname{argmin}} F(\mathbf{Z}) \quad \text{subject to } \|\mathbf{y} - \boldsymbol{\mathcal{A}}(\mathbf{Z})\|_2 \leq \|\mathbf{e}\|_2, \tag{14}$$

where $F$ is a convex function promoting the joint low-rank and bisparsity structure. The negative results from [22] indicate that reducing the sample complexity below $\min\{rn, s^2 \ln(en/s)\}$ is unattainable when $F$ is a positive combination of the $\ell_1$-norm and nuclear norm.

What about a variant of iterative hard thresholding? Consider the sequence $(\mathbf{X}_k)_{k \geq 0}$ defined by

$$\mathbf{X}_{k+1} = P_{(s)}^{[r]}(\mathbf{X}_k + \boldsymbol{\mathcal{A}}^*(\mathbf{y} - \boldsymbol{\mathcal{A}}(\mathbf{X}_k))), \tag{15}$$

where the adjoint of $\boldsymbol{\mathcal{A}}$ is given by

$$\boldsymbol{\mathcal{A}}^* : \mathbf{u} \in \mathbb{R}^m \mapsto \sum_{i=1}^m u_i \mathbf{A}_i \in \mathbb{R}^{n \times n}$$

and where $P_{(s)}^{[r]} : \mathbb{R}^{n \times n} \to \Sigma_{(s)}^{[r]}$ denotes the projection onto $\Sigma_{(s)}^{[r]}$, that is, the operator of best approximation from $\Sigma_{(s)}^{[r]}$. One can show (see Appendix A or [2]) that if $\Delta(\mathbf{y})$ is defined as a cluster point of $(\mathbf{X}_k)_{k \geq 0}$, then

$$\|\mathbf{X} - \Delta(\boldsymbol{\mathcal{A}}(\mathbf{X}) + \mathbf{e})\|_F \leq C\|\mathbf{e}\|_2 \tag{16}$$

holds for all $\mathbf{X} \in \Sigma_{(s)}^{[r]}$ and all $\mathbf{e} \in \mathbb{R}^m$. Here also the issue is that computing $P_{(s)}^{[r]}$ is NP-hard (see Sec. 5), which incidentally justifies the NP-hardness of (11) (think of $\boldsymbol{\mathcal{A}} = \mathbf{I}$). What about replacing $P_{(s)}^{[r]}$ by an operator of near-best approximation from $\Sigma_{(s)}^{[r]}$, as in, e.g., [14]? After all, if there is any chance for (6) to hold, then such an operator must exist (think again of $\boldsymbol{\mathcal{A}} = \mathbf{I}$). We will in fact construct such an operator in Sec. 5.3. But substituting $P_{(s)}^{[r]}$ by such an operator in the proof of Theorem A.2 (see Appendix A) is not enough to do the trick.

## 3. Optimal Sample Complexity with Factorized Measurements

In this section, we show that the optimal sample complexity can be achieved with a practical algorithm in a rather special measurement framework. This framework being restricted to the specific structure of this sensing procedure, the Main Question remains of interest.

We suppose here that matrices $\mathbf{X} \in \Sigma_{(s)}^{[r]}$ are acquired via measurements in factorized form, namely

$$y_i = \langle \mathbf{X}, \mathbf{B}^\top \mathbf{A}_i \mathbf{B} \rangle, \quad i \in [\![1 : m]\!], \tag{17}$$

where $\mathbf{A}_1, \dots, \mathbf{A}_m \in \mathbb{R}^{p \times p}$ allow for low-rank recovery and $\mathbf{B} \in \mathbb{R}^{p \times n}$ allows for sparse recovery. The recovery algorithm proceeds in two steps, which are both practical, i.e. efficiently implementable.

(1) Compute $\mathbf{Y}^\sharp \in \mathbb{R}^{p \times p}$ from $\mathbf{y} \in \mathbb{R}^m$ as a solution of the nuclear norm minimization

$$\underset{\mathbf{Y} \in \mathbb{R}^{p \times p}}{\text{minimize}} \|\mathbf{Y}\|_* \quad \text{subject to } \langle \mathbf{Y}, \mathbf{A}_i \rangle_F = y_i, \ \ i \in [\![1 : m]\!],$$

or as the output of another low-rank recovery algorithm such as iterative hard thresholding.

(2) Compute $\mathbf{X}^\sharp \in \mathbb{R}^{n \times n}$ from $\mathbf{Y}^\sharp$ as the output of the HiHTP algorithm with measurement map $\boldsymbol{\mathcal{B}} : \mathbf{Z} \in \mathbb{R}^{n \times n} \mapsto \mathbf{B}\mathbf{Z}\mathbf{B}^\top \in \mathbb{R}^{p \times p}$.

Although we refer to [24, 25] for the exact formulation of the hierarchically structured sparsity hard thresholding pursuit (HiHTP) algorithm, a few words about the concept of hierarchical sparsity are in order before we state our result about the two-step recovery procedure above. A matrix is said to be $(s,t)$-hierachical sparse (or simply $(s,t)$-sparse) if at most $s$ of its columns are nonzero and each of these columns possesses at most $t$ nonzero entries. Thus, $s$-bisparse matrices are in particular $(s,s)$-sparse. The HiHTP algorithm essentially relies on the possibility to compute the projection (operator of best approximation) onto $(s,t)$-sparse matrices. In contrast to the projection onto $s$-bisparse matrices, this is indeed an easy task: first, select the $t$ largest absolute entries in each column and calculate the resulting $\ell_2$-norm, then select the $s$ columns with the largest of these $\ell_2$-norms.

**Theorem 3.** *Let $\mathbf{A}_1, \dots, \mathbf{A}_m \in \mathbb{R}^{p \times p}$ be independent standard Gaussian matrices and let $\mathbf{B} \in \mathbb{R}^{p \times n}$ be a standard Gaussian matrix independent of $\mathbf{A}_1, \dots, \mathbf{A}_m$. If*

$$p \asymp s \ln(en/s) \quad \text{and} \quad m \asymp rp, \tag{18}$$

*so that $m \asymp rs \ln(en/s)$, then the probability that every $\mathbf{X} \in \Sigma_{(s)}^{[r]}$ is exactly recovered from $y_i = \langle \mathbf{X}, \mathbf{B}^\top \mathbf{A}_i \mathbf{B} \rangle$, $i \in [\![1 : m]\!]$, via the above two-step procedure is at least $1 - 2\exp(-cp)$.*

**Proof.** First, notice that the matrix $\mathbf{B}\mathbf{X}\mathbf{B}^\top \in \mathbb{R}^{p\times p}$ has rank at most $r$, since $\mathbf{X}$ has rank at most $r$, and that it satisfies

$$\begin{aligned}\langle \mathbf{B}\mathbf{X}\mathbf{B}^\top, \mathbf{A}_i\rangle_F &= \mathrm{tr}(\mathbf{A}_i^\top \mathbf{B}\mathbf{X}\mathbf{B}^\top) = \mathrm{tr}(\mathbf{B}^\top \mathbf{A}_i^\top \mathbf{B}\mathbf{X}) \\ &= \langle \mathbf{X}, \mathbf{B}^\top \mathbf{A}_i \mathbf{B}\rangle_F = y_i, \quad i \in [\![1:m]\!]. \end{aligned} \tag{19}$$

Since $\mathbf{A}_1, \dots, \mathbf{A}_m \in \mathbb{R}^{p\times p}$ are independent standard Gaussian matrices and $m \asymp rp$, it is by now well-known (see, e.g., [4, 17]) that, with failure probability at most $\exp(-cm)$, the matrix $\mathbf{B}\mathbf{X}\mathbf{B}^\top$ is recovered via nuclear norm minimization (or another suitable algorithm), so that $\mathbf{Y}^\sharp = \mathbf{B}\mathbf{X}\mathbf{B}^\top$.

Second, since the matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ is $(s,s)$-sparse and satisfies $\boldsymbol{\mathcal{B}}(\mathbf{X}) = \mathbf{B}\mathbf{X}\mathbf{B}^\top = \mathbf{Y}^\sharp$, [24, Theorem 1] implies that the matrix $\mathbf{X}$ will be exactly recovered via HiHTP as long as the so-called HiRIP of order $(3s, 2s)$ holds. According to [25, Theorem 1], the latter is satisfied when $\mathbf{B}$ obeys a standard RIP, and the latter is indeed fulfilled with failure at most $\exp(-cp)$ by the matrix $\mathbf{B}$ (or rather by a renormalization of it), because $\mathbf{B} \in \mathbb{R}^{p\times n}$ is a standard Gaussian matrix with $p \asymp s\ln(en/s)$.

All in all, exact recovery of $\mathbf{X}$ is guaranteed after the two steps with failure probability bounded by $\exp(-cm) + \exp(-cp) \le 2\exp(-cp)$. □

**Remark 4.** It is possible to extend Theorem 3 beyond the strictly Gaussian setting. In particular, if $\mathbf{A}_1, \dots, \mathbf{A}_m$ take the form $\mathbf{A}_i = \mathbf{a}_i\mathbf{a}_i^\top$ for some independent standard Gaussian vectors $\mathbf{a}_i \in \mathbb{R}^p$, then the first-step recovery of $\mathbf{B}\mathbf{X}\mathbf{B}^\top$ can still be achieved via nuclear norm minimization (see [3, 17, 18]) or by some modified iterative hard thresholding algorithm (see [12]). Note that the measurements made on $\mathbf{X} \in \mathbb{R}^{n\times n}$ are in this case rank-one measurements given by $y_i = \langle \mathbf{X}\mathbf{a}_i', \mathbf{a}_i'\rangle$, where $\mathbf{a}_i' := \mathbf{B}^\top \mathbf{a}_i$.

**Remark 5.** Let us mention that sensing strategies similar to (17) have been proposed before for other objects with related structures or for connected problems. For instance, when estimating $k$-row-sparse and $r$-rank matrices $\mathbf{X} \in \mathbb{R}^{n\times n}$ from $m$ "nested" measurements $y_i = \langle \mathbf{W}\mathbf{X}, \mathbf{A}_i\rangle$, [1] showed that RIP conditions imposed on $\mathbf{W} \in \mathbb{R}^{p\times n}$ and on the linear operator associated with $\mathbf{A}_1, \dots, \mathbf{A}_m$ yield a computationally efficient two-stage method that can (nearly) achieve a minimax lower bound from $m \asymp r\max\{p,n\}$ measurements where $p \asymp k\log(n/k)$, i.e. from $m \asymp \max\{rk\log(n/k), rn\}$. A two-stage sensing strategy has been also proposed in [16] for the sparse phase retrieval problem. In this case, the sensing model is factored into a linear operator with robust null space property and a stable phase retrieval matrix — the latter allows to recover a compressed form of the sparse vector, using e.g., PhaseLift [5], and then the former allows to recover this vector via any compressive sensing algorithm.

## 4. Toward Practical Sample Complexity

In most scenarios, the measurement map is not of the factorized type considered in the previous section, so the two-step procedure cannot even be executed. It is therefore still relevant to search for practical recovery algorithms that can be applied with arbitrary measurement schemes and study the sample complexity using, e.g., Gaussian measurements. As mentioned at the end of Sec. 2, a difficulty occurs when one tries to use a near-best approximation operator instead of the best approximation operator $P_{(s)}^{[r]}$ in the iterative hard thresholding algorithm (15). Such a difficulty was also encountered in model-based compressive sensing. A workaround was found in [15]. As we will see below, our attempt to imitate it prompted Question 2.

Let us start with the observation that any of the structures $\Sigma_{(s)}$, $\Sigma^{[r]}$, or $\Sigma_{(s)}^{[r]}$ is a union of subspaces, which we generically write as

$$\Sigma = \bigcup_{V \in \mathcal{V}_\Sigma} V.$$

Then the projection onto $\Sigma$, i.e. the operator of best approximation from $\Sigma$ with respect to the Frobenius norm, acts on any $\mathbf{M} \in \mathbb{R}^{n \times n}$ via

$$P_\Sigma(\mathbf{M}) = P_{V(\mathbf{M})}(\mathbf{M}), \tag{20}$$

where

$$V(\mathbf{M}) = \operatorname*{argmin}_{V \in \mathcal{V}_\Sigma} \|\mathbf{M} - P_V(\mathbf{M})\|_F^2 \tag{21}$$

$$= \operatorname*{argmax}_{V \in \mathcal{V}_\Sigma} \|P_V(\mathbf{M})\|_F^2, \tag{22}$$

and $P_V$ evidently denotes the orthogonal projection onto the subspace $V$. By analogy with the vector case, we can think of (21) as a 'tail' property for the projection $P_\Sigma$ and of (22) as a 'head' property. We keep this terminology introduced in [15] when relaxing the notion of projection. Precisely, we shall call an operator $T : \mathbb{R}^{n \times n} \to \Sigma$ a tail projection for $\Sigma$ with constant $C_T \geq 1$ (or near best approximation from $\Sigma$ with constant $C_T$) if

$$\|\mathbf{M} - T(\mathbf{M})\|_F \leq C_T \|\mathbf{M} - P_\Sigma(\mathbf{M})\|_F \quad \text{for all } \mathbf{M} \in \mathbb{R}^{n \times n}. \tag{23}$$

We may have to relax this notion further by allowing the operator $T$ to map into a bigger set $\Sigma' \supseteq \Sigma$. Thus, by tail projection for $\Sigma$ into $\Sigma'$ with constant $C_T$, we mean an operator $T : \mathbb{R}^{n \times n} \to \Sigma'$ which satisfies the tail condition (23). Similarly, an operator $H : \mathbb{R}^{n \times n} \to \Sigma$ is called a head projection for $\Sigma$ with constant $c_H \leq 1$ if

$$\|H(\mathbf{M})\|_F \geq c_H \|P_\Sigma(\mathbf{M})\|_F \quad \text{for all } \mathbf{M} \in \mathbb{R}^{n \times n}. \tag{24}$$

A head projection for $\Sigma$ into $\Sigma' \supseteq \Sigma$ with constant $c_H$ is an operator $H : \mathbb{R}^{n \times n} \to \Sigma'$ which satisfies the head condition (24).

At this point, it is worth mentioning (see Appendix A) that the (genuine) projection onto $\Sigma^{[r]}_{(s)}$ acts on any $\mathbf{M} \in \mathbb{R}^{n\times n}$ via

$$P^{[r]}_{(s)}(\mathbf{M}) = P^{[r]}(\mathbf{M}_{S_\star \times S_\star}), \quad \text{where } S_\star = \operatorname*{argmax}_{|S|=s} \|P^{[r]}(\mathbf{M}_{S\times S})\|_F. \tag{25}$$

In Sec. 5, we will see that we can produce a tail projection for $\Sigma^{[r]}_{(s)}$.

The size of $s'$ for which one can produce a head projection for $\Sigma^{[r]}_{(s)}$ into $\Sigma^{[r']}_{(s')}$ with $r'$ proportional to $r$ is exactly the focus of Question 2. We state and prove below (in the idealized setting where there is no measurement error) that a variant of iterative hard thresholding — using such a head projection — allows to perform joint low-rank and bisparse recovery via from $m \asymp rs' \ln(en/s)$ measurements. This will be interesting if it can be established that $s' \asymp s^\gamma$ with $\gamma < 2$ is feasible. Then, for small $r$ (and in particular in the case of sparse phaseless recovery where $r = 1$), $m \asymp rs^\gamma \ln(en/s)$ will be of a smaller order than both $rn$ — the sample complexity of rank-$r$ matrices — and $s^2 \ln(en/s)$. This last bound is associated with enforcing only the matrix bisparse structure, as ensured by combining Theorem 6 in the case $r = n$ with Proposition 8 and Theorem 2 (see below). Quite obviously this last context determines that $\gamma = 2$ is feasible (as stated in the abstract) since $s^2 \le rs^2$.

**Theorem 6.** *Let $T$ be a tail projection for $\Sigma^{[r]}_{(s)}$ with constant $C_T \ge 1$ and let $H$ be a head projection for $\Sigma^{[2r]}_{(2s)}$ into $\Sigma^{[r']}_{(s')}$ with constant $c_H \le 1$ which additionally takes the form*

$$H(\mathbf{M}) = P^{[r']}(\mathbf{M}_{S'\times S'}) \quad \text{for some index set } S' \text{ (depending on } \mathbf{M}) \text{ of size } s'. \tag{26}$$

*If $(1 + C_T)^2(1 - c_H^2) < 1$ and if the restricted isometry property (10) holds on $\Sigma^{[2r+r']}_{(2s+s')}$ with constant $\delta > 0$ small enough to have*

$$\rho := (1 + C_T)^2(1 - c_H^2(1 - \delta)^2 + 2\delta(1 + \delta)) < 1, \tag{27}$$

*then any $\mathbf{X} \in \Sigma^{[r]}_{(s)}$ acquired from $\mathbf{y} = \boldsymbol{\mathcal{A}}(\mathbf{X})$ is recovered as the limit of the sequence $(\mathbf{X}_k)_{k\ge 0}$ defined by*

$$\mathbf{X}_{k+1} = T[\mathbf{X}_k + H(\boldsymbol{\mathcal{A}}^*(\mathbf{y} - \boldsymbol{\mathcal{A}}(\mathbf{X}_k)))]. \tag{28}$$

**Proof.** We shall prove that, for any $k \ge 0$,

$$\|\mathbf{X} - \mathbf{X}_{k+1}\|_F^2 \le \rho \, \|\mathbf{X} - \mathbf{X}_k\|_F^2. \tag{29}$$

The tail property guarantees that

$$\|[\mathbf{X}_k + H(\boldsymbol{\mathcal{A}}^*(\mathbf{y} - \boldsymbol{\mathcal{A}}(\mathbf{X}_k)))] - \mathbf{X}_{k+1}\|_F \le C_T \|[\mathbf{X}_k + H(\boldsymbol{\mathcal{A}}^*(\mathbf{y} - \boldsymbol{\mathcal{A}}(\mathbf{X}_k)))] - \mathbf{X}\|_F \tag{30}$$

and the triangle inequality then yields[a]

$$\|\mathbf{X} - \mathbf{X}_{k+1}\|_F \le (1 + C_T)\|[\mathbf{X}_k + H(\boldsymbol{\mathcal{A}}^*(\mathbf{y} - \boldsymbol{\mathcal{A}}(\mathbf{X}_k)))] - \mathbf{X}\|_F. \tag{31}$$

We now concentrate on bounding $\|[\mathbf{X}_k + H(\boldsymbol{\mathcal{A}}^*(\mathbf{y} - \boldsymbol{\mathcal{A}}(\mathbf{X}_k)))] - \mathbf{X}\|_F = \|\mathbf{Z} - H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\|_F$, where we have set $\mathbf{Z} := \mathbf{X} - \mathbf{X}_k \in \Sigma_{(2s)}^{[2r]}$. By expanding the square, we obtain

$$\begin{aligned}\|\mathbf{Z} - H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\|_F^2 &= \|\mathbf{Z}\|_F^2 + \|H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\|_F^2 - 2\langle \mathbf{Z}, H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\rangle_F \\ &= \|\mathbf{Z}\|_F^2 + \|H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\|_F^2 - 2\langle \boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}), H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\rangle_F \\ &\quad - 2\langle \mathbf{Z} - \boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}), H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\rangle_F. \end{aligned} \tag{32}$$

In view of the form (26) of the head projection, followed by the facts that $P^{[r']}$ acts locally as an orthogonal projection and that it preserves the bisupport of a matrix, we observe that

$$\begin{aligned}\|H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\|_F^2 &= \langle P^{[r']}(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z})_{S'\times S'}), P^{[r']}(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z})_{S'\times S'})\rangle_F \\ &= \langle \boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z})_{S'\times S'}, P^{[r']}(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z})_{S'\times S'})\rangle_F \\ &= \langle \boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}), P^{[r']}(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z})_{S'\times S'})\rangle_F \\ &= \langle \boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}), H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\rangle_F. \end{aligned} \tag{33}$$

Substituting the latter into (32) gives

$$\begin{aligned}\|\mathbf{Z} - H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\|_F^2 &= \|\mathbf{Z}\|_F^2 - \|H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\|_F^2 \\ &\quad - 2\langle \mathbf{Z} - \boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}), H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\rangle_F. \end{aligned} \tag{34}$$

The inner product term is small in absolute value. Indeed, in view of Lemma A.1 (see Appendix A), we have

$$|\langle \mathbf{Z} - \boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}), H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\rangle_F| \le \delta\|\mathbf{Z}\|_F\|H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\|_F \le \delta(1+\delta)\|\mathbf{Z}\|_F^2, \tag{35}$$

where the bound on $\|H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\|_F$ followed from the observation (33) and the restricted isometry property (10), according to

$$\begin{aligned}\|H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\|_F^2 &= \langle \boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}), H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\rangle_F = \langle \boldsymbol{\mathcal{A}}(\mathbf{Z}), \boldsymbol{\mathcal{A}}(H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z})))\rangle_F \\ &\le \|\boldsymbol{\mathcal{A}}(\mathbf{Z})\|_F\|\boldsymbol{\mathcal{A}}(H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z})))\|_F \le (1+\delta)\|\mathbf{Z}\|_F\|H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\|_F. \end{aligned} \tag{36}$$

It now remains to prove that $\|H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\|_F^2$ is large, and this is where the head condition comes into play. Precisely, assuming that $\mathbf{Z}$ is supported on $S'' \times S''$ with

[a]It is probably possible to replace $1 + C_T$ by a constant arbitrarily close to 1 if $T$ mapped into $\Sigma_{(s'')}^{[r'']}$ with $r''$ and $s''$ proportional to $r$ and $s$ (with proportionality constant increasing when $C_T$ decreases), as in [26] for the sparse vector case and in [12] for the low-rank matrix case. This would allow us to eliminate the condition $(1 + C_T)^2(1 - c_H^2) < 1$.

$|S''| \leq 2s$, we know on the one hand that

$$\|H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\|_F \geq c_H \|P^{[2r]}(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z})_{S''\times S''})\|_F. \tag{37}$$

On the other hand, using in particular the restricted isometry property (10) and Von Neumann's trace inequality combined with the fact that $\mathbf{Z}$ has rank at most $2r$, we obtain

$$\begin{aligned}(1-\delta)\|\mathbf{Z}\|_F^2 \leq \|\boldsymbol{\mathcal{A}}(\mathbf{Z})\|_2^2 &= \langle \mathbf{Z}, \boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z})\rangle_F = \langle \mathbf{Z}, \boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z})_{S''\times S''}\rangle_F \\ &\leq \sum_{i=1}^{2r} \sigma_i(\mathbf{Z})\sigma_i(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z})_{S''\times S''}) \\ &\leq \left[\sum_{i=1}^{2r}\sigma_i(\mathbf{Z})^2\right]^{1/2}\left[\sum_{i=1}^{2r}\sigma_i(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z})_{S''\times S''})^2\right]^{1/2} \\ &= \|\mathbf{Z}\|_F \|P^{[2r]}(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z})_{S''\times S''})\|_F. \end{aligned} \tag{38}$$

Combining (37) and (38) yields

$$\|H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\|_F \geq c_H(1-\delta)\|\mathbf{Z}\|_F. \tag{39}$$

Substituting (39) and (35) into (34), we deduce that

$$\|\mathbf{Z} - H(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{Z}))\|_F^2 \leq (1 - c_H^2(1-\delta)^2 + 2\delta(1+\delta))\|\mathbf{Z}\|_F^2. \tag{40}$$

Finally, using (31), we arrive that

$$\|\mathbf{X} - \mathbf{X}_{k+1}\|_F^2 \leq (1+C_T)^2(1 - c_H^2(1-\delta)^2 + 2\delta(1+\delta))\|\mathbf{X} - \mathbf{X}_k\|_F^2, \tag{41}$$

which is the objective announced in (29). □

## 5. Tail and Head Projections

In this section, we gather some information about the construction of computable tail and head projections for each of the structures $\Sigma^{[r]}$, $\Sigma_{(s)}$, and $\Sigma_{(s)}^{[r]}$. We work under the implicit assumption that the domain of all these projections is the space of symmetric matrices, i.e. the projections are only applied to matrices $\mathbf{M} \in \mathbb{R}^{n\times n}$ satisfying $\mathbf{M}^\top = \mathbf{M}$.

### 5.1. *Low-rank structure*

There is no difficulty whatsoever here — even the exact projection $P^{[r]} : \mathbb{R}^{n\times n} \to \Sigma^{[r]}$ is accessible. Indeed, it is well known that if $\mathbf{X} \in \mathbb{R}^{n\times n}$ has singular value decomposition

$$\mathbf{X} = \sum_{i=1}^{n} \sigma_i(\mathbf{X})\mathbf{u}_i\mathbf{v}_i^\top, \tag{42}$$

where the singular values $\sigma_1(\mathbf{X}) \geq \cdots \geq \sigma_n(\mathbf{X}) \geq 0$ are arranged in nondecreasing order, then the projection of $\mathbf{X}$ onto the set of rank-$r$ matrices is obtained by

truncating this decomposition to include only the first $r$ summands, i.e.

$$P^{[r]}(\mathbf{X}) = \sum_{i=1}^{r} \sigma_i(\mathbf{X}) \mathbf{u}_i \mathbf{v}_i^\top. \tag{43}$$

Note that $P^{[r]}(\mathbf{M})$ is symmetric whenever $\mathbf{M}$ itself is symmetric.

### 5.2. *Bisparsity structure*

Quickly stated, exact projections for $\Sigma_{(s)}$ are NP-hard, but there are computable tail projections for $\Sigma_{(s)}$. Head projections for $\Sigma_{(s)}$ are still NP-hard if they are forced to map exactly into $\Sigma_{(s)}$.

If they are allowed to map into a larger set $\Sigma_{(s')}$, the situation depends on the order of $s'$ compared to $s$ — Question 1 in fact asks which value of $s' > s$ allows for a computable head projection.

We provide a few incomplete results related to this situation.

**Exact projection.** Finding the exact projection for $\Sigma_{(s)}$ amounts to solving the problem

$$\underset{|S|=s}{\text{maximize}} \, \|\mathbf{M}_{S\times S}\|_F^2. \tag{44}$$

This is NP-hard even with the restriction that $\mathbf{M}$ is an adjacency matrix of a graph because it then reduces to the densest $k$-subgraph problem, which is known to be NP-hard [21].

**Tail projections.** There is a simple procedure to obtain a practical tail projection for $\Sigma_{(s)}$, as described in the following.

**Proposition 7.** *Given a symmetric matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$, let $S_\star$ denote an index set corresponding to $s$ columns of $\mathbf{M}$ with largest $\ell_2$-norms, i.e.*

$$S_\star = \underset{|S|=s}{\operatorname{argmin}} \|\mathbf{M} - \mathbf{M}_{:\times S}\|_F. \tag{45}$$

*Then*

$$\|\mathbf{M} - \mathbf{M}_{S_\star \times S_\star}\|_F \le \sqrt{2} \min_{|S|=s} \|\mathbf{M} - \mathbf{M}_{S\times S}\|_F. \tag{46}$$

**Proof.** For any index set $T$, the symmetry of $\mathbf{M}$ imposes that $\|\mathbf{M}_{\overline{T}\times T}\|_F^2 = \|\mathbf{M}_{T\times\overline{T}}\|_F^2$, hence

$$\begin{aligned} \|\mathbf{M} - \mathbf{M}_{T\times T}\|_F^2 &= \|\mathbf{M}_{T\times\overline{T}}\|_F^2 + \|\mathbf{M}_{\overline{T}\times T}\|_F^2 + \|\mathbf{M}_{\overline{T}\times\overline{T}}\|_F^2 \\ &= 2\|\mathbf{M}_{T\times\overline{T}}\|_F^2 + \|\mathbf{M}_{\overline{T}\times\overline{T}}\|_F^2. \end{aligned} \tag{47}$$

In view of $\|\mathbf{M}_{T\times\overline{T}}\|_F^2 + \|\mathbf{M}_{\overline{T}\times\overline{T}}\|_F^2 = \|\mathbf{M}_{:\times\overline{T}}\|_F^2 = \|\mathbf{M} - \mathbf{M}_{:\times T}\|_F^2$, we deduce that

$$\|\mathbf{M} - \mathbf{M}_{:\times T}\|_F^2 \le \|\mathbf{M} - \mathbf{M}_{T\times T}\|_F^2 \le 2\|\mathbf{M} - \mathbf{M}_{:\times T}\|_F^2. \tag{48}$$

**Algorithm 1.** A head projection $H$ for $\Sigma_{(s)}$ to $\Sigma_{(s^2)}$ with $c_H = 1$

**Input:** A symmetric matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, a sparsity level $s \in [\![1 : n]\!]$.

**for** $i \in [\![1 : n]\!]$ **do**

$C_i := \underset{|C|=s-1, C \not\ni i}{\operatorname{argmax}} \|\mathbf{M}_{i \times (\{i\} \cup C)}\|_2$

$c_i := \|\mathbf{M}_{i \times (\{i\} \cup C_i)}\|_2$

**end**

$R := \underset{|R'|=s}{\operatorname{argmax}} \|\mathbf{c}_{R'}\|$, with $\mathbf{c} = (c_1, \ldots, c_n)^\top$

$S' := R \cup (\cup_{i \in R} C_i)$

**return** $H(\mathbf{M}) := \mathbf{M}_{S' \times S'} \in \Sigma_{(s^2)}$

Applying the latter with $T$ equal to $S_\star$ and with $T$ equal to an arbitrary index set $S$ of size $s$ shows that

$$\|\mathbf{M} - \mathbf{M}_{S_\star \times S_\star}\|_F^2 \le 2\|\mathbf{M} - \mathbf{M}_{:\times S_\star}\|_F^2 \le 2\|\mathbf{M} - \mathbf{M}_{:\times S}\|_F^2 \le 2\|\mathbf{M} - \mathbf{M}_{S \times S}\|_F^2, \tag{49}$$

which yields the required result after taking the square root. □

**Head projections.** The literature on the densest $k$-subgraph problem informs us that finding a head projection for $\Sigma_{(s)}$ is also an NP-hard problem [21]. In our setting, though, there is room to relax the head projection to map into $\Sigma_{(s')}$ with $s' > s$. In this regard, Question 1 asks if one can actually compute a head projection for $\Sigma_{(s)}$ into $\Sigma_{(s')}$. We do not have a definite answer for it, but we prove below that the exponent $\gamma$ in a speculative behavior $s' \asymp s^\gamma$ must lie in $(1, 2]$ — note that a behavior $s' \asymp s\,\mathrm{polylog}(s)$ is not excluded. We then highlight a few observations which feature a nonabsolute constant $c_H$ when $s' \asymp s$.

**Proposition 8.** *The practical algorithm Algorithm* 1 *yields a head projection for* $\Sigma_{(s)}$ *into* $\Sigma_{(s^2)}$ *with constant* $c_H = 1$. *However, there is no practical algorithm that yields a head projection for* $\Sigma_{(s)}$ *into* $\Sigma_{(s')}$ *with absolute constant* $c_H > 0$ *when* $s' = O(s)$.

**Proof.** From the definition of the index sets $\{C_i : 1 \le i \le n\}$, $R$ and $S'$ in Algorithm 1, for any index set $S$ with $|S| = s$, we have

$$\begin{aligned} \|\mathbf{M}_{S \times S}\|_F^2 &= \sum_{i \in S} \|\mathbf{M}_{i \times S}\|_2^2 \le \sum_{i \in S} \|\mathbf{M}_{i \times (\{i\} \cup C_i)}\|_2^2 \le \sum_{i \in R} \|\mathbf{M}_{i \times (\{i\} \cup C_i)}\|_2^2 \\ &= \|\mathbf{M}_{R \times (R \cup \bigcup_{i \in R} C_i)}\|_F^2 \le \|\mathbf{M}_{S' \times S'}\|_F^2, \end{aligned} \tag{50}$$

where the index set $S' = R \cup \bigcup_{i \in R} C_i$ has size at most $s + s(s-1) = s^2$. This proves the first part of the statement.

For the second part of the statement, we shall show that if we could compute, for each $\mathbf{M} \in \mathbb{R}^{n\times n}$, an index set $S'$ with $|S'| \le Cs$ such that

$$\|\mathbf{M}_{S'\times S'}\|_F^2 \ge c_H^2 \max_{|S|=s} \|\mathbf{M}_{S\times S}\|_F^2, \tag{51}$$

then a practical algorithm that yields a head approximation for $\Sigma_{(s)}$ into $\Sigma_{(s)}$ itself would follow, contradiction the NP-hardness of the latter task. So let us assume that we have a computable procedure to construct an index set $S'$ as above. Looking without loss of generality at the case where $s$ is even and $|S'| = Cs$, we consider an index set $R \subseteq S'$ of size $s/2$ corresponding to $s/2$ largest values of $\|\mathbf{M}_{i\times S'}\|_2$. By comparing averages, we see that

$$\frac{1}{s/2}\|\mathbf{M}_{R\times S'}\|_F^2 \ge \frac{1}{Cs}\|\mathbf{M}_{S'\times S'}\|_F^2, \quad \text{i.e. } \|\mathbf{M}_{R\times S'}\|_F^2 \ge \frac{1}{2C}\|\mathbf{M}_{S'\times S'}\|_F^2. \tag{52}$$

Next, we consider an index set $C \subseteq S'$ of size $s/2$ corresponding to $s/2$ largest values of $\|\mathbf{M}_{R\times j}\|_2$. By comparing averages again, we see that

$$\frac{1}{s/2}\|\mathbf{M}_{R\times C}\|_F^2 \ge \frac{1}{Cs}\|\mathbf{M}_{R\times S'}\|_F^2, \quad \text{i.e. } \|\mathbf{M}_{R\times C}\|_F^2 \ge \frac{1}{2C}\|\mathbf{M}_{R\times S'}\|_F^2. \tag{53}$$

Combining (53), (52), and (51), we arrive at

$$\|\mathbf{M}_{R\times C}\|_F^2 \ge \frac{c_H^2}{4C^2} \max_{|S|=s} \|\mathbf{M}_{S\times S}\|_F^2. \tag{54}$$

With $T := R \cup C$, which has size at most $s$, this immediately implies that

$$\|\mathbf{M}_{T\times T}\|_F^2 \ge \frac{c_H^2}{4C^2} \max_{|S|=s} \|\mathbf{M}_{S\times S}\|_F^2, \tag{55}$$

meaning that a head approximation for $\Sigma_{(s)}$ into $\Sigma_{(s)}$ can be produced in a practical way. Since this is not possible, the second part of the statement is proved. □

Now that we have established the impracticability of head approximations for $\Sigma_{(s)}$ into $\Sigma_{(Cs)}$ with an absolute constant $c_H$, we examine what can be done when $c_H$ can depend on specific parameters.

**Proposition 9.** *Given a symmetric matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$, we consider the practical algorithm that returns the matrix $\mathbf{M}_{T\times T}$ for a set $T := R \cup C$ defined by the union of the index sets of size $s$*

$$R = \operatorname*{argmax}_{|S|=s} \|\mathbf{M}_{S\times :}\|_F^2, \tag{56}$$

$$C = \operatorname*{argmax}_{|S|=s} \|\mathbf{M}_{R\times S}\|_F^2. \tag{57}$$

*This algorithm yields a head projection for $\Sigma_{(s)}$ into $\Sigma_{(2s)}$ with constant $c_H = \sqrt{s/n}$.*

**Algorithm 2.** A head projection $H$ for $\Sigma_{(s)}$ with $c_H = 1/\sqrt{s}$

**Input:** A symmetric matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, a sparsity level $s \in [\![1 : n]\!]$.

**for** $j \in [\![1 : n]\!]$ **do**

$S_j := \underset{|S|=s, S \ni j}{\operatorname{argmax}} \|\mathbf{M}_{S \times j}\|_2^2$

**end**

$j_\star := \underset{j \in [\![1:n]\!]}{\operatorname{argmax}} \|\mathbf{M}_{S_j \times j}\|_2^2$

**return** $H(\mathbf{M}) := \mathbf{M}_{S_{j_\star} \times S_{j_\star}} \in \Sigma_{(s)}$

**Proof.** From the definition of $R$ and $C$, it is painless to see that, for an arbitrary index set $S$ of size $s$,

$$\|\mathbf{M}_{T \times T}\|_F^2 \geq \|\mathbf{M}_{R \times C}\|_F^2 \geq \frac{s}{n}\|\mathbf{M}_{R \times :}\|_F^2 \geq \frac{s}{n}\|\mathbf{M}_{S \times :}\|_F^2 \geq \frac{s}{n}\|\mathbf{M}_{S \times S}\|_F^2, \quad (58)$$

which concludes the proof. □

When $n > s^2$ (which is the most realistic situation from our perspective), the previous observation is superseded by the following one.

**Proposition 10.** *The practical algorithm Algorithm* 2 *yields a head projection for* $\Sigma_{(s)}$ *with constant* $c_H = 1/\sqrt{s}$.

**Proof.** It is painless to see that, given the definition of Algorithm 2, for an arbitrary index set $S$ of size $s$,

$$\|\mathbf{M}_{S \times S}\|_F^2 = \sum_{j \in S} \|\mathbf{M}_{S \times j}\|_2^2 \leq \sum_{j \in S} \|\mathbf{M}_{S_j \times j}\|_2^2 \leq s\|\mathbf{M}_{S_{j_\star} \times j_\star}\|_2^2 \leq s\|\mathbf{M}_{S_{j_\star} \times S_{j_\star}}\|_F^2, \quad (59)$$

which concludes the proof. □

As a final remark, we show that head projections can be computed for specific symmetric matrices, e.g., matrices of rank one.

**Proposition 11.** *Given a symmetric matrix* $\mathbf{M} = \sum_{k=1}^r \mathbf{v}_k \mathbf{v}_k^\top \in \mathbb{R}^{n \times n}$ *of rank-r, we consider the practical algorithm that returns the matrix* $\mathbf{M}_{S_\star \times S_\star}$, *with* $S_\star := S_1 \cup \cdots \cup S_r$, *and* $S_k$ *the index set of* $s$ *largest absolute entries of* $\mathbf{v}_k$, $1 \leq k \leq r$. *This algorithm yields a head projection for* $\Sigma_{(s)}$ *into* $\Sigma_{(rs)}$ *with constant* $c_H = 1/\sqrt{r}$ *when applied to* $r$*-rank positive semidefinite matrices.*

**Proof.** Given the definition of $S_\star$, we are going to show that, for any index set $S$ of size $s$,

$$\|\mathbf{M}_{S_\star\times S_\star}\|_F \geq \frac{1}{\sqrt{r}}\|\mathbf{M}_{S\times S}\|_F. \tag{60}$$

To do so, we start by writing

$$M_{i,j}^2 = \left(\sum_{k=1}^r (\mathbf{v}_k)_i(\mathbf{v}_k)_j\right)^2 = \sum_{k,\ell=1}^r (\mathbf{v}_k)_i(\mathbf{v}_k)_j(\mathbf{v}_\ell)_i(\mathbf{v}_\ell)_j. \tag{61}$$

Then, for any index set $T$, in view of

$$\begin{aligned}\|\mathbf{M}_{T\times T}\|_F^2 &= \sum_{i,j\in T}\sum_{k,\ell=1}^r (\mathbf{v}_k)_i(\mathbf{v}_k)_j(\mathbf{v}_\ell)_i(\mathbf{v}_\ell)_j = \sum_{k,\ell=1}^r\sum_{i,j\in T} (\mathbf{v}_k)_i(\mathbf{v}_\ell)_i(\mathbf{v}_k)_j(\mathbf{v}_\ell)_j \\ &= \sum_{k,\ell=1}^r \left(\sum_{i\in T}(\mathbf{v}_k)_i(\mathbf{v}_\ell)_i\right)^2, \end{aligned} \tag{62}$$

we derive on the one hand that

$$\|\mathbf{M}_{T\times T}\|_F^2 \geq \sum_{k=1}^r \left(\sum_{i\in T}(\mathbf{v}_k)_i^2\right)^2 \tag{63}$$

and on the other hand, by the Cauchy–Schwarz inequality applied twice, that

$$\begin{aligned}\|\mathbf{M}_{T\times T}\|_F^2 &\leq \sum_{k,\ell=1}^r \left(\sum_{i\in T}(\mathbf{v}_k)_i^2\right)\left(\sum_{i\in T}(\mathbf{v}_\ell)_i^2\right) \\ &= \left(\sum_{k=1}^r\sum_{i\in T}(\mathbf{v}_k)_i^2\right)^2 \leq r\sum_{k=1}^r \left(\sum_{i\in T}(\mathbf{v}_k)_i^2\right)^2. \end{aligned} \tag{64}$$

Applying (64) with $T = S$ and using the defining property of each $S_k$ and of $S_\star$, we obtain

$$\begin{aligned}\|\mathbf{M}_{S\times S}\|_F^2 &\leq r\sum_{k=1}^r \left(\sum_{i\in S}(\mathbf{v}_k)_i^2\right)^2 \leq r\sum_{k=1}^r \left(\sum_{i\in S_k}(\mathbf{v}_k)_i^2\right)^2 \\ &\leq r\sum_{k=1}^r \left(\sum_{i\in S_\star}(\mathbf{v}_k)_i^2\right)^2 \leq r\|\mathbf{M}_{S_\star\times S_\star}\|_F^2, \end{aligned} \tag{65}$$

the last inequality being (63) applied with $T = S_\star$. The prospective inequality (60) is proved. □

### 5.3. *Joint low-rank and bisparsity structure*

Quickly stated, exact projections for $\Sigma_{(s)}^{[r]}$ are NP-hard, but there are computable tail projections for $\Sigma_{(s)}^{[r]}$. Head projections for $\Sigma_{(s)}^{[r]}$ are still NP-hard if they are forced to map exactly into $\Sigma_{(s)}^{[r]}$.

If they are allowed to map into a larger set $\Sigma_{(s')}^{[r']}$, the situation is not settled — this directly relates to Question 2.

We provide a few incomplete results related to this situation.

**Exact projections.** We already know from Sec. 5.2 that it is NP-hard to find the exact projection onto $\Sigma_{(s)}^{[r]}$ in general, since we are talking about exact projection onto $\Sigma_{(s)}$ when $r = n$. But we are more interested in the case where $r$ is a small constant, say $r = 1$ as a prototype. Then finding the exact projection onto $\Sigma_{(s)}^{[1]}$ amounts to solving the problem

$$\underset{|S|=s}{\text{maximize}} \, \|P^{[1]}(\mathbf{M}_{S\times S})\|_F = \underset{|S|=s}{\text{maximize}} \, \sigma_{\max}(\mathbf{M}_{S\times S}). \tag{66}$$

Thus, when $\mathbf{M}$ is a positive semidefinite matrix, we consider the problem

$$\underset{\|\mathbf{x}\|_0\leq s, \|\mathbf{x}\|_2=1}{\text{maximize}} \langle \mathbf{M}\mathbf{x}, \mathbf{x}\rangle. \tag{67}$$

This is the so-called sparse principal component analysis problem, which is NP-hard [20].

**Tail projections.** There is a fairly simple procedure to create a practical tail projection for $\Sigma_{(s)}^{[r]}$. It is based on the availability of tail projections for both $\Sigma^{[r]}$ and $\Sigma_{(s)}$. The argument is in fact valid for any two 'structures' $\Sigma'$ and $\Sigma''$ such that $\Sigma'$ is compatible with a tail projection $T''$ for $\Sigma''$, in the sense that

$$\mathbf{Z} \in \Sigma' \Rightarrow T''(\mathbf{Z}) \in \Sigma'. \tag{68}$$

The compatibility applies to the low-rank and bisparsity structures in two different ways: first, $\Sigma^{[r]}$ is compatible with the tail projection for $\Sigma_{(s)}$ given in Proposition 7, by virtue of the fact that a matrix $\mathbf{Z}$ of rank at most $r$ has all its submatrices $\mathbf{Z}_{S\times S}$ of rank at most $r$, too; second, $\Sigma_{(s)}$ is compatible with the exact projection for $\Sigma^{[r]}$, by virtue of the fact that a matrix $\mathbf{Z}$ supported on $S \times S$ has all its singular vectors supported on $S$, so that $P^{[r]}(\mathbf{Z})$ is supported on $S \times S$, too. Here is the abstract statement valid for arbitrary structures $\Sigma'$ and $\Sigma''$.

**Proposition 12.** *Let $T'$ and $T''$ be tail projections for $\Sigma'$ and $\Sigma''$ with constants $C_{T'}$ and $C_{T''}$. If $\Sigma'$ is compatible with $T''$, then $T'' \circ T'$ is a tail projection for $\Sigma' \cap \Sigma''$ with constant $C_{T'} + C_{T''} + C_{T'}C_{T''}$.*

**Proof.** We first remark that the compatibility condition ensures that $T'' \circ T'$ maps into $\Sigma' \cap \Sigma''$. Let $\mathbf{M} \in \mathbb{R}^{n\times n}$ and let $P(\mathbf{M})$ denote its exact projection for $\Sigma' \cap \Sigma''$.

The tail condition for $T'$ implies that

$$\|\mathbf{M} - T'(\mathbf{M})\|_F \leq C_{T'} \|\mathbf{M} - P(\mathbf{M})\|_F. \tag{69}$$

As a result, we obtain

$$\begin{aligned} \|T'(\mathbf{M}) - P(\mathbf{M})\|_F &\leq \|T'(\mathbf{M}) - \mathbf{M}\|_F + \|\mathbf{M} - P(\mathbf{M})\|_F \\ &\leq (C_{T'} + 1)\|\mathbf{M} - P(\mathbf{M})\|_F. \end{aligned} \tag{70}$$

The tail condition for $T''$ combined with (70) yields

$$\begin{aligned} \|T'(\mathbf{M}) - T''(T'(\mathbf{M}))\|_F &\leq C_{T''} \|T'(\mathbf{M}) - P(\mathbf{M})\|_F \\ &\leq C_{T''}(C_{T'} + 1)\|\mathbf{M} - P(\mathbf{M})\|_F. \end{aligned} \tag{71}$$

Using (69) and (71), we derive that

$$\begin{aligned} \|\mathbf{M} - T''(T'(\mathbf{M}))\|_F &\leq \|\mathbf{M} - T'(\mathbf{M})\|_F + \|T'(\mathbf{M}) - T''(T'(\mathbf{M}))\|_F \\ &\leq (C_{T'} + C_{T''}(C_{T'} + 1))\|\mathbf{M} - P(\mathbf{M})\|_F, \end{aligned} \tag{72}$$

which proves that $T'' \circ T'$ is a tail projection for $\Sigma' \cap \Sigma''$ with the desired constant. □

**Head projections.** The literature on the sparse principal component analysis problem informs us that finding a head projection for $\Sigma_{(s)}^{[r]}$ is still an NP-hard problem [20, Theorem 2]. In our setting, though, there is room to relax the head projection to map into $\Sigma_{(s')}^{[r']}$ with $r' > r$ and $s' > s$. In this regard, Question 2 asks if one can actually compute a head projection for $\Sigma_{(s)}^{[r]}$ into $\Sigma_{(s')}^{[r']}$ with $r' = Cr$. We do not have a definite answer for it, but we highlight an observation featuring a nonabsolute constant $c_H$, based on what was done for the bisparsity structure.

**Proposition 13.** *Given a symmetric matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$ and $r \leq s$, the practical algorithm that yields $P^{[r]}(H(\mathbf{M}))$ for the operator $H$ defined in Algorithm 2 is a head projection for $\Sigma_{(s)}^{[r]}$ with constant $c_H = \sqrt{r}/s$.*

**Proof.** Given a symmetric matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, we consider the row (or column) index set $S_\star$ of size $s$ supporting the nonzero rows (or columns) of $H(\mathbf{M}) \in \Sigma_{(s)}$ for the operator $H$ defined in Algorithm 2. By Proposition 10 for any index set $S$ of size $s$, we have

$$\|\mathbf{M}_{S_\star \times S_\star}\|_F^2 \geq \frac{1}{s}\|\mathbf{M}_{S \times S}\|_F^2. \tag{73}$$

Then, by noticing that the average of the $r$ largest squared singular values of $\mathbf{M}_{S_\star \times S_\star}$ is larger than the average of all the squared singular values of $\mathbf{M}_{S_\star \times S_\star}$, we derive

$$\|P^{[r]}(\mathbf{M}_{S_\star \times S_\star})\|_F^2 \geq \frac{r}{s}\|\mathbf{M}_{S_\star \times S_\star}\|_F^2 \geq \frac{r}{s^2}\|\mathbf{M}_{S \times S}\|_F^2 \geq \frac{r}{s^2}\|P^{[r]}(\mathbf{M}_{S \times S})\|_F^2. \tag{74}$$

The desired result is now proved. □

A similar argument, based on Proposition 9 instead of Proposition 10, would yield a head projection for $\Sigma^{[r]}_{(s)}$ into $\Sigma^{[r]}_{(2s)}$ with constant $c_H = \sqrt{r/n}$.

## 6. Sample Complexity with Rank-One Measurements

The specific (rank-one) measurements (5) do not result in a measurement map $\boldsymbol{\mathcal{A}} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ obeying the standard restricted isometry property (10). However, it will satisfy the following version featuring the $\ell_1$-norm as an inner norm. This was established in [3] when considering the low-rank structure alone. The proof sketch is deferred to the appendix. Note that the rank-one measurements (5) also satisfy a version of the null space property ensuring recovery via nuclear norm minimization, see [17, 18].

**Theorem 14.** *Suppose $\mathbf{a}_1, \dots, \mathbf{a}_m \in \mathbb{R}^m$ are independent vectors with independent $\mathcal{N}(0, 1/m)$ entries. Then, with failure probability at most $2\exp(-cm)$,*

$$\alpha\|\mathbf{Z}\|_F \le \|(\mathbf{a}_i^\top \mathbf{Z}\mathbf{a}_i)_{i=1}^m\|_1 \le \beta\|\mathbf{Z}\|_F \quad \text{for all } \mathbf{Z} \in \Sigma^{[r]}_{(s)}, \tag{75}$$

*provided $m \ge Crs\ln(en/s)$. The constants $\beta \ge \alpha > 0$ are absolute.*

The restricted isometry property (75) already guarantees that the specific-sample complexity — the theoretical one — is $m \asymp rs\ln(en/s)$, as expected. Indeed, given $\mathbf{y} = \boldsymbol{\mathcal{A}}(\mathbf{X}) + \mathbf{e}$ for some $\mathbf{X} \in \Sigma^{[r]}_{(s)}$, consider the unpractical recovery scheme

$$\Delta(\mathbf{y}) = \operatorname*{argmin}_{\mathbf{Z}\in\Sigma^{[r]}_{(s)}} \|\mathbf{y} - \boldsymbol{\mathcal{A}}(\mathbf{Z})\|_1. \tag{76}$$

In a similar spirit to (12)–(13), we can derive that

$$\|\mathbf{X} - \Delta(\boldsymbol{\mathcal{A}}(\mathbf{X}) + \mathbf{e})\|_F \le \frac{2}{\alpha}\|\mathbf{e}\|_1. \tag{77}$$

For a practical algorithm scheme, we have in mind an algorithm belonging to the iterative hard thresholding family. Namely, we can think of constructing a sequence $(\mathbf{X}_k)$ of matrices in $\Sigma^{[r']}_{(s')}$ by the recursion[b]

$$\mathbf{X}_{k+1} = T[\mathbf{X}_k + \nu_k H(\boldsymbol{\mathcal{A}}^*\operatorname{sgn}(\mathbf{y} - \boldsymbol{\mathcal{A}}\mathbf{X}_k))], \quad \nu_k = \frac{\|\mathbf{y} - \boldsymbol{\mathcal{A}}\mathbf{X}_k\|_1}{\beta^2}. \tag{78}$$

Here, the operators $T : \mathbb{R}^{n\times n} \to \Sigma^{[r']}_{(s')}$ and $H : \mathbb{R}^{n\times n} \to \Sigma^{[r'']}_{(s'')}$, depending on parameters $r'$, $s'$, $r''$, and $s''$, may be tail and head projections. It could also be useful to require the operator $T$ to satisfy the property[c] that, for all $\mathbf{X} \in \Sigma^{[r]}_{(s)}$ and

[b] It is 'natural' to include the sgn operator in order to exploit the restricted isometry property with $\ell_1$ inner norm.

[c] The inequality of (79) implies that $T$ is a tail projection with $C_T = 1 + \eta(C')$, since

$$\|\mathbf{M} - T(\mathbf{M})\|_F \le \|\mathbf{M} - P^{[r]}_{(s)}(\mathbf{M})\|_F + \|P^{[r]}_{(s)}(\mathbf{M}) - T(\mathbf{M})\|_F \le \|\mathbf{M} - P^{[r]}_{(s)}(\mathbf{M})\|_F$$

$$+ \eta(C')\|P^{[r]}_{(s)}(\mathbf{M}) - \mathbf{M}\|_F = C_T\|\mathbf{M} - P^{[r]}_{(s)}(\mathbf{M})\|_F.$$

all $\mathbf{Z} \in \mathbb{R}^{n \times n}$,

$$\|\mathbf{X} - T(\mathbf{Z})\|_F \le \eta(C)\|\mathbf{X} - \mathbf{Z}\|_F \quad \text{with } \eta(C') \underset{C' \to \infty}{\longrightarrow} 1. \tag{79}$$

With $T = P_{(s')}^{[r']}$, this inequality seems rather intuitive, but it needs to be formalized — keep in mind, however, that $P_{(s')}^{[r']}$ is not accessible. When considering the low-rank structure alone, such an inequality has been established and exploited in [12] to prove that an iterative hard thresholding algorithm of the type (78) presents the same recovery guarantees as nuclear norm minimization for recovery from measurements of type (5). The type of inequality (79) was first put forward for the sparse vector case in [26] and it has been exploited in [10] to propose and analyze an iterative hard thresholding algorithm designed for the case when the standard restricted isometry property fails.

There is an additional property that we could require about the operator $T$. Namely, given a matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, if $T(\mathbf{M})$ is supported on $S \times S$, then

$$T(\mathbf{M}) = T(\mathbf{M}_{S' \times S'}) \quad \text{whenever } S' \supseteq S. \tag{80}$$

This property is true (see Appendix A) for $T = P_{(s')}^{[r']}$, which again is inaccessible.

## Appendix A. Proofs of Auxiliary Results

This section collects the detailed arguments for some facts that have been stated but not proved in the narrative.

### A.1. Restricted isometry properties

First, let us concentrate on Theorem 2 and briefly justify that Gaussian measurements of type (4) satisfy the standard restricted isometry property (10). Without going into details, we simply mention that the classical proof consisting of a concentration inequality followed by a covering argument works — the key being to estimate the covering number of the 'ball' of $\Sigma_{(s)}^{[r]}$ essentially as in [4, Lemma 3.1] with the addition of a union bound.

Next, let us concentrate on Theorem 14 and briefly justify that Gaussian rank-one measurements of type (5) satisfy the modified restricted isometry property (75). Again, without going into details, we point out that the proof is in the spirit of [9]: for a fixed $\mathbf{Z} \in \mathbb{R}^{n \times n}$, establish a concentration inequality for $\|(\mathbf{a}_i^\top \mathbf{Z} \mathbf{a}_i)_{i=1}^m\|_1$ around its expectation $/\!/\mathbf{Z}/\!/$, prove that this slanted norm is equivalent to the Frobenius norm, and conclude with a covering argument.

### A.2. Convergence of the idealized iterative hard thresholding

We now establish that the naive (and impractical) iterative hard thresholding algorithm (15) allows for stable and robust recovery of jointly low-rank and bisparse matrices under the standard restricted isometry property. The precise statement appears after the important observation below.

**Lemma A.1.** *Suppose that* $\boldsymbol{\mathcal{A}} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ *satisfies the restricted isometry property* (10) *on* $\Sigma^{[2r]}_{(2s)}$ *with constant* $\delta \in (0,1)$. *Then, for all* $\mathbf{Z}, \mathbf{Z}' \in \Sigma^{[r]}_{(s)}$, *one has*

$$|\langle \mathbf{Z}, (\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}} - \mathbf{I})(\mathbf{Z}')\rangle| \le \delta \|\mathbf{Z}\|_F \|\mathbf{Z}'\|_F. \tag{A.1}$$

**Proof.** Assuming without loss of generality that $\|\mathbf{Z}\|_F = \|\mathbf{Z}'\|_F = 1$, we use in particular the parallelogram identity to write

$$\begin{aligned}
|\langle \mathbf{Z}, (\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}} - \mathbf{I})(\mathbf{Z}')\rangle| &= |\langle \boldsymbol{\mathcal{A}}(\mathbf{Z}), \boldsymbol{\mathcal{A}}(\mathbf{Z}')\rangle - \langle \mathbf{Z}, \mathbf{Z}'\rangle| \\
&= \left| \frac{1}{4}(\|\boldsymbol{\mathcal{A}}(\mathbf{Z}+\mathbf{Z}')\|_2^2 - \|\boldsymbol{\mathcal{A}}(\mathbf{Z}-\mathbf{Z}')\|_2^2) \right. \\
&\quad \left. - \frac{1}{4}(\|\mathbf{Z}+\mathbf{Z}'\|_F^2 - \|\mathbf{Z}-\mathbf{Z}'\|_F^2) \right| \\
&\le \frac{1}{4}|\|\boldsymbol{\mathcal{A}}(\mathbf{Z}+\mathbf{Z}')\|_2^2 - \|\mathbf{Z}+\mathbf{Z}'\|_F^2| \\
&\quad + \frac{1}{4}|\|\boldsymbol{\mathcal{A}}(\mathbf{Z}-\mathbf{Z}')\|_2^2 - \|\mathbf{Z}-\mathbf{Z}'\|_F^2| \\
&\le \frac{1}{4}\delta\|\mathbf{Z}+\mathbf{Z}'\|_F^2 + \frac{1}{4}\delta\|\mathbf{Z}-\mathbf{Z}'\|_F^2 \\
&= \frac{1}{4}\delta(2\|\mathbf{Z}\|_F^2 + 2\|\mathbf{Z}'\|_F^2) = \delta,
\end{aligned} \tag{A.2}$$

which is the required result. □

**Theorem A.2.** *If the restricted isometry property* (10) *holds on* $\Sigma^{[4r]}_{(4s)}$ *with constant* $\delta \in (0, 1/2)$, *then any* $\mathbf{X} \in \Sigma^{[r]}_{(s)}$ *is approximated from* $\mathbf{y} = \boldsymbol{\mathcal{A}}\mathbf{X} + \mathbf{e} \in \mathbb{R}^m$ *as a cluster point* $\mathbf{X}_\infty$ *of the sequence* $(\mathbf{X}_k)_{k\ge 0}$ *defined by*

$$\mathbf{X}_{k+1} = P^{[r]}_{(s)}(\mathbf{X}_k + \boldsymbol{\mathcal{A}}^*(\mathbf{y} - \boldsymbol{\mathcal{A}}\mathbf{X}_k)) \tag{A.3}$$

*with error*

$$\|\mathbf{X} - \mathbf{X}_\infty\|_F \le C\|\mathbf{e}\|_2. \tag{A.4}$$

**Proof.** It is enough to prove that, for all $k \ge 0$,

$$\|\mathbf{X} - \mathbf{X}_{k+1}\|_F \le \rho\|\mathbf{X} - \mathbf{X}_k\|_F + \tau\|\mathbf{e}\|_2, \quad \text{with } \rho := 2\delta < 1 \quad \text{and} \quad \tau > 0. \tag{A.5}$$

To start, notice that $\mathbf{X}_{k+1}$ better approximates $\mathbf{X}_k + \boldsymbol{\mathcal{A}}^*(\mathbf{y} - \boldsymbol{\mathcal{A}}\mathbf{X}_k) = \mathbf{X}_k + \boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{X} - \mathbf{X}_k) + \boldsymbol{\mathcal{A}}^*\mathbf{e}$ as an element from $\Sigma^{[r]}_{(s)}$ than $\mathbf{X}$ does, so that

$$\|\mathbf{X}_k + \boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{X} - \mathbf{X}_k) + \boldsymbol{\mathcal{A}}^*\mathbf{e} - \mathbf{X}_{k+1}\|_F^2 \le \|\mathbf{X}_k + \boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}}(\mathbf{X} - \mathbf{X}_k) + \boldsymbol{\mathcal{A}}^*\mathbf{e} - \mathbf{X}\|_F^2. \tag{A.6}$$

Introducing $\mathbf{X}$ in the left-hand side, expanding the squares, and simplifying leads to

$$\|\mathbf{X} - \mathbf{X}_{k+1}\|_F^2 \le -2\langle \mathbf{X} - \mathbf{X}_{k+1}, (\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}} - \mathbf{I})(\mathbf{X} - \mathbf{X}_k) + \boldsymbol{\mathcal{A}}^*\mathbf{e}\rangle. \tag{A.7}$$

Thanks to Lemma A.1, we have

$$|\langle \mathbf{X} - \mathbf{X}_{k+1}, (\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}} - \mathbf{I})(\mathbf{X} - \mathbf{X}_k)\rangle| \leq 2\delta \|\mathbf{X} - \mathbf{X}_{k+1}\|_F \|\mathbf{X} - \mathbf{X}_k\|_F, \quad \text{(A.8)}$$

while the restricted isometry property (10) also guarantees that

$$\begin{aligned} |\langle \mathbf{X} - \mathbf{X}_{k+1}, \boldsymbol{\mathcal{A}}^*\mathbf{e}\rangle| &= |\langle \boldsymbol{\mathcal{A}}(\mathbf{X} - \mathbf{X}_{k+1}), \mathbf{e}\rangle| \leq \|\boldsymbol{\mathcal{A}}(\mathbf{X} - \mathbf{X}_{k+1})\|_2 \|\mathbf{e}\|_2 \\ &\leq \sqrt{1+\delta}\|\mathbf{X} - \mathbf{X}_{k+1}\|_F \|\mathbf{e}\|_2. \end{aligned} \quad \text{(A.9)}$$

Therefore, using (A.8) and (A.9) in (A.7), we obtain

$$\|\mathbf{X} - \mathbf{X}_{k+1}\|_F^2 \leq 2\delta \|\mathbf{X} - \mathbf{X}_{k+1}\|_F \|\mathbf{X} - \mathbf{X}_k\|_F + \sqrt{1+\delta}\|\mathbf{X} - \mathbf{X}_{k+1}\|_F \|\mathbf{e}\|_2, \quad \text{(A.10)}$$

which clearly implies the required estimates (A.5) with $\tau = \sqrt{1+\delta}$ and (A.4) with $C = \tau/(1-\rho)$. □

**The exact projection for $\boldsymbol{\Sigma}_{(s)}^{[r]}$.** Here, we prove the statement (25) about the form of $P_{(s)}^{[r]}$ before justifying that property (80) holds for $T = P_{(s)}^{[r]}$.

**Proposition A.3.** *For $\mathbf{M} \in \mathbb{R}^{n\times n}$, the projection $P_{(s)}^{[r]}(\mathbf{M})$ of $\mathbf{M}$ onto $\Sigma_{(s)}^{[r]}$ has the form $P^{[r]}(\mathbf{M}_{S_\star \times S_\star})$, where $S_\star$ maximizes $\|P^{[r]}(\mathbf{M}_{S\times S})\|_F$ over all index sets $S$ of size $s$.*

**Proof.** Let us remark that, for any index set $T$,

$$\begin{aligned} \|\mathbf{M} - P^{[r]}(\mathbf{M}_{T\times T})\|_F^2 &= \|\mathbf{M}_{\overline{T\times T}} + \mathbf{M}_{T\times T} - P^{[r]}(\mathbf{M}_{T\times T})\|_F^2 \\ &= \|\mathbf{M}_{\overline{T\times T}}\|_F^2 + \|\mathbf{M}_{T\times T} - P^{[r]}(\mathbf{M}_{T\times T})\|_F^2 \\ &= \|\mathbf{M}_{\overline{T\times T}}\|_F^2 + \|\mathbf{M}_{T\times T}\|_F^2 - \|P^{[r]}(\mathbf{M}_{T\times T})\|_F^2 \\ &= \|\mathbf{M}\|_F^2 - \|P^{[r]}(\mathbf{M}_{T\times T})\|_F^2. \end{aligned} \quad \text{(A.11)}$$

Now, let $\mathbf{Z} \in \Sigma_{(s)}^{[r]}$ and consider an index set $S$ of size $s$ such that $\mathbf{Z}$ is supported on $S \times S$. The defining property of $S_\star$, together with (A.11), implies that

$$\begin{aligned} \|\mathbf{M} - P^{[r]}(\mathbf{M}_{S_\star\times S_\star})\|_F^2 &\leq \|\mathbf{M}\|_F^2 - \|P^{[r]}(\mathbf{M}_{S\times S})\|_F^2 \\ &= \|\mathbf{M}_{\overline{S\times S}}\|_F^2 + \|\mathbf{M}_{S\times S} - P^{[r]}(\mathbf{M}_{S\times S})\|_F^2 \\ &\leq \|\mathbf{M}_{\overline{S\times S}}\|_F^2 + \|\mathbf{M}_{S\times S} - \mathbf{Z}\|_F^2 = \|\mathbf{M} - \mathbf{Z}\|_F^2, \end{aligned} \quad \text{(A.12)}$$

where we have taken into account the facts that $P^{[r]}(\mathbf{M}_{S\times S})$ is the best $r$-rank approximation to $\mathbf{M}_{S\times S}$ and that $\mathbf{M}_{\overline{S\times S}}$ and $\mathbf{M}_{S\times S} - \mathbf{Z}$ are disjointly supported. Thus, we have proved that $\|\mathbf{M} - P^{[r]}(\mathbf{M}_{S_\star\times S_\star})\|_F \leq \|\mathbf{M} - \mathbf{Z}\|_F$ for all $\mathbf{Z} \in \Sigma_{(s)}^{[r]}$, which is the desired result. □

**Proposition A.4.** *For* $\mathbf{M} \in \mathbb{R}^{n\times n}$, *considering an index set* $S_\star$ *of size* $s$ *with* $P^{[r]}_{(s)}(\mathbf{M}) = P^{[r]}(\mathbf{M}_{S_\star\times S_\star})$, *one has*

$$P^{[r]}_{(s)}(\mathbf{M}) = P^{[r]}_{(s)}(\mathbf{M}_{S'\times S'}) \quad \textit{whenever } S' \supseteq S_\star. \tag{A.13}$$

**Proof.** According to Proposition A.3, it is enough to verify that, for any index set $S$ of size $s$,

$$\|P^{[r]}((\mathbf{M}_{S'\times S'})_{S_\star\times S_\star})\|_F \geq \|P^{[r]}((\mathbf{M}_{S'\times S'})_{S\times S})\|_F. \tag{A.14}$$

But this is true because $(\mathbf{M}_{S'\times S'})_{S_\star\times S_\star} = \mathbf{M}_{S_\star\times S_\star}$ and $(\mathbf{M}_{S'\times S'})_{S\times S} = (\mathbf{M}_{S\times S})_{S'\times S'}$, so that

$$\|P^{[r]}((\mathbf{M}_{S'\times S'})_{S\times S})\|_F \leq \|P^{[r]}(\mathbf{M}_{S\times S})\|_F \leq \|P^{[r]}(\mathbf{M}_{S_\star\times S_\star})\|_F, \tag{A.15}$$

where the last inequality follows from the defining property of $S_\star$. □

## Acknowledgments

S. F. is partially supported by NSF grants DMS-1622134 and DMS-1664803, and also acknowledges the NSF grant CCF-1934904. L. J. is funded by the FNRS, Belgium.

## References

[1] S. Bahmani and J. Romberg, Near-optimal estimation of simultaneously sparse and low-rank matrices from nested linear measurements, *Inf. Inference* **5**(3) (2016) 331–351.
[2] T. Blumensath, Sampling and reconstructing signals from a union of linear subspaces, *IEEE Trans. Inform. Theory* **57**(7) (2011) 4660–4671.
[3] T. Cai and A. Zhang, ROP: Matrix recovery via rank-one projections, *Ann. Statist.* **43**(1) (2015) 102–138.
[4] E. J. Candès and Y. Plan, Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements, *IEEE Trans. Inform. Theory* **57**(4) (2011) 2342–2359.
[5] E. J. Candès, T. Strohmer and V. Voroninski, Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming, *Commun. Pure Appl. Math.* **66**(8) (2013) 1241–1274.
[6] R. DeVore, P. Guergana and P. Wojtaszczyk, Approximation of functions of few variables in high dimensions, *Constr. Approx.* **33**(1) (2011) 125–143.
[7] M. Fornasier, J. Maly and V. Naumova, Sparse PCA from inaccurate and incomplete measurements, preprint (2018), arXiv:1801.06240.
[8] S. Foucart, Sampling Schemes and Recovery Algorithms for Functions of Few Coordinate Variables, preprint (2019).
[9] S. Foucart and M.-J. Lai, Sparse recovery with pre-Gaussian random matrices, *Studia Math.* **200**(1) (2010) 91–102.
[10] S. Foucart and G. Lecué, An IHT algorithm for sparse recovery from sub-exponential measurements, *IEEE Signal Process. Lett.* **24**(9) (2017) 1280–1283.
[11] S. Foucart and H. Rauhut, A Mathematical Introduction to Compressive Sensing, *Appl. Numer. Harmon. Analysis* (Birkhäuser, 2013).

[12] S. Foucart and S. Subramanian, Iterative hard thresholding for low-rank recovery from rank-one projections, *Linear Algebra Appl.* **572** (2019) 117–134.
[13] J. Geppert, F. Krahmer and D. Stöger, Sparse power factorization: Balancing peakiness and sample complexity, *Adv. Comput. Math.* **45**(3) (2019) 1711–1728.
[14] M. Golbabaee and M. E. Davies, Inexact gradient projection and fast data driven compressed sensing, *IEEE Trans. Inf. Theory* **64**(10) (2018) 6707–6721.
[15] C. Hegde, P. Indyk and L. Schmidt, Approximation algorithms for model-based compressive sensing, *IEEE Trans. Inform. Theory* **61**(9) (2015) 5129–5147.
[16] M. Iwen, A. Viswanathan and Y. Wang, Robust sparse phase retrieval made easy, *Appl. Comput. Harmon. Anal.* **42**(1) (2017) 135–142.
[17] M. Kabanava, R. Kueng, H. Rauhut and U. Terstiege, Stable low-rank matrix recovery via null space properties, *Inf. Inference* **5**(4) (2016) 405–441.
[18] R. Kueng, H. Rauhut and U. Terstiege, Low rank matrix recovery from rank one measurements, *Appl. Comput. Harmon. Anal.* **42**(1) (2017) 88–116.
[19] K. Lee, Y. Wu and Y. Bresler, Near-optimal compressed sensing of a class of sparse low-rank matrices via sparse power factorization, *IEEE Trans. Inf. Theory* **64**(3) (2017) 1666–1698.
[20] M. Magdon-Ismail, NP-hardness and inapproximability of sparse PCA, *Inform. Proc. Lett.* **126** (2017) 35–38.
[21] P. Manurangsi, Almost-polynomial ratio ETH-hardness of approximating densest $k$-subgraph, *Proc. 49th Annual ACM SIGACT Symp. Theory of Computing* (*STOC'17*) (ACM, Montreal, QC, Canada, 2017), pp. 954–961.
[22] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar and B. Hassibi, Simultaneously structured models with application to sparse and low-rank matrices, *IEEE Trans. Inform. Theory* **61**(5) (2015) 2886–2908.
[23] H. Rauhut, R. Schneider and Z. Stojanac, Low rank tensor recovery via iterative hard thresholding, *Linear Algebra Appl.* **523** (2018) 220–262.
[24] I. Roth, M. Kliesch, G. Wunder and J. Eisert, Reliable recovery of hierarchically sparse signals and application in machine-type communications, preprint (2016), arXiv:1612.07806.
[25] I. Roth, A. Flinth, R. Kueng, J. Eisert and G. Wunder, Hierarchical restricted isometry property for Kronecker product measurements, *56th Annual Allerton Conf. Communication, Control, and Computing* (IEEE, Monticello, IL, 2018), pp. 632–638.
[26] J. Shen and P. Li, A tight bound of hard thresholding, *JMLR* **18**(1) (2017) 7650–7691.