arXiv:1011.1839v1 [math.OC] 8 Nov 2010

Finding approximately rank-one submatrices with the nuclear norm and ℓ_1-norm*

Xuan Vinh Doan† Stephen Vavasis‡

November 2010

Abstract

We propose a convex optimization formulation with the nuclear norm and ℓ_1-norm to find a large approximately rank-one submatrix of a given nonnegative matrix. We develop optimality conditions for the formulation and characterize the properties of the optimal solutions. We establish conditions under which the optimal solution of the convex formulation has a specific sparse structure. Finally, we show that, under certain hypotheses, with high probability, the approach can recover the rank-one submatrix even when it is corrupted with random noise and inserted as a submatrix into a much larger random noise matrix.

1 Introduction

Given a nonnegative matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $a_{ij} \geq 0$ for all $i = 1, \ldots, m$, $j = 1, \ldots, n$, we consider the problem of finding $\mathcal{I} \subset \{1, \ldots, m\}$ and $\mathcal{J} \subset \{1, \ldots, n\}$ such that $\boldsymbol{A}(\mathcal{I}, \mathcal{J})$ is close to a rank-one matrix, and such that $\|\boldsymbol{A}(\mathcal{I}, \mathcal{J})\|$ is large. We shall call this problem the LAROS problem (for "large approximately rank-one submatrix").

The main application of the LAROS problem is for finding features in data. For example, suppose $\boldsymbol{A}$ represents a corpus of documents in some language. Each column of $\boldsymbol{A}$ is in correspondence with

*Supported in part by the U. S. Air Force Office of Scientific Research, a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada, and a grant from MITACS.

†Department of Combinatorics and Optimization, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada, vanxuan@uwaterloo.ca.

‡Department of Combinatorics and Optimization, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada, vavasis@uwaterloo.ca.

one document, and each row is in correspondence with a term used in the corpus. Here, "term" means a word in the language, excluding common words such as articles and prepositions. The (i, j) entry of $\boldsymbol{A}$ is the number of occurrences of term i in document j, perhaps normalized. Such a matrix is called the *term-document matrix* of the underlying corpus.

In this case, an approximately rank-one submatrix of $\boldsymbol{A}$ corresponds to a subset of terms and a subset of documents in which the selected terms occur with proportional frequencies in the selected documents. Such a submatrix may correspond to the intuitive notion of a topic that recurs in several documents, since a topic may manifest itself as a particular group of relevant terms that occur roughly in the same proportions.

As another example, the matrix $\boldsymbol{A}$ may correspond to a database of pixelated grayscale images, where each image has the same pixel size. Each column of $\boldsymbol{A}$ corresponds to one image, and each row to one pixel position. The (i, j) entry of $\boldsymbol{A}$ is the intensity of the ith pixel in the jth image. In this case, the approximately rank-one submatrix corresponds to a visual feature that recurs in a certain position in some subset of the images.

If one wanted to find more than one topic in a term-document matrix or more than one feature in an image database matrix, then one could iteratively find an approximately rank-one submatrix, subtract it from $\boldsymbol{A}$ (perhaps modifying the result of the subtraction to ensure that $\boldsymbol{A}$ remains nonnegative), and then repeat the procedure p times. Let the submatrices discovered be denoted $(\mathcal{I}_1, \mathcal{J}_1)$, $\ldots$, $(\mathcal{I}_p, \mathcal{J}_p)$. Suppose $A(\mathcal{I}_i, \mathcal{J}_i) \approx \boldsymbol{w}_i \boldsymbol{h}_i^T$ for $i = 1, \ldots, p$, and let $\bar{\boldsymbol{w}}_i$, $\bar{\boldsymbol{h}}_i$ denote the extension of $\boldsymbol{w}_i, \boldsymbol{h}_i$ to vectors of length m, n by inserting zeros for entries not in $\mathcal{I}_i, \mathcal{J}_i$ respectively.

It is known that if $\hat{\boldsymbol{A}}$ is a nonnegative matrix representing a submatrix of $\boldsymbol{A}$, then the minimizer $\boldsymbol{w}\boldsymbol{h}^T$ of $\|\hat{\boldsymbol{A}} - \boldsymbol{w}\boldsymbol{h}^T\|$ is the dominant singular vector pair in either the Frobenius or 2-norm (a consequence of the Eckart–Young theorem, Theorem 2.5.3 of [11]) and furthermore, $\boldsymbol{w} \geq \boldsymbol{0}$ and $\boldsymbol{h} \geq \boldsymbol{0}$ (a consequence of the Perron–Frobenius theorem.) Thus, without loss of generality, we may assume that each $\boldsymbol{w}_i \boldsymbol{h}_i^T$ determined by the iterative computation is nonnegative.

In this case, one has an approximate factorization

$$\boldsymbol{A} \approx [\bar{\boldsymbol{w}}_1, \ldots, \bar{\boldsymbol{w}}_p][\bar{\boldsymbol{h}}_1, \ldots, \bar{\boldsymbol{h}}_p]^T,$$

where we can write the right-hand side as $\boldsymbol{W}\boldsymbol{H}^T$ with $\boldsymbol{W} \geq \boldsymbol{0}$, $\boldsymbol{H} \geq \boldsymbol{0}$. This factorization is called a *nonnegative matrix factorization* of $\boldsymbol{A}$. The earliest reference known to us concerning nonnegative matrix factorization is Thomas' solution [18] to a problem posed by A. Berman and R. Plemmons (which, according to a remark in the journal, was also solved by A. Ben-Israel). Cohen and Rothblum

[8] describe applications for NMF in probability, quantum mechanics and other fields. Lee and Seung [13] showed that NMF can find features in image databases, and Hofmann [12] showed that probabilistic latent semantic analysis, a variant of NMF, can effectively cluster documents according to their topics.

Nonnegative matrix factorization is sometimes posed as an optimization problem: find $\boldsymbol{W} \in \mathbb{R}^{m\times p}$ and $\boldsymbol{H} \in \mathbb{R}^{n\times p}$, both nonnegative, such that $\|\boldsymbol{A} - \boldsymbol{W}\boldsymbol{H}^T\|$ is minimized in some matrix norm. It is known that this optimization problem is NP-hard [19]. Therefore, it is not surprising that most algorithms for the problem are heuristic in the sense that they do not make guarantees about the quality of the approximation.

One class of heuristic NMF algorithms are the 'greedy' algorithms [2, 3, 4, 10] that follow the framework described above. In a greedy algorithm, the columns of $\boldsymbol{W}$ and $\boldsymbol{H}$ are generated sequentially, with each new pair of columns accounting for one feature in the original $\boldsymbol{A}$. These greedy algorithms give rise to the LAROS subproblem addressed in this paper, namely, find one pair $\boldsymbol{w}_i, \boldsymbol{h}_i$ nonnegative such that $\boldsymbol{w}_i\boldsymbol{h}_i^T$, is a good approximation for a submatrix of $\boldsymbol{A}$ in the positions (i, j) where $\boldsymbol{A}$ is positive.

The LAROS subproblem, however, is itself NP-hard as observed by [10]. This is because the maximum-edge biclique problem can be naturally expressed as a rank-one submatrix problem. The biclique problem takes as input a bipartite graph $G = (U, V, E)$. The output is composed of two subsets $U^* \subset U$ and $V^* \subset V$ such that $U^* \times V^* \subset E$ (i.e., all possible $|U^*| \cdot |V^*|$ edges between U^* and V^* are present in G) and such that $|U^*| \cdot |V^*|$ is maximum with this property. This problem was shown by Peeters [15] to be NP-hard.

Maximum-edge biclique can be expressed as finding a large rank-one submatrix using the following construction. Let $\boldsymbol{A}$ be a $|U| \times |V|$ matrix with rows in correspondence to U and columns in correspondence to V. Entry (i, j) of $\boldsymbol{A}$ for $(i, j) \in U \times V$ is 1 if $(i, j) \in E$, else this entry is 0. Then a biclique corresponds exactly of a $|U^*| \cdot |V^*|$ submatrix of all 1's. A submatrix of all 1's is a rank-one matrix of norm $(|U^*| \cdot |V^*|)^{1/2}$, and there is no other kind of rank-one submatrix of $\boldsymbol{A}$.

We will provide a formal definition of the LAROS problem, i.e., exactly what is the desired output in Section 2. (Some authors mentioned earlier, e.g., [4] and [10] have provided other formal definitions.) We will also propose an convex optimization problem in Section 2 that, for matrices $\boldsymbol{A}$ constructed in a certain way, successfully finds large, approximately rank-one submatrices. We present two such theorems. One case is when the approximately rank-one submatrix dominates the rest of the matrix; this is presented in Section 3.

The second case is when $\boldsymbol{A}$ is constructed as follows:

$$\boldsymbol{A} = \boldsymbol{A}_0 + \boldsymbol{R},$$

where there are two index sets $\mathcal{I}, \mathcal{J}$ such that $\text{rank}(\boldsymbol{A_0}(\mathcal{I}, \mathcal{J})) = 1$, $\boldsymbol{A}_0(i,j) = 0$ for all $(i,j) \notin \mathcal{I} \times \mathcal{J}$, and $\boldsymbol{R}$ is a random matrix representing noise. In this case, under certain assumptions, our algorithm recovers $(\mathcal{I}, \mathcal{J})$ from $\boldsymbol{A}$ as proved in Section 4.

2 Matrix norm minimization

For reasons that will become clear, we start this section by presenting the convex relaxation of the LAROS problem, and only later will we present a nonconvex exact optimization formulation. In particular, we propose the following convex optimization problem that in some cases solves the LAROS problem:

$$\begin{aligned} \min \quad & \|\boldsymbol{X}\|_* + \theta \|\boldsymbol{X}\|_1 \\ \text{s.t.} \quad & \langle \boldsymbol{A}, \boldsymbol{X} \rangle \geq 1. \end{aligned} \tag{1}$$

The matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ is the unknown. Norm $\|\boldsymbol{X}\|_*$ is the *nuclear norm*, also called the *trace norm*; it is the sum of the singular values of $\boldsymbol{X}$. We use the notation $\|\boldsymbol{X}\|_1$ to mean the sum of the absolute values of entries of $\boldsymbol{X}$, that is, the ℓ_1^{mn}-norm applied to $\text{vec}(\boldsymbol{X})$, the concatenation of the columns of $\boldsymbol{X}$ into a long vector. Finally, $\langle \boldsymbol{A}, \boldsymbol{X} \rangle$ means the inner product of the two matrices. We note that an objective function involving a sum of the nuclear and ℓ_1-norms was used for a different purpose by [7]. Two other norms used extensively in this paper are $\|\boldsymbol{X}\|$, which is the spectral or 2-norm, i.e., $\sigma_1(\boldsymbol{X})$, and $\|\boldsymbol{X}\|_\infty$, which is the ℓ_∞^{mn}-norm applied to $\text{vec}(\boldsymbol{X})$, i.e., the maximum absolute entry of $\boldsymbol{X}$.

Before beginning a detailed analysis of this optimization problem, we first provide some motivation. Consider first the simplification obtained by taking $\theta = 0$:

$$\begin{aligned} \min \quad & \|\boldsymbol{X}\|_* \\ \text{s.t.} \quad & \langle \boldsymbol{A}, \boldsymbol{X} \rangle \geq 1. \end{aligned}$$

It follows from Proposition 1 below that the optimal solution is found using the singular value decomposition. In particular, if $\boldsymbol{A}$ is factored as $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ is orthogonal, $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times n}$ is diagonal, and $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ is orthogonal, then an optimizer is $\boldsymbol{X} = \boldsymbol{U}(:,1)\boldsymbol{V}(:,1)^T/\sigma_1$. Thus, when $\theta = 0$, the above formulation successfully finds the best rank-one approximation to the whole matrix A.

This approximation, however, is not always well suited for identifying submatrices. Consider e.g.

the following 6×6 matrix:

$$\boldsymbol{A} = \begin{pmatrix} 0.8 & 0.9 & 1.1 & 0.1 & 0.2 & 0.2 \\ 0.8 & 1.1 & 0.8 & 0 & 0 & 0 \\ 1.0 & 1.0 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 & 0.9 & 1.0 \\ 0 & 0 & 0 & 0.9 & 1.0 & 0.8 \\ 0 & 0 & 0 & 1.0 & 1.1 & 0.8 \end{pmatrix}$$

It is apparent from inspection that this matrix has two 3×3 approximately rank-one blocks in positions $\{1,2,3\} \times \{1,2,3\}$ and $\{4,5,6\} \times \{4,5,6\}$. If the 'noise' entries in the upper right $\{1,2,3\} \times \{4,5,6\}$ block were absent, then the two dominant singular vectors would exactly identify the two diagonal blocks. Once the noise entries are inserted, however, the dominant left singular vector of the above matrix $\boldsymbol{A}$ accurate to two decimal places is $[.45, .37, .37, .40, .40, .43]^T$. In other words, there is no separation at all between the rows numbered $1, 2, 3$ and those numbered $4, 5, 6$, so no submatrix is identified.

Armed with a preliminary understanding of the convex relaxed formulation, we now present and motivate an exact (nonconvex) formulation of LAROS, which is as follows.

$$\begin{aligned} \min \quad & \|\boldsymbol{X}\|_* + \theta\, |\mathcal{I}|\, |\mathcal{J}| \\ \text{s.t.} \quad & \langle \boldsymbol{A}, \boldsymbol{X} \rangle \geq 1, \\ & x_{ij} = 0, \qquad\qquad \forall\, (i,j) \notin \mathcal{I} \times \mathcal{J}, \end{aligned}$$

where $\mathcal{I} \subset \{1, \ldots, m\}$ and $\mathcal{J} \subset \{1, \ldots, n\}$ are unknowns (as well as $\boldsymbol{X}$). For fixed $\mathcal{I}$ and $\mathcal{J}$, the optimal solution would be the rank-one approximation of the submatrix $\boldsymbol{A}(\mathcal{I}, \mathcal{J})$ given by the SVD, as explained above, and the optimal value is $\|\boldsymbol{A}(\mathcal{I}, \mathcal{J})\|^{-1} + \theta\, |\mathcal{I}|\, |\mathcal{J}|$. The first term of the optimal value is $\|\boldsymbol{A}(\mathcal{I}, \mathcal{J})\|^{-1}$; therefore, for appropriate selection of θ, a large submatrix (in terms of 2-norm) will be selected. The second term is the size of the submatrix, which is a nonconvex function. Thus, the two terms balance the twin objectives of selecting a submatrix with a large first singular value and selecting a submatrix that has a relatively small number of entries.

This now motivates (1): we relax the above nonconvex formulation by replacing the cardinality term in the objective function with the ℓ_1-norm. The relaxed term $\theta \|\boldsymbol{X}\|_1$ in the objective function has the well-known effect of favoring sparser matrices $\boldsymbol{X}$ (those with fewer nonzero entries). Thus, the combination of the two terms seeks a low rank matrix with many entries equal to 0. For example, our formulation (1) applied to $\boldsymbol{A}$ above identifies the $\{4,5,6\} \times \{4,5,6\}$ submatrix when $\theta = 0.5$. In particular, the solution $\boldsymbol{X}$ has zeros in all positions except $\{4,5,6\} \times \{4,5,6\}$; in these positions it has positive entries ranging from 0.08 to 0.16.

To start our more formal analysis, let us consider the following general norm minimization problem.

$$\begin{aligned} \min \quad & |||\boldsymbol{X}||| \\ \text{s.t.} \quad & \langle \boldsymbol{A}, \boldsymbol{X} \rangle \geq 1, \end{aligned} \tag{2}$$

where $|||\cdot|||$ is an arbitrary norm function on $\mathbb{R}^{m\times n}$. (For example, the objective function $\|\boldsymbol{X}\|_* + \theta\|\boldsymbol{X}\|_1$ appearing in (1) is a norm.) Using the associated dual norm $|||\cdot|||^*$, we can relate Problem (2) to an equivalent problem as follows.

Lemma 1. *Consider $\boldsymbol{A} \neq \boldsymbol{0}$. Matrix $\boldsymbol{X}^*$ is an optimal solution of Problem (2) if and only if $\boldsymbol{Y}^* = |||\boldsymbol{A}|||^* \boldsymbol{X}^*$ is an optimal solution of the following problem:*

$$\begin{aligned} \max \quad & \langle \boldsymbol{A}, \boldsymbol{Y} \rangle \\ \text{s.t.} \quad & |||\boldsymbol{Y}||| \leq 1. \end{aligned} \tag{3}$$

Proof. Let $\boldsymbol{X}^*$ be an optimal solution of Problem (2). Clearly, $\boldsymbol{X}^* \neq \boldsymbol{0}$ and $\langle \boldsymbol{A}, \boldsymbol{X}^* \rangle = 1$. Apply the norm inequality, we have

$$|||\boldsymbol{X}^*||| \cdot |||\boldsymbol{A}|||^* \geq \langle \boldsymbol{A}, \boldsymbol{X}^* \rangle = 1 \Leftrightarrow |||\boldsymbol{X}^*||| \geq \frac{1}{|||\boldsymbol{A}|||^*}.$$

According to Boyd and Vandenberghe [5], the dual of the dual norm is the original norm and the norm inequality is tight: for any $\boldsymbol{A}$, there is always an $\boldsymbol{X} \neq \boldsymbol{0}$ such that the equality holds (for finite-dimensional vector spaces). Since $\boldsymbol{X}^*$ is an optimal solution of Problem (2),

$$|||\boldsymbol{X}^*||| = \frac{1}{|||\boldsymbol{A}|||^*}.$$

Let $\boldsymbol{Y}^* = |||\boldsymbol{A}|||^* \boldsymbol{X}^*$, we then have: $|||\boldsymbol{Y}^*||| = 1$ and $\langle \boldsymbol{A}, \boldsymbol{Y}^* \rangle = |||\boldsymbol{A}|||^*$. We also have:

$$\begin{aligned} |||\boldsymbol{A}|||^* = \max \quad & \langle \boldsymbol{A}, \boldsymbol{Y} \rangle \\ \text{s.t.} \quad & |||\boldsymbol{Y}||| \leq 1. \end{aligned}$$

Thus $\boldsymbol{Y}^*$ is indeed an optimal solution of Problem (3). Using similar arguments, we can prove that conversely, if $\boldsymbol{Y}^*$ is an optimal solution of Problem (3), then $\boldsymbol{X}^* = (|||\boldsymbol{A}|||^*)^{-1}\boldsymbol{Y}^*$ is an optimal solution of Problem (3). □

The next lemma characterizes the set of all optimal solutions of Problem (3).

Lemma 2. *The set of all optimal solutions of Problem (3) with $\boldsymbol{A} \neq \boldsymbol{0}$ is the subgradient of the dual norm function $|||\cdot|||^*$ at $\boldsymbol{A}$, $\partial|||\boldsymbol{A}|||^*$.*

Proof. Let $\boldsymbol{Y}^*$ be an optimal solution of Problem (3), we have $|||\boldsymbol{Y}^*||| = 1$ since $\boldsymbol{A} \neq \boldsymbol{0}$. Thus we have: $|||\boldsymbol{A}|||^* = \langle \boldsymbol{A}, \boldsymbol{Y}^* \rangle$. For an arbitrary matrix $\boldsymbol{B} \in \mathbb{R}^{m\times n}$,

$$|||\boldsymbol{A}+\boldsymbol{B}|||^* \geq \langle \boldsymbol{A}+\boldsymbol{B}, \boldsymbol{Y}^* \rangle = |||\boldsymbol{A}|||^* + \langle \boldsymbol{B}, \boldsymbol{Y}^* \rangle.$$

Thus $\boldsymbol{Y}^* \in \partial|||\boldsymbol{A}|||^*$.

Now consider $\boldsymbol{Y} \in \partial|||\boldsymbol{A}|||^*$:

$$|||\boldsymbol{B}|||^* \geq |||\boldsymbol{A}|||^* + \langle \boldsymbol{B}-\boldsymbol{A}, \boldsymbol{Y} \rangle \Leftrightarrow \langle \boldsymbol{A}, \boldsymbol{Y} \rangle - |||\boldsymbol{A}|||^* \geq \langle \boldsymbol{B}, \boldsymbol{Y} \rangle - |||\boldsymbol{B}|||^*, \quad \forall\, \boldsymbol{B} \in \mathbb{R}^{m\times n}.$$

With $\boldsymbol{B} = \boldsymbol{0}$ and $\boldsymbol{B} = 2\boldsymbol{A}$, we obtain the equality $\langle \boldsymbol{A}, \boldsymbol{Y} \rangle = |||\boldsymbol{A}|||^* > 0$. We have:

$$\langle \boldsymbol{A}, \boldsymbol{Y} \rangle = |||\boldsymbol{A}|||^* \leq |||\boldsymbol{A}|||^* |||\boldsymbol{Y}||| \Leftrightarrow (|||\boldsymbol{Y}||| - 1)|||\boldsymbol{A}|||^* \geq 0 \Rightarrow |||\boldsymbol{Y}|||_* \geq 1.$$

In addition, $\langle \boldsymbol{B}, \boldsymbol{Y} \rangle - |||\boldsymbol{B}|||^* \leq 0$ for all $\boldsymbol{B} \in \mathbb{R}^{m\times n}$. The norm inequality $\langle \boldsymbol{B}, \boldsymbol{Y} \rangle \leq |||\boldsymbol{B}|||^* |||\boldsymbol{Y}|||$ is tight and $\boldsymbol{Y} \neq \boldsymbol{0}$ ($|||\boldsymbol{Y}||| \geq 1$); therefore, there exists $\boldsymbol{B} \neq \boldsymbol{0}$ such that $\langle \boldsymbol{B}, \boldsymbol{Y} \rangle = |||\boldsymbol{B}|||^* |||\boldsymbol{Y}|||$. Thus we have:

$$|||\boldsymbol{B}|||^* |||\boldsymbol{Y}||| - |||\boldsymbol{B}|||^* \leq 0 \Leftrightarrow (|||\boldsymbol{Y}||| - 1)|||\boldsymbol{B}|||^* \leq 0 \Rightarrow |||\boldsymbol{Y}||| \leq 1.$$

Thus $|||\boldsymbol{Y}||| = 1$ and $\langle \boldsymbol{A}, \boldsymbol{Y} \rangle = |||\boldsymbol{A}|||^*$, the optimal value of Problem (3), which means $\boldsymbol{Y}$ is an optimal solution of Problem (3). □

Lemma 1 and 2 show that the set of all optimal solutions of Problem (2) is $(|||\boldsymbol{A}|||^*)^{-1}\partial|||\boldsymbol{A}|||^*$. The uniqueness of the optimal solution of Problem (2) is equivalent to the differentiability of the dual norm function $|||\cdot|||^*$ at $\boldsymbol{A}$. These results are summarized in the following theorem.

Theorem 1. *Consider $\boldsymbol{A} \neq \boldsymbol{0}$. The following statements are true:*

(i) The set of optimal solutions of Problem (2) is $(|||\boldsymbol{A}|||^*)^{-1}\partial|||\boldsymbol{A}|||^*$.

(ii) Problem (2) has a unique optimal solution if and only if the dual norm function $|||\cdot|||^*$ is differentiable at $\boldsymbol{A}$.

If the norm is set to be the nuclear norm, we obtain the following minimization problem, which has been used [9, 16, 6, 1] as a relaxation of rank minimization optimization problems:

$$\begin{aligned} \min \quad & \|\boldsymbol{X}\|_* \\ \text{s.t.} \quad & \langle \boldsymbol{A}, \boldsymbol{X} \rangle \geq 1. \end{aligned} \tag{4}$$

The dual norm of the nuclear norm is the spectral norm. According to Ziętak [20], if $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ is a singular value decomposition of $\boldsymbol{A}$ and s is the multiplicity of the largest singular value of $\boldsymbol{A}$, the subgradient $\partial \|\boldsymbol{A}\|$ is written as follows:

$$\partial \|\boldsymbol{A}\| = \left\{ \boldsymbol{U} \begin{bmatrix} \boldsymbol{S} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{V}^T : \boldsymbol{S} \in \mathcal{S}_+^s, \|\boldsymbol{S}\|_* = 1 \right\}.$$

Clearly, the largest rank-one approximation of $\boldsymbol{A}$, $\boldsymbol{u}_1\boldsymbol{v}_1^T$, always belongs to the subgradient $\partial \|\boldsymbol{A}\|$. The description of the subgradient $\partial \|\boldsymbol{A}\|$ shows that the maximum possible rank of an optimal solution of Problem (4) is the multiplicity of the largest singular value of $\boldsymbol{A}$. In addition, the spectral norm function $\|\cdot\|$ is not differentiable in general. The uniqueness of the optimal solution of Problem (4) is equivalent to the differentiability of the spectral norm function $\|\cdot\|$ at $\boldsymbol{A}$. The necessary and sufficient condition is $s = 1$ or equivalently, $\sigma_1(\boldsymbol{A}) > \sigma_2(\boldsymbol{A})$. In the case of unique optimal solution, we obtain the largest rank-one approximation of $\boldsymbol{A}$ (up to the scaling factor $\|\boldsymbol{A}\|^{-1}$). These results are stated in the following proposition:

Proposition 1. *Consider $\boldsymbol{A} \neq \boldsymbol{0}$. The following statements are true:*

(i) The set of optimal solutions of Problem (4) is $\|\boldsymbol{A}\|^{-1}\,\partial \|\boldsymbol{A}\|$.

(ii) The largest rank-one approximation of $\boldsymbol{A}$ is an optimal solution of Problem (4) and it is the unique solution if and only if $\sigma_1(\boldsymbol{A}) > \sigma_2(\boldsymbol{A})$.

Similar to low-rank minimization problems with nuclear norm approximation, sparse optimization problems can be approximately handled by the (vector) ℓ_1-norm function $\|\cdot\|_1$. Let us consider the following problem

$$\begin{aligned} \min \quad & \|\boldsymbol{X}\|_1 \\ \text{s.t.} \quad & \langle \boldsymbol{A}, \boldsymbol{X} \rangle \geq 1. \end{aligned} \tag{5}$$

The dual norm of ℓ_1-norm is the (vector) infinity norm $\|\cdot\|_\infty$, i.e., the maximum absolute entry of the matrix, and the subgradient $\partial \|\boldsymbol{A}\|_\infty$ can be written as follows,

$$\partial \|\boldsymbol{A}\|_\infty = \operatorname{conv} \left\{ \operatorname{sgn}(a_{ij})\boldsymbol{E}_{ij} \mid (i,j) \in \arg\max_{(k,l)} |a_{kl}| \right\},$$

where $\boldsymbol{E}_{ij}$ is the unit matrix in $\mathbb{R}^{m\times n}$ with $\boldsymbol{E}_{ij}(i,j) = 1$. The sparsity of the optimal solution $\boldsymbol{X}^*$ of Problem (5) is clearly related to the multiplicity of the maximum absolute value of elements of $\boldsymbol{A}$. Applying Theorem 1 for this particular ℓ_1-norm, we obtain the following results:

Proposition 2. *Consider $\boldsymbol{A} \neq \boldsymbol{0}$. The following statements are true:*

(i) The set of optimal solutions of Problem (5) is $\|\boldsymbol{A}\|_\infty^{-1} \partial \|\boldsymbol{A}\|_\infty$.

(ii) The matrix $\operatorname{sgn}(a_{ij})\boldsymbol{E}_{ij}$, where $(i,j) \in \arg\max_{(k,l)} |a_{kl}|$, and $\operatorname{sgn}(\cdot)$ is the usual sign function, is an optimal solution of Problem (5) and it is the unique solution if and only if $|a_{ij}| > |a_{kl}|$ for all $(k,l) \neq (i,j)$.

As mentioned above, finding a low-rank submatrix clearly involves both low-rank and sparse optimization (with a specific sparse structure). Let us return to the parametric optimization problem (1) proposed at the beginning of this section

$$\begin{aligned} \min \quad & \|\boldsymbol{X}\|_* + \theta \|\boldsymbol{X}\|_1 \\ \text{s.t.} \quad & \langle \boldsymbol{A}, \boldsymbol{X} \rangle \geq 1, \end{aligned}$$

where $\theta \geq 0$. Clearly, if $\theta = 0$, we obtain Problem (4) and when $\theta \to \infty$, we approach Problem (5). This optimization problem clearly addresses both low-rank and sparse requirements of the solution $\boldsymbol{X}$. We now would like to characterize the set of optimal solutions of the problem.

The objective function $\|\boldsymbol{X}\|_* + \theta \|\boldsymbol{X}\|_1$ is a norm function since $\theta \geq 0$. Denote $\|\boldsymbol{X}\|_\theta$ to be this parametric norm of $\boldsymbol{X}$,

$$\|\boldsymbol{X}\|_\theta := \|\boldsymbol{X}\|_* + \theta \|\boldsymbol{X}\|_1$$

and consider its dual norm function $\|\cdot\|_\theta^*$. Clearly, Problem (1) is a special case of Problem (2). The set of optimal solutions of Problem (1) can therefore be characterized as follows:

Proposition 3. *Consider $\boldsymbol{A} \neq \boldsymbol{0}$. The following statements are true:*

(i) The set of optimal solutions of Problem (1) is $(\|\boldsymbol{A}\|_\theta^*)^{-1} \partial \|\boldsymbol{A}\|_\theta^*$.

(ii) There is a unique solution if and only if the dual norm function $\|\cdot\|_\theta^*$ is differentiable at $\boldsymbol{A}$.

We now focus on deriving some properties of the dual norm $\|\cdot\|_\theta^*$. We have:

$$\begin{aligned} \|\boldsymbol{A}\|_\theta^* = \max \quad & \langle \boldsymbol{A}, \boldsymbol{X} \rangle \\ \text{s.t.} \quad & \|\boldsymbol{X}\|_\theta \leq 1. \end{aligned} \tag{6}$$

We will use the gauge function and its dual polar function (see Rockafellar [17] for more details) to compute this dual norm.

Proposition 4. *The dual norm $\|\boldsymbol{A}\|_\theta^*$ with $\theta > 0$ is the optimal value of the following optimization problem:*

$$\begin{aligned} \|\boldsymbol{A}\|_\theta^* = \min \quad & \max\left\{\|\boldsymbol{Y}\|, \theta^{-1}\|\boldsymbol{Z}\|_\infty\right\} \\ \text{s.t.} \quad & \boldsymbol{Y} + \boldsymbol{Z} = \boldsymbol{A}. \end{aligned} \tag{7}$$

Proof. Consider the closed unit ball $\mathcal{C}_* = \{\boldsymbol{X} \in \mathbb{R}^{m\times n} \mid \|\boldsymbol{X}\|_* \leq 1\}$ with respect to the nuclear norm and similarly, the unit ball $\mathcal{C}_1$ with respect to the ℓ_1-norm $\|\cdot\|_1$. We have the polar of $\mathcal{C}_*$ is the closed unit ball with respect to the spectral norm, $\mathcal{C}_*^\circ = \mathcal{C}$. Similarly, we have: $\mathcal{C}_1^\circ = \mathcal{C}_\infty$, the unit ball with respect to the infinity norm.

Using the definition of gauge functions, we have: $\|\boldsymbol{X}\|_* = \gamma_{\mathcal{C}_*}(\boldsymbol{X}) = \min\{\lambda \geq 0 \mid \boldsymbol{X} \in \lambda\mathcal{C}_*\}$. In addition, the support function $\sigma_{\mathcal{S}}(\boldsymbol{X}) = \max\{\langle \boldsymbol{X}, \boldsymbol{Y}\rangle \mid \boldsymbol{Y} \in \mathcal{S}\}$ is the gauge function of $\mathcal{S}^\circ$ for all symmetric closed bounded convex set with $\mathbf{0} \in \operatorname{int}(\mathcal{S})$. All unit balls satisfy these conditions; therefore, we obtain the well-known results $\|\boldsymbol{X}\|_* = \gamma_{\mathcal{C}_*}(\boldsymbol{X}) = \sigma_{\mathcal{C}_*^\circ}(\boldsymbol{X}) = \sigma_{\mathcal{C}}(\boldsymbol{X})$ and $\|\boldsymbol{X}\|_1 = \gamma_{\mathcal{C}_1}(\boldsymbol{X}) = \sigma_{\mathcal{C}_1^\circ}(\boldsymbol{X}) = \sigma_{\mathcal{C}_\infty}(\boldsymbol{X})$.

Now consider the unit ball $\mathcal{C}_\theta = \{\boldsymbol{X} \in \mathbb{R}^{m\times n} \mid \|\boldsymbol{X}\|_* + \theta\|\boldsymbol{X}\|_1 \leq 1\}$, we have:

$$\mathcal{C}_\theta = \{\boldsymbol{X} \in \mathbb{R}^{m\times n} \mid \sigma_{\mathcal{C}}(\boldsymbol{X}) + \theta\sigma_{\mathcal{C}_\infty}(\boldsymbol{X}) \leq 1\}.$$

Applying the definition of support functions, we have: $\sigma_{\mathcal{C}}(\boldsymbol{X}) + \theta\sigma_{\mathcal{C}_\infty}(\boldsymbol{X}) = \sigma_{\mathcal{C}+\theta\mathcal{C}_\infty}(\boldsymbol{X})$, where $\mathcal{C} + \theta\mathcal{C}_\infty$ is the Minkowski sum of two sets, $\mathcal{C}$ and $\theta\mathcal{C}_\infty$. This set satisfies all the conditions above; therefore, $\sigma_{\mathcal{C}+\theta\mathcal{C}_\infty}(\boldsymbol{X}) = \gamma_{(\mathcal{C}+\theta\mathcal{C}_\infty)^\circ}(\boldsymbol{X})$. Thus

$$\mathcal{C}_\theta = \{\boldsymbol{X} \in \mathbb{R}^{m\times n} \mid \gamma_{(\mathcal{C}+\theta\mathcal{C}_\infty)^\circ}(\boldsymbol{X}) \leq 1\} = (\mathcal{C} + \theta\mathcal{C}_\infty)^\circ.$$

We also have: $\|\boldsymbol{A}\|_\theta^* = \sigma_{\mathcal{C}_\theta}(\boldsymbol{A})$. Thus

$$\|\boldsymbol{A}\|_\theta^* = \gamma_{\mathcal{C}_\theta^\circ}(\boldsymbol{A}) = \gamma_{\mathcal{C}+\theta\mathcal{C}_\infty}(\boldsymbol{A}).$$

We have: $\gamma_{\mathcal{C}+\theta\mathcal{C}_\infty}(\boldsymbol{A}) = \min\{\lambda \geq 0 \mid \boldsymbol{A} \in \lambda(\mathcal{C} + \theta\mathcal{C}_\infty)\}$ or equivalently,

$$\begin{aligned} \gamma_{\mathcal{C}+\theta\mathcal{C}_\infty}(\boldsymbol{A}) = \min \quad & \lambda \\ \text{s.t.} \quad & \boldsymbol{A} = \boldsymbol{Y} + \boldsymbol{Z}, \\ & \|\boldsymbol{Y}\| \leq \lambda, \\ & \theta^{-1}\|\boldsymbol{Z}\|_\infty \leq \lambda, \\ & \lambda \geq 0. \end{aligned}$$

Rewriting the minimization problem above, we obtain the final result as shown in (7):

$$\begin{aligned} \|\boldsymbol{A}\|_\theta^* = \min \quad & \max\left\{\|\boldsymbol{Y}\|, \theta^{-1}\|\boldsymbol{Z}\|_\infty\right\} \\ \text{s.t.} \quad & \boldsymbol{Y}+\boldsymbol{Z}=\boldsymbol{A}. \end{aligned}$$

□

We can now derive the optimality conditions for both problems (6) and (7):

Lemma 3. *Nonzero feasible solutions $\boldsymbol{X}$ and $(\boldsymbol{Y},\boldsymbol{Z})$ are optimal for Problem (6) and (7) respectively if and only if they satisfy the conditions below:*

(i) $\|\boldsymbol{Y}\| = \theta^{-1}\|\boldsymbol{Z}\|_\infty$,

(ii) $\boldsymbol{X} \in \alpha\partial\|\boldsymbol{Y}\|$, $\alpha \geq 0$,

(iii) $\boldsymbol{X} \in \beta\partial\|\boldsymbol{Z}\|_\infty$, $\beta \geq 0$, *and*

(iv) $\alpha + \theta\beta = 1$.

Proof. We first prove the weak duality result. Consider feasible solutions $\boldsymbol{X}$ and $(\boldsymbol{Y},\boldsymbol{Z})$ for Problem (6) and (7) respectively, we have:

$$\begin{aligned} \langle \boldsymbol{X},\boldsymbol{A}\rangle &= \langle \boldsymbol{X},\boldsymbol{Y}\rangle + \langle \boldsymbol{X},\boldsymbol{Z}\rangle \\ &\leq \|\boldsymbol{X}\|_*\|\boldsymbol{Y}\| + \|\boldsymbol{X}\|_1\|\boldsymbol{Z}\|_\infty \\ &\leq \|\boldsymbol{X}\|_*\max\{\|\boldsymbol{Y}\|,\theta^{-1}\|\boldsymbol{Z}\|_\infty\} + \theta\|\boldsymbol{X}\|_1\max\{\|\boldsymbol{Y}\|,\theta^{-1}\|\boldsymbol{Z}\|_\infty\} \\ &= (\|\boldsymbol{X}\|_* + \theta\|\boldsymbol{X}\|_1)\max\{\|\boldsymbol{Y}\|,\theta^{-1}\|\boldsymbol{Z}\|_\infty\} \\ &\leq \max\{\|\boldsymbol{Y}\|,\theta^{-1}\|\boldsymbol{Z}\|_\infty\}. \end{aligned}$$

The strong duality result shows that $\boldsymbol{X}$ and $(\boldsymbol{Y},\boldsymbol{Z})$ are the optimal solutions if and only if $\langle \boldsymbol{X},\boldsymbol{A}\rangle = \max\{\|\boldsymbol{Y}\|,\theta^{-1}\|\boldsymbol{Z}\|_\infty\}$. This happens if and only if all the conditions below are satisfied:

(i) $\langle \boldsymbol{X},\boldsymbol{Y}\rangle = \|\boldsymbol{X}\|_*\|\boldsymbol{Y}\|$ and $\langle \boldsymbol{X},\boldsymbol{Z}\rangle = \|\boldsymbol{X}\|_1\|\boldsymbol{Z}\|_\infty$,

(ii) $\|\boldsymbol{Y}\| = \max\{\|\boldsymbol{Y}\|,\theta^{-1}\|\boldsymbol{Z}\|_\infty\} = \theta^{-1}\|\boldsymbol{Z}\|_\infty$, and

(iii) $\|\boldsymbol{X}\|_* + \theta\|\boldsymbol{X}\|_1 = 1$.

The first two conditions are equivalent to the fact that $\boldsymbol{X} = \alpha\partial\|\boldsymbol{Y}\|$, where $\alpha = \|\boldsymbol{X}\|_*$, and $\boldsymbol{X} = \beta\partial\|\boldsymbol{Z}\|_\infty$, where $\beta = \|\boldsymbol{X}\|_1$. The second condition is simply $\|\boldsymbol{Y}\| = \theta^{-1}\|\boldsymbol{Z}\|_\infty$ and the third condition is equivalent to $\alpha + \theta\beta = 1$. Thus we have proved the necessary and sufficient optimality conditions for Problem (6) and (7). □

Using these optimality conditions, we can obtain simple sufficient conditions for the uniqueness of the optimal solution $\boldsymbol{X}$:

Proposition 5. *Consider the feasible solution $\boldsymbol{X}$ of Problem (6). If there exists $(\boldsymbol{Y}, \boldsymbol{Z})$ that satisfies the conditions below,*

(i) $\boldsymbol{Y} + \boldsymbol{Z} = \boldsymbol{A}$ *and* $\|\boldsymbol{Y}\| = \theta^{-1} \|\boldsymbol{Z}\|_\infty$,

(ii) $\boldsymbol{X} \in \alpha \partial \|\boldsymbol{Y}\|$, $\alpha \geq 0$,

(iii) $\boldsymbol{X} \in \beta \partial \|\boldsymbol{Z}\|_\infty$, $\beta \geq 0$,

(iv) $\alpha + \theta\beta = 1$, *and*

(v) $\|\cdot\|$ *is differentiable at* $\boldsymbol{Y}$ *or* $\|\cdot\|_\infty$ *is differentiable at* $\boldsymbol{Z}$,

then $\boldsymbol{X}$ is the unique optimal solution of Problem (6).

Proof. Using the first four conditions, we can prove that $\boldsymbol{X}$ is an optimal solution of Problem (6) and $(\boldsymbol{Y}, \boldsymbol{Z})$ is an optimal solution of Problem (7). Now assume that $\|\cdot\|$ is differentiable at $\boldsymbol{Y}$, we have: $\partial \|\boldsymbol{Y}\|$ is a singleton, $\partial \|\boldsymbol{Y}\| = \{\boldsymbol{V}\}$. Thus we have:

$$\|\boldsymbol{X}\|_1 = \alpha \|\boldsymbol{V}\|_1 = \beta \Rightarrow \alpha(1 + \theta \|\boldsymbol{V}\|_1) = 0 \Rightarrow \alpha = \frac{1}{1 + \theta \|\boldsymbol{V}\|_1}.$$

Assume there is another optimal solution $\bar{\boldsymbol{X}} \neq \boldsymbol{X}$ of Problem (6). Applying Lemma 3, we will have $\bar{\boldsymbol{X}} \in \bar{\alpha} \partial \|\boldsymbol{Y}\|$ and similarly $\bar{\boldsymbol{X}} \in \bar{\beta} \partial \|\boldsymbol{Z}\|_\infty$ with $\bar{\alpha} + \theta\bar{\beta} = 1$. Same calculation results in $\bar{\alpha} = \alpha$ (contradiction). Thus $\boldsymbol{X}$ is the unique optimal solution of Problem (6). Similar arguments can be used to prove the uniqueness of $\boldsymbol{X}$ if $\|\cdot\|_\infty$ is differentiable at $\boldsymbol{Z}$. □

Proposition 5 relies on dual solutions $\boldsymbol{Y}$ and $\boldsymbol{Z}$ to show the uniqueness of the primal solution $\boldsymbol{X}$. Next, we will focus on the low-rank and sparse property of the optimal solution $\boldsymbol{X}$ for different values of θ. The following theorem provides the sufficient conditions on matrix $\boldsymbol{A}$ for the rank-one property (and uniqueness) of the optimal solution $\boldsymbol{X}$ when θ is small enough.

Theorem 2. *If $\boldsymbol{A}$ satisfies the condition $\sigma_1(\boldsymbol{A}) > \sigma_2(\boldsymbol{A})$, then Problem (6) has a (unique) rank-one optimal solution $\boldsymbol{X}$ for all $0 \leq \theta < \theta_A$, where $\theta_A = \frac{1}{\sqrt{mn}} \left(\frac{\sigma_1(\boldsymbol{A}) - \sigma_2(\boldsymbol{A})}{3\sigma_1(\boldsymbol{A}) - \sigma_2(\boldsymbol{A})} \right)$.*

Proof. The optimality conditions in Lemma 3 show that there exist $\boldsymbol{Y}$ and $\boldsymbol{Z}$ such that $\boldsymbol{A} = \boldsymbol{Y} + \boldsymbol{Z}$, $\|\boldsymbol{Z}\|_\infty = \theta \|\boldsymbol{Y}\|$ and $\boldsymbol{X} \in \alpha \partial \|\boldsymbol{Y}\|$. Applying a standard perturbation theorem of singular values (see Cor. 8.6.2 of [11]), we have:

$$|\sigma_i(\boldsymbol{A}) - \sigma_i(\boldsymbol{Y})| \leq \|\boldsymbol{Z}\|, \quad i = 1, 2.$$

We also have: $\|\boldsymbol{Z}\| \leq \sqrt{mn}\,\|\boldsymbol{Z}\|_\infty$. Thus

$$\|\boldsymbol{Z}\| \leq \sqrt{mn}\,(\theta\,\|\boldsymbol{Y}\|) = \sqrt{mn}\,(\theta\sigma_1(\boldsymbol{Y}))\,.$$

For all $0 \leq \theta < \theta_A$, we have:

$$\sigma_1(\boldsymbol{Y}) \leq \sigma_1(\boldsymbol{A}) + \|\boldsymbol{Z}\| \leq \sigma_1(\boldsymbol{A}) + \sqrt{mn}\,(\theta\sigma_1(\boldsymbol{Y})) < \sigma_1(\boldsymbol{A}) + \sqrt{mn}\,(\theta_A\sigma_1(\boldsymbol{Y}))\,.$$

This implies

$$(1-\theta_A\sqrt{mn})\sigma_1(\boldsymbol{Y}) < \sigma_1(\boldsymbol{A}) \Leftrightarrow \frac{2\sigma_1(\boldsymbol{A})}{3\sigma_1(\boldsymbol{A})-\sigma_2(\boldsymbol{A})}\sigma_1(\boldsymbol{Y}) < \sigma_1(\boldsymbol{A}) \Leftrightarrow \sigma_1(\boldsymbol{Y}) < \frac{1}{2}(3\sigma_1(\boldsymbol{A})-\sigma_2(\boldsymbol{A})).$$

We then have:

$$\|\boldsymbol{Z}\| \leq \sqrt{mn}\,(\theta\sigma_1(\boldsymbol{Y})) < \sqrt{mn}\,(\theta_A\sigma_1(\boldsymbol{Y})) < \frac{1}{2}(\sigma_1(\boldsymbol{A})-\sigma_2(\boldsymbol{A})).$$

Thus

$$\sigma_1(\boldsymbol{Y}) \geq \sigma_1(\boldsymbol{A}) - \|\boldsymbol{Z}\| > \frac{1}{2}(\sigma_1(\boldsymbol{A})+\sigma_2(\boldsymbol{A})) > \sigma_2(\boldsymbol{A}) + \|\boldsymbol{Z}\| \geq \sigma_2(\boldsymbol{Y}).$$

We have $\sigma_1(\boldsymbol{Y}) > \sigma_2(\boldsymbol{Y})$; therefore, $\|\cdot\|$ is differentiable at $\boldsymbol{Y}$. According to Proposition 5, we have $\boldsymbol{X}$ is the unique rank-one optimal solution of Problem (6). □

The last result of this section concerns the nonnegativity of $\boldsymbol{X}$. If $\boldsymbol{A}$ is nonnegative, then one might expect $\boldsymbol{X}$ to be nonnegative. For $\theta = 0$ or $\theta = \infty$, this is certainly true by preceding results in this section. It is not always necessarily true for intermediate values of θ. The following theorem shows that, at least in the rank-one case, nonnegativity is assured.

Theorem 3. *Consider the set of optimal solutions of Problem (6) when $\boldsymbol{A} \geq \boldsymbol{0}$. We have:*

(i) If Problem (6) has a rank-one optimal solution, then there exists a nonnegative rank-one optimal solution.

(ii) If $\theta > 1$, then all optimal solutions of Problem (6) are nonnegative.

Proof.

(i) Consider a rank-one optimal solution $\boldsymbol{X}$, $\boldsymbol{X} = \sigma\boldsymbol{u}\boldsymbol{v}^T$, of Problem (6). We prove that $|\boldsymbol{X}| = \sigma\,|\boldsymbol{u}|\,|\boldsymbol{v}|^T \geq \boldsymbol{0}$ is also an optimal solution. Let $\tilde{\boldsymbol{X}}$ denote $|\boldsymbol{X}|$. We have:

$$\left\|\tilde{\boldsymbol{X}}\right\|_\theta = \left\|\tilde{\boldsymbol{X}}\right\|_* + \theta\left\|\tilde{\boldsymbol{X}}\right\|_1 = \|\boldsymbol{X}\|_* + \theta\,\|\boldsymbol{X}\|_1\,.$$

In addition, $\langle\boldsymbol{A},\tilde{\boldsymbol{X}}\rangle \geq \langle\boldsymbol{A},\boldsymbol{X}\rangle$ since $\boldsymbol{A} \geq \boldsymbol{0}$. Thus clearly $\tilde{\boldsymbol{X}}$ is also an optimal solution.

(ii) Assume that there exists an optimal solution $\boldsymbol{X}$ of Problem (6) is not nonnegative. Without loss of generality, assume $x_{11} < 0$. Consider $\boldsymbol{X}(\epsilon) = \boldsymbol{X} + \epsilon \boldsymbol{E}_{11}$, where $\epsilon > 0$ and $\boldsymbol{E}_{11}$ is the matrix of all zeros except the element $\boldsymbol{E}_{11}(1,1) = 1$, we have:

$$\|\boldsymbol{X}(\epsilon)\|_* \leq \|\boldsymbol{X}\|_* + \epsilon \|\boldsymbol{E}_{11}\|_* = \|\boldsymbol{X}\|_* + \epsilon.$$

In addition, $\|\boldsymbol{X}(\epsilon)\|_1 = \|\boldsymbol{X}\|_1 - \epsilon$ if $\epsilon \leq |x_{11}|$. Therefore, we have:

$$\|\boldsymbol{X}(\epsilon)\|_\theta \leq \|\boldsymbol{X}\| + (1-\theta)\epsilon = 1 + (1-\theta)\epsilon < 1, \quad \forall\, 0 < \epsilon \leq |x_{11}|\,.$$

Here we assume that $\boldsymbol{A} \neq \boldsymbol{0}$; therefore, $\|\boldsymbol{X}\|_\theta = 1$. We also have

$$\langle \boldsymbol{A}, \boldsymbol{X}(\epsilon)\rangle = \langle \boldsymbol{A}, \boldsymbol{X}\rangle + \epsilon a_{11} \geq \langle \boldsymbol{A}, \boldsymbol{X}\rangle.$$

Now consider $\bar{\boldsymbol{X}} = \dfrac{1}{1+(1-\theta)\epsilon}\boldsymbol{X}(\epsilon)$. Clearly, $\left\|\bar{\boldsymbol{X}}\right\|_\theta = 1$ and $\langle \boldsymbol{A}, \bar{\boldsymbol{X}}\rangle > \langle \boldsymbol{A}, \boldsymbol{X}\rangle > 0$ (contradiction). Thus all optimal solutions of Problem (6) are nonnegative if $\theta > 1$.

□

3 Sparsity

As mentioned in the introduction, the penalty term $\theta\|\boldsymbol{X}\|_1$ in the objective function of (1) is intended to promote sparsity of $\boldsymbol{X}$. For some very simple convex optimization problems with an ℓ_1 penalty term, e.g., the unconstrained problem of minimizing $\|\boldsymbol{x} - \boldsymbol{c}\|_2 + \theta\|\boldsymbol{x}\|_1$ for a given vector $\boldsymbol{c}$, it is known that sparsity increases monotonically with θ (i.e., if $\boldsymbol{x}_1^*$ is the optimizer for θ_1 and $\boldsymbol{x}_2^*$ is the optimizer for θ_2 with $\theta_1 \leq \theta_2$, then the indices of nonzeros of $\boldsymbol{x}_2^*$ are a subset of the indices of nonzeros of $\boldsymbol{x}_1^*$).

For a more complicated problem such as (1), monotonicity does not hold in general. But nonetheless, some weaker statements about the relationship between θ and sparsity are possible. Two such results are derived in this section. We start with a lemma that leads to a sparsity result.

Lemma 4. *Assume $\boldsymbol{X} = \sigma \boldsymbol{u}\boldsymbol{v}^T$, where $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1$, $\boldsymbol{u} \geq \boldsymbol{0}$, and $\boldsymbol{v} \geq \boldsymbol{0}$, is the optimal solution of Problem (6). If $u_i > u_j = 0$ then*

$$\|A\|_\theta^* = \frac{\boldsymbol{a}_i^T \boldsymbol{v}}{\theta\,\|\boldsymbol{v}\|_1 + u_i} \geq \frac{\boldsymbol{a}_j^T \boldsymbol{v}}{\theta\,\|\boldsymbol{v}\|_1}, \tag{8}$$

where $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$ are the ith and jth row of $\boldsymbol{A}$.

Proof. We again assume here $\boldsymbol{A} \neq \boldsymbol{0}$, which means $\|\boldsymbol{X}\|_\theta = 1$. We have:

$$\|\boldsymbol{X}\|_\theta = \sigma + \theta\sigma \|\boldsymbol{u}\|_1 \|\boldsymbol{v}\|_1 = 1 \Leftrightarrow \sigma = \frac{1}{1 + \theta \|\boldsymbol{u}\|_1 \|\boldsymbol{v}\|_1}.$$

Consider $\boldsymbol{X}(\epsilon) = \sigma(\boldsymbol{u} + \epsilon \boldsymbol{e}_i)\boldsymbol{v}^T$, where $\epsilon \geq 0$ and $\boldsymbol{e}_i$ is the ith unit vector, we have:

$$\|\boldsymbol{u} + \epsilon \boldsymbol{e}_i\|_2^2 = \|\boldsymbol{u}\|_2^2 + (u_i + \epsilon)^2 - u_i^2 = 1 + 2u_i\epsilon + \epsilon^2.$$

Thus $\|\boldsymbol{X}(\epsilon)\|_* = \sigma\sqrt{1 + 2u_i\epsilon + \epsilon^2}$. On the other hand, $\|\boldsymbol{X}(\epsilon)\|_1 = \sigma(\|\boldsymbol{u}\|_1 + \epsilon) \|\boldsymbol{v}\|_1$. So we have:

$$\|\boldsymbol{X}(\epsilon)\|_\theta = 1 + \sigma\left(\sqrt{1 + 2u_i\epsilon + \epsilon^2} - 1 + \theta\epsilon \|\boldsymbol{v}\|_1\right) = 1 + \sigma\epsilon\left(\frac{2u_i + \epsilon}{\sqrt{1 + 2u_i\epsilon + \epsilon^2} + 1} + \theta \|\boldsymbol{v}\|_1\right).$$

Let $\alpha = \dfrac{2u_i + \epsilon}{\sqrt{1 + 2u_i\epsilon + \epsilon^2} + 1} + \theta \|\boldsymbol{v}\|_1$ and consider $\bar{\boldsymbol{X}} = \dfrac{\boldsymbol{X}(\epsilon)}{1 + \sigma\epsilon\alpha}$, we have: $\left\|\bar{\boldsymbol{X}}\right\|_\theta = 1$ and

$$\langle \boldsymbol{A}, \bar{\boldsymbol{X}} \rangle = \frac{\langle \boldsymbol{A}, \boldsymbol{X}(\epsilon) \rangle}{1 + \sigma\epsilon\alpha} = \frac{\|\boldsymbol{A}\|_\theta^* + \sigma\epsilon \boldsymbol{a}_i^T \boldsymbol{v}}{1 + \sigma\epsilon\alpha} \leq \|\boldsymbol{A}\|_\theta^*.$$

With $\sigma > 0$ and $\epsilon > 0$, we obtain the following inequality

$$\boldsymbol{a}_i^T \boldsymbol{v} \leq \left(\frac{2u_i + \epsilon}{\sqrt{1 + 2u_i\epsilon + \epsilon^2} + 1} + \theta \|\boldsymbol{v}\|_1\right) \|\boldsymbol{A}\|_\theta^*.$$

Taking the limit $\epsilon \to 0^+$, we have:

$$\|\boldsymbol{A}\|_\theta^* \geq \frac{\boldsymbol{a}_i^T \boldsymbol{v}}{\theta \|\boldsymbol{v}\|_1 + u_i}.$$

Now consider the case in which $u_i > 0$ and set $\boldsymbol{X}(\epsilon) = \sigma(\boldsymbol{u} - \epsilon \boldsymbol{e}_i)\boldsymbol{v}^T$, where $0 \leq \epsilon \leq u_i$. Similarly, we have:

$$\|\boldsymbol{X}(\epsilon)\|_\theta = 1 - \sigma\epsilon\left(\frac{2u_i - \epsilon}{\sqrt{1 - 2u_i\epsilon + \epsilon^2} + 1} + \theta \|\boldsymbol{v}\|_1\right).$$

This implies the following inequality

$$\boldsymbol{a}_i^T \boldsymbol{v} \geq \left(\frac{2u_i - \epsilon}{\sqrt{1 - 2u_i\epsilon + \epsilon^2} + 1} + \theta \|\boldsymbol{v}\|_1\right) \|\boldsymbol{A}\|_\theta^*.$$

Again, taking the limit $\epsilon \to 0^+$, we have:

$$\|\boldsymbol{A}\|_\theta^* \leq \frac{\boldsymbol{a}_i^T \boldsymbol{v}}{\theta \|\boldsymbol{v}\|_1 + u_i}.$$

From these two results, we can see that if $u_i > u_j = 0$, then

$$\|A\|_\theta^* = \frac{\boldsymbol{a}_i^T \boldsymbol{v}}{\theta \|\boldsymbol{v}\|_1 + u_i} \geq \frac{\boldsymbol{a}_j^T \boldsymbol{v}}{\theta \|\boldsymbol{v}\|_1}.$$

□

Since the roles of columns and rows are interchangeable, we also have the following result. If $v_k > v_l = 0$, then

$$\|A\|_\theta^* = \frac{\boldsymbol{u}^T \boldsymbol{A}_k}{\theta \|\boldsymbol{u}\|_1 + v_k} \geq \frac{\boldsymbol{u}^T \boldsymbol{A}_l}{\theta \|\boldsymbol{u}\|_1}, \tag{9}$$

where $\boldsymbol{A}_k$ and $\boldsymbol{A}_l$ are k-th and l-th column of $\boldsymbol{A}$.

The sparsity structure of $\boldsymbol{X} = \sigma \boldsymbol{u}\boldsymbol{v}^T$ depends on the sparsity structure of $\boldsymbol{u}$ and $\boldsymbol{v}$. The results obtained above help us derive some conditions under which a row (or column) of $\boldsymbol{X}$ is zero.

Corollary 1. *Consider two rows $\boldsymbol{a}_i^T$ and $\boldsymbol{a}_j^T$ of matrix $\boldsymbol{A}$. If $\min_k \{a_{ik}\} \geq \alpha \max_k \{a_{jk}\}$, where $\alpha > 1$, then for every $\theta > \dfrac{1}{\alpha - 1}$ and for every nonnegative rank-one optimal solution $\boldsymbol{X}$ of Problem (6), the jth row of $\boldsymbol{X}$ is zero.*

Proof. Assume that $\boldsymbol{X} = \sigma \boldsymbol{u}\boldsymbol{v}^T$, where $\boldsymbol{u} \geq \boldsymbol{0}$ and $\boldsymbol{v} \geq \boldsymbol{0}$, and $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1$. We have:

$$\frac{\boldsymbol{a}_i^T \boldsymbol{v}}{\theta \|\boldsymbol{v}\|_1 + u_i} \geq \frac{\min_k\{a_{ik}\} \|\boldsymbol{v}\|_1}{\theta \|\boldsymbol{v}\|_1 + 1} \geq \frac{\alpha \max_k\{a_{jk}\} \|\boldsymbol{v}\|_1}{\theta \|\boldsymbol{v}\|_1 + 1} = \frac{\alpha \max_k\{a_{jk}\}}{\theta + \dfrac{1}{\|\boldsymbol{v}\|_1}} \geq \frac{\alpha \max_k\{a_{jk}\}}{\theta + 1},$$

since $\|\boldsymbol{v}\|_1 \geq \|\boldsymbol{v}\|_2 = 1$. On the other hand, we also have the following inequality

$$\frac{\boldsymbol{a}_j^T \boldsymbol{v}}{\theta \|\boldsymbol{v}\|_1 + u_j} \leq \frac{\max_k\{a_{jk}\} \|\boldsymbol{v}\|_1}{\theta \|\boldsymbol{v}\|_1} = \frac{\max_k\{a_{jk}\}}{\theta}.$$

We have:

$$\frac{\alpha \max_k\{a_{jk}\}}{\theta + 1} - \frac{\max_k\{a_{jk}\}}{\theta} = \max_k\{a_{kj}\} \frac{(\alpha - 1)\theta - 1}{\theta(\theta + 1)} > 0, \quad \forall\, \theta > \frac{1}{\alpha - 1}.$$

Thus we have:

$$\frac{\boldsymbol{a}_i^T \boldsymbol{v}}{\theta \|\boldsymbol{v}\|_1 + u_i} > \frac{\boldsymbol{a}_j^T \boldsymbol{v}}{\theta \|\boldsymbol{v}\|_1 + u_j}, \quad \forall\, \theta > \frac{1}{\alpha - 1},$$

which means $u_j = 0$ according to Lemma 4. Thus the j-th row of $\boldsymbol{X}$ is zero. □

We would like to use these results to build up results for columns and rows simultaneously. More exactly, consider a subset $\mathcal{I} \subset \{1, \ldots, m\}$ and $\mathcal{J} \subset \{1, \ldots, n\}$, we would like to obtain conditions on magnitudes of elements of $\boldsymbol{A}(\mathcal{I}, \mathcal{J})$ as compared to those of the remaining elements of $\boldsymbol{A}$ to guarantee that all rows and columns that are not in $\mathcal{I}$ and $\mathcal{J}$ have to be zero in the nonnegative rank-one optimal matrix $\boldsymbol{X}$ of Problem (6) for $\theta \geq \theta_0$. One of the difficulties here is that under these conditions, there is a coupling relationship between rows and columns. More exactly, in order to prove the rows that are not in $\mathcal{I}$ are zero, we need to prove the columns that are not in $\mathcal{J}$ are small or zero at the same time.

Lemma 4 and Corollary 1 are based on local optimality conditions with respect to rows or columns. We can obtain additional results on the sparsity of the optimal solution $\boldsymbol{X}$ using the global optimality conditions.

The following theorem states that if the weight of nonnegative matrix $\boldsymbol{A}$ is concentrated in a particular subblock then for θ sufficiently large, the optimal solution $\boldsymbol{X}$ will have nonzero entries only in that subblock. "Concentration of weight" in this sense means that the average of those entries dominates all the other entries of the matrix.

As a special case, this theorem implies that if the maximum entry of $\boldsymbol{A}$ is unique, then for θ sufficiently large, $\boldsymbol{X}$ will be a singleton matrix whose unique nonzero entry corresponds to the maximum entry of $\boldsymbol{A}$.

Theorem 4. *Assume $\boldsymbol{A} \geq \boldsymbol{0}$. Let $\mathcal{I}$ and $\mathcal{J}$ be subsets of $\{1, \ldots, m\}$ and $\{1, \ldots, n\}$, respectively; $|\mathcal{I}| = M$ and $|\mathcal{J}| = N$. Define $\bar{a}(\mathcal{I}, \mathcal{J}) = \dfrac{1}{MN} \sum\limits_{(i,j) \in (\mathcal{I}, \mathcal{J})} a_{ij}$ and $a_{\max}(\overline{\mathcal{I}, \mathcal{J}}) = \max\limits_{(i,j) \notin (\mathcal{I}, \mathcal{J})} a_{ij}$. If $\bar{a}(\mathcal{I}, \mathcal{J}) > a_{\max}(\overline{\mathcal{I}, \mathcal{J}})$ then all optimal solutions $\boldsymbol{X}$ of Problem (6) are sparse, $x_{ij} = 0$ for all $(i, j) \notin (\mathcal{I}, \mathcal{J})$, for all $\theta > \theta_B$, where $\theta_B = \dfrac{1}{\sqrt{MN}} \left(\dfrac{\bar{a}(\mathcal{I}, \mathcal{J})\sqrt{MM} + a_{\max}(\overline{\mathcal{I}, \mathcal{J}})}{\bar{a}(\mathcal{I}, \mathcal{J}) - a_{\max}(\overline{\mathcal{I}, \mathcal{J}})} \right)$.*

Proof. Assume there exists an optimal solution $\boldsymbol{X}$ of Problem (6) such that $x_{ij} \neq 0$ for some $(i, j) \notin (\mathcal{I}, \mathcal{J})$ when $\theta > \theta_B$. We have: $\theta_B > 1$; therefore, according to Theorem 3, $\boldsymbol{X} \geq \boldsymbol{0}$. Thus $x_{ij} > 0$. We also have: $\boldsymbol{A} \neq \boldsymbol{0}$; therefore $\boldsymbol{X} \neq \boldsymbol{0}$ and $\|\boldsymbol{X}\|_\theta = 1$. Consider two cases, $a_{ij} = 0$ and $a_{ij} > 0$.

If $a_{ij} = 0$, let $\boldsymbol{X}_0 = \boldsymbol{X} - x_{ij}\boldsymbol{E}_{ij}$, where $\boldsymbol{E}_{ij}$ is the matrix of all zeros but the element $\boldsymbol{E}_{ij}(i, j) = 1$. We have: $\|\boldsymbol{X}_0\| \leq \|\boldsymbol{X}\| + x_{ij}$ and $\|\boldsymbol{X}_0\|_1 = \|\boldsymbol{X}\|_1 - x_{ij}$. Thus

$$\|\boldsymbol{X}_0\|_\theta \leq \|\boldsymbol{X}\|_\theta + (1 - \theta)x_{ij} < 1, \quad \forall\, \theta > \theta_B > 1.$$

We also have: since $a_{ij} = 0$, $\langle \boldsymbol{A}, \boldsymbol{X}_0 \rangle = \langle \boldsymbol{A}, \boldsymbol{X} \rangle = \|\boldsymbol{A}\|_\theta^* > 0$. Thus $\boldsymbol{X}_0 \neq \boldsymbol{0}$ or $\|\boldsymbol{X}_0\|_\theta > 0$. Define $\boldsymbol{X}_0^s = \dfrac{1}{\|\boldsymbol{X}_0\|_\theta}\boldsymbol{X}_0$, we have: $\boldsymbol{X}_0^s$ is a feasible solution of Problem (6) with the objective value $\langle \boldsymbol{A}, \boldsymbol{X}_0^s \rangle = \dfrac{\|\boldsymbol{A}\|_\theta^*}{\|\boldsymbol{X}_0\|_\theta} > \|\boldsymbol{A}\|_\theta^*$ (contradiction).

Now consider the case when $a_{ij} > 0$. Define $\boldsymbol{D} = \boldsymbol{e}_{\mathcal{I}}\boldsymbol{e}_{\mathcal{J}}^T - MNr\boldsymbol{e}_i\boldsymbol{e}_j^T$, where $r = \dfrac{\bar{a}(\mathcal{I}, \mathcal{J})}{a_{ij}}$, $\boldsymbol{e}_{\mathcal{I}} = \sum\limits_{i \in \mathcal{I}} \boldsymbol{e}_i$, $\boldsymbol{e}_i \in \mathbb{R}^m$ is the ith unit vector in $\mathbb{R}^m$, and similarly, $\boldsymbol{e}_{\mathcal{J}} = \sum\limits_{j \in \mathcal{J}} \boldsymbol{e}_j$, $\boldsymbol{e}_j \in \mathbb{R}^n$ is the j-th unit vector in $\mathbb{R}^n$. We have r is well-defined since $\boldsymbol{A} > \boldsymbol{0}$ and $r > 1$. We now consider a new solution $\boldsymbol{X}_\alpha = \boldsymbol{X} + \alpha\boldsymbol{D}$, where $0 < \alpha \leq \dfrac{x_{ij}}{MNr}$. Clearly, $\boldsymbol{X}_\alpha \geq \boldsymbol{0}$. Thus we have: $\|\boldsymbol{X}_\alpha\|_1 = \|\boldsymbol{X}\|_1 + \alpha MN(1 - r)$. Applying the triangle inequality, we can bound $\|\boldsymbol{X}_\alpha\|_*$ as follows:

$$\|\boldsymbol{X}_\alpha\|_* \leq \|\boldsymbol{X}\|_* + \alpha\,\|\boldsymbol{D}\|_* \leq \|\boldsymbol{X}\|_* + \alpha(\sqrt{MN} + MNr).$$

Thus we have:

$$\|\boldsymbol{X}_\alpha\|_\theta \leq \|\boldsymbol{X}\|_\theta + \alpha\sqrt{MN}\left[(1 + r\sqrt{MN}) + \theta\sqrt{MN}(1 - r)\right].$$

Since $\theta > \theta_B = \dfrac{1}{\sqrt{MN}}\left(\dfrac{\bar{a}(\mathcal{I},\mathcal{J})\sqrt{MM} + a_{\max}(\overline{\mathcal{I},\mathcal{J}})}{\bar{a}(\mathcal{I},\mathcal{J}) - a_{\max}(\overline{\mathcal{I},\mathcal{J}})}\right)$ and $0 < a_{ij} \leq a_{\max}(\overline{\mathcal{I},\mathcal{J}}) < \bar{a}(\mathcal{I},\mathcal{J})$, we have:

$$\theta > \frac{1}{\sqrt{MN}}\left(\frac{1 + r\sqrt{MN}}{r-1}\right).$$

This implies that $\|\boldsymbol{X}_\alpha\|_\theta < \|\boldsymbol{X}\|_\theta = 1$ for all $0 < \alpha \leq \dfrac{x_{ij}}{MNr}$. Now consider the scaled solution $\boldsymbol{X}^s_\alpha = \dfrac{1}{\|\boldsymbol{X}_\alpha\|_\theta}\boldsymbol{X}_\alpha$, which is also a feasible solution of Problem (6). In terms of the objective, we have: $\langle \boldsymbol{A}, \boldsymbol{D}\rangle = \boldsymbol{e}_{\mathcal{I}}^T \boldsymbol{A}\boldsymbol{e}_{\mathcal{J}} - MNra_{ij} = 0$. Thus $\langle \boldsymbol{A}, \boldsymbol{X}_\alpha\rangle = \langle \boldsymbol{A}, \boldsymbol{X}\rangle = \|\boldsymbol{A}\|^*_\theta$ for all α. We then have:

$$\langle \boldsymbol{A}, \boldsymbol{X}^s_\alpha\rangle = \frac{\|\boldsymbol{A}\|^*_\theta}{\|\boldsymbol{X}_\alpha\|_\theta} > \|\boldsymbol{A}\|^*_\theta, \quad \forall\, \alpha \in \left(0, \frac{x_{ij}}{MNr}\right],$$

which is a contradiction because $\|\boldsymbol{A}\|^*_\theta$ is the optimal value of Problem (6).

Thus we can conclude that if $\boldsymbol{A} \geq \boldsymbol{0}$, all optimal solutions $\boldsymbol{X}$ of Problem (6) are sparse with $x_{ij} = 0$ for all $(i,j) \notin (\mathcal{I},\mathcal{J})$ when $\theta > \theta_B$. □

4 Random noise

The main technical result of this article is that the proposed algorithm can find a large rank-one submatrix hidden in a substantial amount of noise. The noise takes two forms: the rank-one submatrix itself has random noise added to it (so that its rank is no longer 1), and the entries outside the rank-one submatrix are generated by a random process.

First, we recall the following definition: a random variable x is *b-subgaussian* if its mean is zero, and if there exists a $b > 0$ such that for all $t \geq 0$,

$$\mathbb{P}(|x| \geq t) \leq \exp(-t^2/(2b^2)). \tag{10}$$

For example, a normally distributed variable or any mean-zero variable with a discrete distribution is subgaussian.

The result of this section is the following bound. For this entire section, we adopt the notation that $\boldsymbol{e}_M$ denotes the vector of all 1's of length M, and similarly for $\boldsymbol{e}_N$.

Theorem 5. *Let $\boldsymbol{A}$ be an $m \times n$ matrix defined as follows.*

$$\boldsymbol{A} = \begin{pmatrix} \sigma_0 \boldsymbol{u}_0 \boldsymbol{v}_0^T & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \boldsymbol{R}_{21} & \boldsymbol{R}_{22} \end{pmatrix}, \tag{11}$$

where $\sigma_0 > 0$, $\boldsymbol{u}_0 \in \mathbb{R}^M$, $\boldsymbol{u}_0 \geq \boldsymbol{0}$, $M < m$, *and* $\boldsymbol{v}_0 \in \mathbb{R}^N$, $\boldsymbol{v}_0 \geq \boldsymbol{0}$, $N < n$. *Furthermore, assume that* $\boldsymbol{u}_0 = \boldsymbol{e}_M + \boldsymbol{p}$ *with* $\|\boldsymbol{p}\|_2 \leq c_1\sqrt{M}$, *and* $\boldsymbol{v}_0 = \boldsymbol{e}_N + \boldsymbol{q}$ *with* $\|\boldsymbol{q}\|_2 \leq c_2\sqrt{N}$. *The matrix* $\boldsymbol{R}$ *is a random matrix with i.i.d. nonnegative elements* r_{ij} *with mean* $c_3\sigma_0$, *where* $c_3 > 0$ *is a constant, such that* $r_{ij}/\sigma_0 - c_3$ *is b-subgaussian. Here* c_1, c_2, c_3 *are positive constants. Assume that these scalar constants* c_1, c_2, c_3 *satisfy the following relations*

$$c_5 \leq 1/3, \qquad c_3 + c_5 < 1. \tag{12}$$

where c_5 *is chosen to satisfy*

$$c_5 > c_1 + c_2 + c_1 c_2. \tag{13}$$

Under these hypotheses concerning $\boldsymbol{A}$, *and assuming* θ *satisfies*

$$\theta \leq \min\left(\frac{1}{c_3 + c_5}, \frac{1 + c_3 - 3c_5}{2c_5}\right) \cdot \frac{1}{\sqrt{MN}}, \tag{14}$$

$$\theta \geq \frac{2c_3}{1 - c_3 - c_5} \cdot \frac{1}{\sqrt{MN}}, \tag{15}$$

the solution $\boldsymbol{X}$ *to problem (1) is a rank-one matrix with positive entries in positions that are indexed by* $\{1, \ldots, M\} \times \{1, \ldots, N\}$ *and zeros elsewhere with probability exponentially close to* 1 *(i.e., of the form* $1 - \exp(-(M+N)^{\text{const}})$*) provided that* $MN \geq \Omega\left((M+N)^{4/3}\right)$ *and* $MN \geq \Omega(m+n)$. *Here, the constants implicit in the* $\Omega(\cdot)$ *notation depend on* b *and* c_5. *See (32)–(35) below for a detailed presentation of these constants.*

Remarks.

1. Naturally, the theorem also applies if the MN distinguished entries occur as any $M \times N$ submatrix of $\boldsymbol{A}$; we have numbered the distinguished submatrix first in order to simplify notation.

2. It is not enough to assume simply that $\boldsymbol{u}_0 > \boldsymbol{0}$ and $\boldsymbol{v}_0 > \boldsymbol{0}$ because if these vectors have very small entries, then they cannot be distinguished from the noise.

3. This theorem is not a consequence of Theorem 4 because the hypotheses do not force entries outside the distinguished block to be smaller than the average of the distinguished block's entries.

4. The relationships among the constants as well as (14), (15) can all be satisfied provided c_3, c_5 are sufficiently small.

5. The result holds with probability exponentially close to 1 as long as $M \sim N$ and $M \geq \Omega(m^{1/2})$, $N \geq \Omega(n^{1/2})$. Thus, the rank-one submatrix can be much smaller than the entire matrix $\boldsymbol{A}$.

Before beginning the proof of the theorem, we require the following key lemma regarding matrices constructed from independent b-subgaussian random variables.

Lemma 5. *Let $\boldsymbol{B} \in \mathbb{R}^{m\times n}$ be a random matrix, where b_{ij} are independent b-subgaussian random variables for all $i = 1, \ldots, m$, and $j = 1, \ldots, n$. Then for any $u > 0$,*

(i) $\mathbb{P}\left(\|\boldsymbol{B}\| \geq u\right) \leq \exp\left(-\left(\frac{8u^2}{81b^2} - (\log 7)(m+n)\right)\right)$

(ii) $\mathbb{P}\left(\|\boldsymbol{C}\boldsymbol{B}\| \geq u\right) \leq \exp\left(-\left(\frac{8u^2}{81b^2\|\boldsymbol{C}\|^2} - (\log 7)(m+n)\right)\right)$, *where $\boldsymbol{C}$ is a deterministic matrix.*

The proof of this lemma follows the proof techniques by Litvak et al. [14]. Major steps are shown as follows.

Proof.

(i) We have: $\|\boldsymbol{B}\| = \max\limits_{\|\boldsymbol{x}\|_2=\|\boldsymbol{y}\|_2=1} \boldsymbol{y}^T\boldsymbol{B}\boldsymbol{x}$. We discretize the unit balls in $\mathbb{R}^n$ and $\mathbb{R}^m$ by finite ϵ-nets, where $\epsilon \in (0,1)$. An ϵ-net of a set $\mathcal{K}$ is the subset $\mathcal{N}$ such that for all $\boldsymbol{x} \in \mathcal{K}$, there exists $\boldsymbol{y} \in \mathcal{N}$ such that $\|\boldsymbol{x}-\boldsymbol{y}\|_2 \leq \epsilon$. Using a construction proof, we can prove that there exists a finite ϵ-net of the unit ball in $\mathbb{R}^n$ with the cardinality of no more than $\left(\frac{2}{\epsilon}+1\right)^n$. Let $\mathcal{N}$ and $\mathcal{M}$ be the finite ϵ-nets of the unit balls in $\mathbb{R}^n$ and $\mathbb{R}^m$ with minimum cardinality, respectively. Applying the triangle inequality, we have:

$$\|\boldsymbol{B}\| \leq \frac{1}{(1-\epsilon)^2} \max_{\boldsymbol{x}\in\mathcal{N},\boldsymbol{y}\in\mathcal{M}} \boldsymbol{y}^T\boldsymbol{B}\boldsymbol{x}.$$

We can bound the tail probability $P(\|\boldsymbol{B}\| \geq u)$ as follows.

$$\mathbb{P}\left(\|\boldsymbol{B}\| \geq u\right) \leq \left(\frac{2}{\epsilon}+1\right)^{m+n} \max_{\boldsymbol{x}\in\mathcal{N},\boldsymbol{y}\in\mathcal{M}} \mathbb{P}\left(\boldsymbol{y}^T\boldsymbol{B}\boldsymbol{x} \geq (1-\epsilon)^2 u\right).$$

We have, b_{ij} are independent b-subgaussian random variables; therefore, $\boldsymbol{y}^T\boldsymbol{B}\boldsymbol{x} = \sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n}(x_j y_i)\, b_{ij}$ is also a b-subgaussian random variable since $\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}(x_j y_i)^2 = \|\boldsymbol{x}\|_2^2\,\|\boldsymbol{y}\|_2^2 = 1$. Thus we have:

$$\mathbb{P}\left(\|\boldsymbol{B}\| \geq u\right) \leq \left(\frac{2}{\epsilon}+1\right)^{m+n} e^{-\frac{(1-\epsilon)^4 u^2}{2b^2}}.$$

Letting $\epsilon = 1/3$, we obtain the inequality

$$\mathbb{P}\left(\|\boldsymbol{B}\| \geq u\right) \leq e^{-\left(\frac{8u^2}{81b^2} - (\log 7)(m+n)\right)}$$

(ii) We have: $\boldsymbol{y}^T\boldsymbol{C}\boldsymbol{B}\boldsymbol{x} = \left(\boldsymbol{C}^T\boldsymbol{y}\right)^T \boldsymbol{B}\boldsymbol{x}$, thus:

$$\mathbb{P}\left(\boldsymbol{y}^T\boldsymbol{C}\boldsymbol{B}\boldsymbol{x} \geq u\right) \leq e^{-\frac{u^2}{2b^2\|\boldsymbol{C}^T\boldsymbol{y}\|_2^2\|\boldsymbol{x}\|_2^2}} \leq e^{-\frac{u^2}{2b^2\|\boldsymbol{C}\|^2}}.$$

Applying similar arguments, we can then obtain the inequality (ii) of the Lemma.

□

We now turn to the proof of the main theorem. We now would like to find conditions on θ and the constants so that Problem (1) has an optimal solution $\boldsymbol{X}$ of the form

$$\boldsymbol{X} = \begin{pmatrix} \sigma_1\boldsymbol{u}_1\boldsymbol{v}_1^T & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix},$$

where $\boldsymbol{u}_1 > \boldsymbol{0}$, $\|\boldsymbol{u}_1\|_2 = 1$, and $\boldsymbol{v}_1 > \boldsymbol{0}$, $\|\boldsymbol{v}_1\|_2 = 1$. If $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ are determined, σ_1 can be easily calculated in order to satisfy the condition $\langle \boldsymbol{A}, \boldsymbol{X}\rangle = 1$ of the optimal solution. Thus the main task is to find $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ if they exist. We will construct them using optimality conditions derived in the previous section for Problem (6) (equivalent to Problem (1)) and its dual, Problem (7). Defining $\boldsymbol{u} = [\boldsymbol{u}_1; \boldsymbol{0}] \in \mathbb{R}^m$ and $\boldsymbol{v} = [\boldsymbol{v}_1; \boldsymbol{0}] \in \mathbb{R}^n$, we can then write the optimality conditions as follows:

There exists $\boldsymbol{Y}$ and $\boldsymbol{Z}$ such that $\boldsymbol{A} = \boldsymbol{Y} + \boldsymbol{Z}$ and

$$\boldsymbol{Y} = \|\boldsymbol{A}\|_\theta^* (\boldsymbol{u}\boldsymbol{v}^T + \boldsymbol{W}), \quad \boldsymbol{Z} = \theta \|\boldsymbol{A}\|_\theta^* \boldsymbol{V},$$

where $\|\boldsymbol{W}\| \leq 1$, $\boldsymbol{W}^T\boldsymbol{u} = \boldsymbol{0}$, $\boldsymbol{W}\boldsymbol{v} = \boldsymbol{0}$, and $\|\boldsymbol{V}\|_\infty \leq 1$, $\boldsymbol{V}_{11} = \boldsymbol{e}_M\boldsymbol{e}_N^T$.

These conditions come from the properties of the subgradient $\partial\|\boldsymbol{Y}\|$ and $\partial\|\boldsymbol{Z}\|_\infty$ and the fact that $\boldsymbol{X}$ belongs to these sets (up to appropriate scaling factors).

In the following analysis, we will construct $(\boldsymbol{V}, \boldsymbol{W})$ so that the optimality conditions are satisfied. The entries of these matrices will be constructed separately for the four subblocks of $\boldsymbol{A}$, starting with the $(1,1)$ block. Breaking the equation $\boldsymbol{Y} + \boldsymbol{Z} = \boldsymbol{A}$ into blocks and scaling by $1/\|\boldsymbol{A}\|_\theta^*$, we obtain the following more detailed optimality conditions.

$$\boldsymbol{u}_1\boldsymbol{v}_1^T + \boldsymbol{W}_{11} + \theta\boldsymbol{e}_M\boldsymbol{e}_N^T = \boldsymbol{A}_{11}/\|\boldsymbol{A}\|_\theta^* = (\sigma_0\boldsymbol{u}_0\boldsymbol{v}_0^T + \boldsymbol{R}_{11})/\|\boldsymbol{A}\|_\theta^*, \tag{16}$$

$$\boldsymbol{W}_{ij} + \theta\boldsymbol{V}_{ij} = \boldsymbol{A}_{ij}/\|\boldsymbol{A}\|_\theta^* = \boldsymbol{R}_{ij}/\|\boldsymbol{A}\|_\theta^*, \tag{17}$$

where the second line applies to (i,j) equal to $(1,2)$, $(2,1)$ and $(2,2)$. Following this block matrix notation, the remaining optimality conditions to be established are $\boldsymbol{W}_{11}^T\boldsymbol{u}_1 = \boldsymbol{0}$, $\boldsymbol{W}_{21}^T\boldsymbol{u}_1 = \boldsymbol{0}$, $\boldsymbol{W}_{11}\boldsymbol{v}_1 =$

$\mathbf{0}$, $\boldsymbol{W}_{12}\boldsymbol{v}_1 = \mathbf{0}$, $\|\boldsymbol{V}\|_\infty \leq 1$, $\|\boldsymbol{W}\| \leq 1$. We shall establish the latter inequality by proving more specifically that $\|\boldsymbol{W}_{ij}\| \leq 1/2$ for $(i,j) \in \{1,2\} \times \{1,2\}$.

We begin with the $(1,1)$ block of this equation. The conditions $\|\boldsymbol{W}_{11}\| \leq 1/2$, $\boldsymbol{W}_{11}^T\boldsymbol{u}_1 = \mathbf{0}$, $\boldsymbol{W}_{11}\boldsymbol{v}_1 = \mathbf{0}$ imply that $\left\|\boldsymbol{u}_1\boldsymbol{v}_1^T + \boldsymbol{W}_{11}\right\| = 1$. This is because the dominant singular triple of $\boldsymbol{u}_1\boldsymbol{v}_1^T + \boldsymbol{W}_{11}$ must be $(1, \boldsymbol{u}_1, \boldsymbol{v}_1)$ by the conditions. Equivalent to $\left\|\boldsymbol{u}_1\boldsymbol{v}_1^T + \boldsymbol{W}_{11}\right\| = 1$ is

$$\left\| \frac{1}{\|A\|_\theta^*}\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11} \right\| = 1,$$

where, as noted earlier, we are required to take $\boldsymbol{V}_{11} = \boldsymbol{e}_M\boldsymbol{e}_N^T$.

Thus the first necessary condition for $\boldsymbol{X}$ to be the optimal solution is that there exists $\lambda > 0$ such that $f(\lambda) = \|\lambda\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11}\| = 1$. If such a λ is identified, then $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ can be easily found since $\boldsymbol{u}_1\boldsymbol{v}_1^T$ is the rank-one approximation of $\lambda\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11}$. Note that the nonnegativity of $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ will require additional conditions which will be discussed later. We have:

$$\begin{aligned}
\lambda\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11} &= \lambda\left[\sigma_0\boldsymbol{u}_0\boldsymbol{v}_0^T + \boldsymbol{R}_{11}\right] - \theta\boldsymbol{e}_M\boldsymbol{e}_N^T \\
&= \lambda\left[\sigma_0(\boldsymbol{e}_M + \boldsymbol{p})(\boldsymbol{e}_N + \boldsymbol{q})^T + \boldsymbol{R}_{11}\right] - \theta\boldsymbol{e}_M\boldsymbol{e}_N^T \\
&= (\lambda\sigma_0 - \theta)\boldsymbol{e}_M\boldsymbol{e}_N^T + \lambda\left[\sigma_0(\boldsymbol{e}_M\boldsymbol{q}^T + \boldsymbol{p}\boldsymbol{e}_N^T + \boldsymbol{p}\boldsymbol{q}^T) + \boldsymbol{R}_{11}\right] \\
&= [\lambda\sigma_0(1 + c_3) - \theta]\boldsymbol{e}_M\boldsymbol{e}_N^T + \lambda\left[\sigma_0(\boldsymbol{e}_M\boldsymbol{q}^T + \boldsymbol{p}\boldsymbol{e}_N^T + \boldsymbol{p}\boldsymbol{q}^T) + (\boldsymbol{R}_{11} - c_3\sigma_0\boldsymbol{e}_M\boldsymbol{e}_N^T)\right].
\end{aligned}$$

We have: $f(\lambda) \to +\infty$ when $\lambda \to +\infty$ since $\boldsymbol{A}_{11} \neq \mathbf{0}$. Now define $\lambda_0 = \dfrac{\theta}{\sigma_0(1 + c_3)}$ to make the first term vanish, yielding

$$\begin{aligned}
\lambda_0\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11} &= \lambda_0\left[\sigma_0(\boldsymbol{e}_M\boldsymbol{q}^T + \boldsymbol{p}\boldsymbol{e}_N^T + \boldsymbol{p}\boldsymbol{q}^T) + (\boldsymbol{R}_{11} - c_3\sigma_0\boldsymbol{e}_M\boldsymbol{e}_N^T)\right] \\
&= \lambda_0\left[\sigma_0\boldsymbol{P} + \boldsymbol{Q}\right],
\end{aligned}$$

where $\boldsymbol{P} = \boldsymbol{e}_M\boldsymbol{q}^T + \boldsymbol{p}\boldsymbol{e}_N^T + \boldsymbol{p}\boldsymbol{q}^T$ and $\boldsymbol{Q} = \boldsymbol{R}_{11} - c_3\sigma_0\boldsymbol{e}_M\boldsymbol{e}_N^T$. We now bound the spectral norm of $\boldsymbol{P}$ and $\boldsymbol{Q}$ as follows.

$$\begin{aligned}
\|\boldsymbol{P}\| &\leq \left\|\boldsymbol{e}_M\boldsymbol{q}^T\right\| + \left\|\boldsymbol{p}\boldsymbol{e}_N^T\right\| + \left\|\boldsymbol{p}\boldsymbol{q}^T\right\| \\
&= \|\boldsymbol{e}\|_2\|\boldsymbol{q}\|_2 + \|\boldsymbol{p}\|_2\|\boldsymbol{e}_N\|_2 + \|\boldsymbol{p}\|_2\|\boldsymbol{q}\|_2 \\
&\leq c_1\sqrt{MN} + c_2\sqrt{MN} + c_1c_2\sqrt{MN} \\
&= (c_1 + c_2 + c_1c_2)\sqrt{MN}.
\end{aligned}$$

Matrix $\boldsymbol{Q}/\sigma_0$ is random with i.i.d. elements that are b-subgaussian. Thus by Lemma 5(i),

$$\mathbb{P}\left(\|\boldsymbol{Q}\| \geq u\sigma_0\right) \leq \exp\left(-\left(\frac{8u^2}{81b^2} - (\log 7)(M + N)\right)\right),$$

for any $u > 0$. Let us fix $u = (MN)^{3/8}$ to obtain

$$\mathbb{P}\left(\|\boldsymbol{Q}\| \geq \sigma_0 (MN)^{3/8}\right) \leq \exp\left(-\left(\frac{(MN)^{3/4}}{81b^2} - (\log 7)(M+N)\right)\right). \tag{18}$$

For the remainder of this analysis, we will impose the assumption that the event in (18) does not happen. At the end of the proof the right-hand side (18) will be one of the terms in the failure probability of identifying the optimal $\boldsymbol{X}$.

Thus, $\|\boldsymbol{Q}\| \leq o(1)\sigma_0\sqrt{MN}$, so applying the triangle inequality,

$$\begin{aligned} \|\sigma_0 \boldsymbol{P} + \boldsymbol{Q}\| &\leq \sigma_0(c_1 + c_2 + c_1 c_2 + o(1))\sqrt{MN} \\ &\leq \sigma_0 c_5 \sqrt{MN}. \end{aligned} \tag{19}$$

by (13). (The strict inequality '>' in (13) is used in order to absorb the $o(1)$ term.) Therefore,

$$\begin{aligned} f(\lambda_0) &= \lambda_0 \|\sigma_0 \boldsymbol{P} + \boldsymbol{Q}\| \\ &\leq \frac{c_5 \theta \sqrt{MN}}{1 + c_3}. \end{aligned}$$

Thus with high probability, $f(\lambda_0) \leq 1$ if

$$\theta < \left(\frac{1 + c_3}{c_5}\right) \frac{1}{\sqrt{MN}}, \tag{20}$$

Inequality (20) is a consequence of (14) stated in the theorem. This inequality implies $f(\lambda_0) \leq 1$, and, due to the continuity of function f, there exists $\lambda^* \geq \lambda_0$ such that $f(\lambda^*) = 1$. We will prove that under some additional conditions, this value λ^* satisfies all other optimality conditions of Problem (1) and indeed $\|\boldsymbol{A}\|_\theta^* = \dfrac{1}{\lambda^*}$.

Let us recall that $\|\lambda^* \boldsymbol{A}_{11} - \theta \boldsymbol{V}_{11}\| = 1$, i.e., $\|(\lambda^* \sigma_0 (1 + c_3) - \theta)\boldsymbol{e}_M \boldsymbol{e}_N^T + \lambda^*(\sigma_0 \boldsymbol{P} + \boldsymbol{Q})\| = 1$. Applying the fact that $\|\boldsymbol{e}_M \boldsymbol{e}_N^T\| = \sqrt{MN}$ and the triangle inequality twice to this equation yields

$$[\lambda^* \sigma_0 (1 + c_3) - \theta]\sqrt{MN} - \lambda^* \|\sigma_0 \boldsymbol{P} + \boldsymbol{Q}\| \leq 1 \leq [\lambda^* \sigma_0 (1 + c_3) - \theta]\sqrt{MN} + \lambda^* \|\sigma_0 \boldsymbol{P} + \boldsymbol{Q}\|.$$

Applying (19) yields

$$[\lambda^* \sigma_0 (1 + c_3 - c_5) - \theta]\sqrt{MN} \leq 1 \leq [\lambda^* \sigma_0 (1 + c_3 + c_5) - \theta]\sqrt{MN}.$$

Rearranging this chain of inequalities and using the fact that $1 + c_3 - c_5 > 0$, which follows from (12) stated in the theorem, yields

$$\frac{1 + \theta\sqrt{MN}}{\sigma_0 (1 + c_3 + c_5)\sqrt{MN}} \leq \lambda^* \leq \frac{1 + \theta\sqrt{MN}}{\sigma_0 (1 + c_3 - c_5)\sqrt{MN}}, \tag{21}$$

with high probability.

We wish to establish that $\lambda^*\sigma_0 - \theta \geq 0$. Using the left inequality in (21) yields:

$$\begin{aligned}\lambda^*\sigma_0 - \theta &\geq \frac{1+\theta\sqrt{MN}}{(1+c_3+c_5)\sqrt{MN}} - \theta \\ &= \frac{1-(c_3+c_5)\theta\sqrt{MN}}{(1+c_3+c_5)\sqrt{MN}}.\end{aligned}$$

Thus, nonnegativity of $\lambda^*\sigma_0 - \theta$ is implied by the inequality $\theta \leq 1/((c_3+c_5)\sqrt{MN})$, which is a consequence of assumption (14).

Since $\lambda^*\sigma_0 - \theta \geq 0$,

$$\lambda^*\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11} = (\lambda^*\sigma_0 - \theta)\boldsymbol{e}_M\boldsymbol{e}_N^T + \lambda^*(\sigma_0\boldsymbol{P} + \boldsymbol{R}_{11}) > \boldsymbol{0}.$$

Applying the Perron-Frobenius theorem, we obtain the positivity of $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$.

We also need $\|\boldsymbol{W}_{11}\| \leq 1/2$. Recall $\|\boldsymbol{W}_{11}\| = \sigma_2(\lambda^*\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11})$, the second largest singular value of $\lambda^*\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11}$, since $\lambda^*\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11} = \boldsymbol{u}_1\boldsymbol{v}_1^T + \boldsymbol{W}_{11}$. Using the well-known fact that

$$\sigma_2(\boldsymbol{A}) = \min\{\|\boldsymbol{A} - \boldsymbol{S}\| : \mathrm{rank}(\boldsymbol{S}) \leq 1\},$$

we obtain

$$\|\boldsymbol{W}_{11}\| \leq \|\lambda^*(\sigma_0\boldsymbol{P} + \boldsymbol{Q})\|.$$

Here we selected $\boldsymbol{S}$ to be $[\lambda^*\sigma_0(1+c_3) - \theta]\boldsymbol{e}_M\boldsymbol{e}_N^T$. With high probability, we obtain the bound $\|\boldsymbol{W}_{11}\| \leq \lambda^*\sigma_0 c_5\sqrt{MN}$ from (19).

Using the upper bound on λ^* from (21), we have:

$$\|\boldsymbol{W}_{11}\| \leq \frac{(1+\theta\sqrt{MN})c_5}{1+c_3-c_5}.$$

In order to obtain $\|\boldsymbol{W}_{11}\| \leq 1/2$, a sufficient condition is

$$\frac{(1+\theta\sqrt{MN})c_5}{1+c_3-c_5} \leq \frac{1}{2} \tag{22}$$

which is rearranged as

$$\theta \leq \left(\frac{1+c_3-3c_5}{2c_5}\right)\frac{1}{\sqrt{MN}}.$$

The latter inequality follows from (14); the numerator of the right-hand side is positive by (12).

Turning to (17) when $(i,j) = (2,2)$, we need to find $\boldsymbol{W}_{22}$ and $\boldsymbol{V}_{22}$ that satisfy

$$\lambda^*\boldsymbol{R}_{22} = \boldsymbol{W}_{22} + \theta\boldsymbol{V}_{22},$$

$\|\boldsymbol{W}_{22}\| \leq 1/2$, and $\|\boldsymbol{V}_{22}\|_\infty \leq 1$. Consider the assignment $\boldsymbol{V}_{22} = \dfrac{\lambda^*\sigma_0 c_3}{\theta}\boldsymbol{e}_{m-M}\boldsymbol{e}_{n-N}^T$ and $\boldsymbol{W}_{22} = \lambda^*\boldsymbol{R}_{22} - \theta\boldsymbol{V}_{22}$. The coefficient $\lambda^*\sigma_0 c_3/\theta$ is chosen for the definition of $\boldsymbol{V}_{22}$ so that the entries of the remainder term $\boldsymbol{W}_{22}$ have mean zero.

The requirement $\|\boldsymbol{V}_{22}\|_\infty \leq 1$ is satisfied if and only if $\lambda^*\sigma_0 c_3/\theta \leq 1$. Because of the upper bound on λ^* established by (21), this requirement is satisfied if

$$\frac{(1+\theta\sqrt{MN})c_3}{\theta(1+c_3-c_5)\sqrt{MN}} \leq 1 \Leftrightarrow \theta \geq \left(\frac{c_3}{1-c_5}\right)\frac{1}{\sqrt{MN}}, \quad c_5 < 1.$$

This inequality is assured by (12) and (15). (In particular, (12) implies $c_5 < 1$.)

To bound $\|\boldsymbol{W}_{22}\|$, consider $\boldsymbol{W}_{22}/(\lambda^*\sigma_0)$, which is a random matrix with i.i.d. elements that are b-subgaussian. Applying Lemma 5(i) to $\boldsymbol{W}_{22}/(\lambda^*\sigma_0)$ and taking $u = 1/(2\lambda^*\sigma_0)$ yields

$$\mathbb{P}\left(\|\boldsymbol{W}_{22}\| \geq 1/2\right) \leq \exp\left(-\left(\frac{2}{81b^2(\lambda^*\sigma_0)^2} - (\log 7)(m-M+n-N)\right)\right).$$

Use the upper bound on λ^* from (21) to obtain the following tail bound:

$$\mathbb{P}\left(\|\boldsymbol{W}_{22}\| \geq \frac{1}{2}\right) \leq \exp\left(-\left(\frac{2(1+c_3-c_5)^2}{81b^2(1+\theta\sqrt{MN})^2}MN - (\log 7)(m-M+n-N)\right)\right).$$

From (22) we obtain

$$\frac{(1+c_3-c_5)^2}{(1+\theta\sqrt{MN})^2} \geq 4c_5^2 \tag{23}$$

hence

$$\mathbb{P}\left(\|\boldsymbol{W}_{22}\| \geq \frac{1}{2}\right) \leq \exp\left(-\left(\frac{8c_5^2 MN}{81b^2} - (\log 7)(m-M+n-N)\right)\right). \tag{24}$$

Now consider (17) when $(i,j) = (1,2)$. Again we need to find $\boldsymbol{W}_{12}$ and $\boldsymbol{V}_{12}$ such that

$$\lambda^*\boldsymbol{R}_{12} = \boldsymbol{W}_{12} + \theta\boldsymbol{V}_{12},$$

$\|\boldsymbol{W}_{12}\| \leq 1/2$, $\|\boldsymbol{V}_{12}\|_\infty \leq 1$, and $\boldsymbol{W}_{12}^T\boldsymbol{u}_1 = \boldsymbol{0}$. We construct $\boldsymbol{W}_{12}$ and $\boldsymbol{V}_{12}$ column by column as follows:

$$\boldsymbol{V}_{12}(:,i) \quad = \quad \frac{\lambda^*\boldsymbol{R}_{12}(:,i)^T\boldsymbol{u}_1}{\theta\|\boldsymbol{u}_1\|_1}\boldsymbol{e}_M, \quad \boldsymbol{W}_{12}(:,i) = \lambda^*\boldsymbol{R}_{12}(:,i) - \theta\boldsymbol{V}_{12}(:,i)$$

all $i = 1,\ldots,n-N$. By construction we have $\boldsymbol{W}_{12}(:,i)^T\boldsymbol{u}_1 = 0$ for all $i = 1,\ldots,n-N$. Now consider the requirement that $\|\boldsymbol{V}_{12}(:,i)\|_\infty \leq 1$ for all $i = 1,\ldots,n-N$. The requirement is equivalent to

$$\frac{\lambda^*\boldsymbol{R}_{12}(:,i)^T\boldsymbol{u}_1}{\theta\|\boldsymbol{u}_1\|_1} \leq 1$$

for all $i = 1,\ldots,n-N$. Subtract $\lambda^* c_3\sigma_0/\theta$ from both sides and apply the identity $\boldsymbol{e}_M^T\boldsymbol{u}_1 = \|\boldsymbol{u}_1\|_1$ to obtain

$$\frac{\lambda^*(\boldsymbol{R}_{12}(:,i)^T - c_3\sigma_0\boldsymbol{e}_M^T)\boldsymbol{u}_1}{\theta\|\boldsymbol{u}_1\|_1} \leq 1 - \frac{\lambda^* c_3\sigma_0}{\theta}.$$

We will establish this inequality in two steps. First, we establish that $\lambda^* c_3 \sigma_0/\theta \leq 1/2$. Because of (21), it suffices to establish that

$$\frac{c_3(1+\theta\sqrt{MN})}{\theta(1+c_3-c_5)\sqrt{MN}} \leq \frac{1}{2}. \tag{25}$$

This can be rearranged into

$$\theta \geq \frac{2c_3}{(1-c_3-c_5)\sqrt{MN}},$$

which follows from (15).

Second, we establish that with probability exponentially close to 1,

$$\frac{\lambda^*(\boldsymbol{R}_{12}(:,i)^T - c_3\sigma_0 \boldsymbol{e}_M^T)\boldsymbol{u}_1}{\theta \left\|\boldsymbol{u}_1\right\|_1} \leq \frac{1}{2}.$$

Notice that $r_{ji}/\sigma_0 - c_3$ is b-subgaussian; thus,

$$\frac{1}{\sigma_0}\left(\boldsymbol{R}_{12}(:,i) - c_3\sigma_0\boldsymbol{e}_M\right)^T \boldsymbol{u}_1 = \frac{1}{\sigma_0}\sum_{j=1}^{M} u_1(j)\left(R_{12}(j,i) - c_3\sigma_0\right)$$

is also b-subgaussian since $\left\|\boldsymbol{u}_1\right\|_2 = 1$. Thus, by (10), taking $x = \left(\boldsymbol{R}_{12}(:,i) - c_3\sigma_0\boldsymbol{e}_M\right)^T \boldsymbol{u}_1/\sigma_0$ and taking $t = \theta\|\boldsymbol{u}_1\|_1/(2\lambda^*\sigma_0)$,

$$\begin{aligned}\mathbb{P}\left(\frac{\lambda^*(\boldsymbol{R}_{12}(:,i)^T - c_3\sigma_0\boldsymbol{e}_M)\boldsymbol{u}_1}{\theta\left\|\boldsymbol{u}_1\right\|_1} > \frac{1}{2}\right) &\leq \exp\left(-\theta^2\|\boldsymbol{u}_1\|_1^2/(8b^2(\lambda^*\sigma_0)^2)\right). \\ &\leq \exp\left(-c_3^2\|\boldsymbol{u}_1\|_1^2/(2b^2)\right). \end{aligned} \tag{26}$$

since, as noted above $\theta/(\lambda^*\sigma_0) \geq 2c_3$.

To proceed, we now need a lower bound for $\left\|\boldsymbol{u}_1\right\|_1$. Let $\boldsymbol{F}$ denote $\lambda^*\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11}$, which is equal to $[\lambda^*\sigma_0(1+c_3)-\theta]\boldsymbol{e}_M\boldsymbol{e}_N^T + \lambda^*(\sigma_0\boldsymbol{P}+\boldsymbol{Q})$. We know that $\boldsymbol{u}_1$ is the first (left) singular vector of $\boldsymbol{F}$. Letting $\boldsymbol{X}_0 = [\lambda^*\sigma_0(1+c_3)-\theta]\boldsymbol{e}_M\boldsymbol{e}_N^T$ and $\boldsymbol{E} = \lambda^*(\sigma_0\boldsymbol{P}+\boldsymbol{Q})$, we then have $\boldsymbol{F} = \boldsymbol{X}_0 + \boldsymbol{E}$, and $\boldsymbol{X}_0$ is a rank-one matrix with a single nonzero singular value equal to $(\lambda^*\sigma_0(1+c_3)-\theta)\sqrt{MN}$ and with left singular vector $\boldsymbol{e}_M/\sqrt{M}$ and right singular vector $\boldsymbol{e}_N/\sqrt{N}$. Furthermore, since $\|\boldsymbol{F}\| = 1$, we know that the singular value of $\boldsymbol{X_0}$ is at least $1 - \|\boldsymbol{E}\|$ by Corollary 8.6.2 of [11]. Thus, by Theorem 8.6.5 of [11],

$$\left\|\boldsymbol{u}_1 - \frac{\boldsymbol{e}_M}{\sqrt{M}}\right\| \leq \frac{4\|\boldsymbol{E}\|}{1-\|\boldsymbol{E}\|} \leq 4/5, \tag{27}$$

provided that $\|\boldsymbol{E}\| \leq 1/6$. Thus, the next step in the analysis is to show that $\|\boldsymbol{E}\| \leq 1/6$. This follows from the following sequence of inequalities:

$$\begin{aligned}\|\boldsymbol{E}\| &= \lambda^*\|\sigma_0\boldsymbol{P}+\boldsymbol{Q}\| \\ &\leq \frac{c_5(1+\theta\sqrt{MN})}{1+c_3-c_5} \\ &\leq 1/6,\end{aligned}$$

where the second line holds with high probability according to (19) and the third line follows from (22) and (12).

Thus, we have established that $\|\boldsymbol{E}\| \leq 1/6$ with high probability, which in turn implies that

$$\begin{aligned}
\|\boldsymbol{u}_1\|_1 &= \boldsymbol{e}_M^T \boldsymbol{u}_1 \\
&= (\boldsymbol{e}_M - M^{1/2}\boldsymbol{u}_1 + M^{1/2}\boldsymbol{u}_1)^T \boldsymbol{u}_1 \\
&\geq M^{1/2}\boldsymbol{u}_1^T\boldsymbol{u}_1 - |(\boldsymbol{e}_M - M^{1/2}\boldsymbol{u}_1)^T\boldsymbol{u}_1| \\
&\geq M^{1/2}\boldsymbol{u}_1^T\boldsymbol{u}_1 - \|\boldsymbol{e}_M - M^{1/2}\boldsymbol{u}_1\| \cdot \|\boldsymbol{u}_1\| \\
&= M^{1/2} - M^{1/2}\|M^{-1/2}\boldsymbol{e}_M - \boldsymbol{u}_1\| \\
&\geq M^{1/2} - (4/5)M^{1/2},
\end{aligned}$$

where the last line is obtained from (27). This gives a lower bound of $M^{1/2}/5$ on $\|\boldsymbol{u}\|_1$. Thus, substituting this into (26) yields

$$\mathbb{P}\left(\frac{\lambda^*(\boldsymbol{R}_{12}(:,i)^T - c_3\sigma_0\boldsymbol{e}_M)\boldsymbol{u}_1}{\theta\,\|\boldsymbol{u}_1\|_1} > \frac{1}{2}\right) \leq \exp\left(-\sigma_0^2 c_3^2 M/(50b^2)\right).$$

This shows that one column of $\boldsymbol{V}_{12}$ exceeds norm 1/2 with exponentially small probability. Applying the union bound over all the columns, we find

$$\mathbb{P}\left(\|\boldsymbol{V}_{12}\|_\infty \geq 1\right) \leq (n-N)\exp\left(-\sigma_0^2 c_3^2 M/(50b^2)\right). \tag{28}$$

Thus, we have established that $\|\boldsymbol{V}_{12}\|_\infty \leq 1$ with probability exponentially close to 1.

We now consider the matrix $\boldsymbol{W}_{12}$, which can be written as $\boldsymbol{W}_{12} = \lambda^*\boldsymbol{D}\boldsymbol{R}_{12}$, where $\boldsymbol{D} \in \mathbb{R}^{M\times M}$ is given by

$$\boldsymbol{D} = \boldsymbol{I} - \frac{1}{\|\boldsymbol{u}_1\|_1}\boldsymbol{e}_M\boldsymbol{u}_1^T.$$

Notice that we can equivalently write

$$\boldsymbol{W}_{12} = \lambda^*\boldsymbol{D}(\boldsymbol{R}_{12} - c_3\sigma_0\boldsymbol{e}_M\boldsymbol{e}_{n-N}^T),$$

since $\boldsymbol{D}\boldsymbol{e}_M = \boldsymbol{0}$. The matrix $\boldsymbol{R}_{12} - c_3\sigma_0\boldsymbol{e}_M\boldsymbol{e}_{n-N}^T$ is a subgaussian matrix scaled by σ_0. Furthermore,

$$\|\boldsymbol{D}_{12}\| \leq 1 + \frac{\sqrt{M}}{\|\boldsymbol{u}\|_1} \leq 6,$$

with high probability, since $\|\boldsymbol{u}_1\|_1 \geq M^{1/2}/5$. Thus, Lemma 5(ii) applied to $\boldsymbol{W}_{12}/(\lambda^*\sigma_0)$, taking $u = 1/(2\lambda^*\sigma_0)$, yields

$$\mathbb{P}\left(\|\boldsymbol{W}_{12}\| \geq 1/2\right) \leq \exp\left(-\left(\frac{2}{36\cdot 81b^2\sigma_0^2(\lambda^*)^2} - (\log 7)(M+n-N)\right)\right).$$

Apply the upper bound on λ^* from (21) to obtain

$$\mathbb{P}\left(\|\boldsymbol{W}_{12}\| \geq 1/2\right) \leq \exp\left(-\left(\frac{2(1+c_3-c_5)^2 MN}{36 \cdot 81 b^2 (1+\theta\sqrt{MN})^2} - (\log 7)(M+n-N)\right)\right).$$

Now finally we apply (23) to obtain

$$\mathbb{P}\left(\|\boldsymbol{W}_{12}\| \geq 1/2\right) \leq \exp\left(-\left(\frac{8c_5^2 MN}{36 \cdot 81 b^2} - (\log 7)(M+n-N)\right)\right). \tag{29}$$

The same construction and analysis applies to $\boldsymbol{V}_{21}$ and $\boldsymbol{W}_{21}$, and the same results are obtained except with the roles of (M, m) and (N, n) interchanged. Thus,

$$\mathbb{P}\left(\|\boldsymbol{V}_{21}\|_\infty \geq 1\right) \leq (m-M)\exp\left(-\sigma_0^2 c_3^2 N/(50b^2)\right), \tag{30}$$

and

$$\mathbb{P}\left(\|\boldsymbol{W}_{21}\| \geq 1/2\right) \leq \exp\left(-\left(\frac{8c_5^2 MN}{36 \cdot 81 b^2} - (\log 7)(N+m-M)\right)\right). \tag{31}$$

From the analysis of all four blocks of $\boldsymbol{V}$ and $\boldsymbol{W}$, we have:

$$\|\boldsymbol{V}\|_\infty = \max\left\{\|\boldsymbol{V}_{11}\|_\infty, \|\boldsymbol{V}_{12}\|_\infty, \|\boldsymbol{V}_{21}\|_\infty, \|\boldsymbol{V}_{22}\|_\infty\right\} \leq 1,$$

where $\boldsymbol{V}_{11} = \boldsymbol{e}_M \boldsymbol{e}_N^T$. With a high probability, we also have $\|\boldsymbol{W}\| \leq 1$ since

$$\|\boldsymbol{W}\|^2 \leq \|\boldsymbol{W}_{11}\|^2 + \|\boldsymbol{W}_{12}\|^2 + \|\boldsymbol{W}_{22}\|^2 + \|\boldsymbol{W}_{21}\|^2 \leq 1.$$

By the union bound, the probability of failure of the main result is at most the sum of the probabilities of the failure at each step. Therefore, the failure of the convex relaxation to find the claimed optimal $\boldsymbol{X}$ is at most the sum of the right-hand sides of (18), (24), (30), (29), (28), and (31). We require these probabilities to be exponentially small. We assure that (18) is exponentially small by requiring

$$MN \geq k_1 (M+N)^{4/3} \tag{32}$$

where

$$k_1 > \left((\log 7) 81 b^2\right)^{4/3}. \tag{33}$$

Next, all of (24), (29), (31) are exponentially small provided that

$$MN \geq k_2 (m+n) \tag{34}$$

where

$$k_2 > \frac{(\log 7) 36 \cdot 81 b^2}{8c_5^2}. \tag{35}$$

Finally, to ensure that (28) and (30) tend to 0 exponentially fast requires that M grow as fast as $\Omega(\log(n-N))$ and similarly N grows as fast as $\Omega(\log(m-M))$, but this is already a consequence of (32) and (34).

5 Conclusions

We have shown that a convex relaxation can find a large, approximately rank-one submatrix of a much larger noisy matrix provided that the dimensions of the larger matrix are no larger than the square of the dimensions of the smaller matrix, and provided certain upper bounds are satisfied on the level of the noise.

It is interesting to note that our result also applies to the maximum biclique problem, which was introduced in Section 1 as a special case of LAROS. In particular, if G is a bipartite graph (U, V, E) containing a biclique given by $U^* \times V^*$, where $|U| = m$, $|V| = n$, $|U^*| = M$, $|V^*| = N$, and if the remaining edges of E (i.e., those not in $U^* \times V^*$) are inserted at random with probability $1/2$, then the U-to-V adjacency matrix has the form (11) in which $\sigma = 1$, $c_1 = c_2 = 0$, $c_3 = 1/2$, $b = 1/(8 \log 2)^{1/2}$. (This is not quite correct since in this case $\boldsymbol{R}_{11} = \mathbf{0}$. However, our analysis covers this case as well.) Thus, our algorithm with parameter $\theta = O(1/(MN)^{1/2})$ finds the planted biclique when $M \sim N$, $m \sim n$, and $M \geq \Omega(m^{1/2})$. The same result was obtained earlier by Ames and Vavasis [1] using a different convex relaxation. Theirs has the advantage that M, N do not need to be known or estimated in advance, but ours solves a more general class of problems.

References

[1] B. Ames and S. Vavasis. Nuclear norm minimization for the planted clique and biclique problems. Under review, Math. Prog., URL: `http://arxiv.org/abs/0901.3348`, 2009.

[2] N. Asgarian and R. Greiner. Using rank-1 biclusters to classify microarray data. Technical report, Department of Computing Science and the Alberta Ingenuity Center for Machine Learning, University of Alberta, Edmonton, AB, Canada, 2008.

[3] S. Bergmann, J. Ihmels, and N. Barkai. Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review E*, 67:031902, 2003.

[4] Michael Biggs, Ali Ghodsi, and Stephen A. Vavasis. Nonnegative matrix factorization via rank-one downdating. In *Proceedings of the 2008 International Conference on Machine Learning*, 2008. Proceedings published online at `http://icml2008.cs.helsinki.fi/abstracts.shtml`, Preliminary version of the full paper in arxiv.org, 0808.0120.

[5] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

[6] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[7] V. Chandrasekaran, S. Sanghavi, P.A. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. See `http://arxiv.org/abs/0906.2220`, 2009.

[8] J. Cohen and U. Rothblum. Nonnegative ranks, decompositions and factorizations of nonnegative matrices. *Linear Algebra and its Applications*, 190:149–168, 1993.

[9] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, Stanford, CA, 2001.

[10] N. Gillis and F. Glineur. Using underapproximations for sparse nonnegative matrix factorization. *Pattern Recognition*, 43:1676–1687, 2010.

[11] G. H. Golub and C. F. Van Loan. *Matrix Computations, 3rd Edition*. Johns Hopkins University Press, Baltimore, 1996.

[12] T. Hofmann. Probabilistic latent semantic analysis. In Kathryn B. Laskey and Henri Prade, editors, *UAI '99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, July 30-August 1, 1999*, pages 289–296. Morgan Kaufmann, 1999.

[13] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[14] A. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann. Smallest singular values of random matrices and geometry of random polytopes. *Advances in Mathematics*, 195:491–523, 2005.

[15] R. Peeters. The maximum edge biclique problem is NP-complete. *Discrete Applied Mathematics*, 131:651–654, 2003.

[16] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[17] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[18] L. B. Thomas. Problem 73-14, rank factorization of nonnegative matrices by Berman and Plemmons. *SIAM Review*, 16:393–394, 1974.

[19] Stephen A. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM J. Optim.*, 20(3):1364–1377, 2009.

[20] K. Zięţak. Properties of linear approximations of matrices in the spectral norm. *Linear Algebra Applications*, 183:41–60, 1993.