Spurious Valleys, Spurious Minima and NP-hardness of Sparse Matrix Factorization With Fixed Support

Quoc-Tung Le, Elisa Riccietti, Rémi Gribonval

Spurious Valleys, Spurious Minima and NP-hardness of Sparse Matrix Factorization With Fixed Support

Quoc-Tung Le **Elisa Riccietti** **Rémi Gribonval**
Univ Lyon, ENS de Lyon, UCBL, CNRS, Inria, LIP, F-69342, LYON Cedex 07, France.

Abstract

The problem of approximating a dense matrix by a product of sparse factors is a fundamental problem for many signal processing and machine learning tasks. It can be decomposed into two subproblems: finding the position of the non-zero coefficients in the sparse factors, and determining their values. While the first step is usually seen as the most challenging one due to its combinatorial nature, this paper focuses on the second step, referred to as sparse matrix approximation with fixed support. First, we show its NP-hardness, while also presenting a nontrivial family of supports making the problem practically tractable with a dedicated algorithm. Then, we investigate the landscape of its natural optimization formulation, proving the absence of spurious local valleys and spurious local minima, whose presence could prevent local optimization methods to achieve global optimality. The advantages of the proposed algorithm over state-of-the-art first-order optimization methods are discussed.

1 Introduction

Matrix factorization with sparsity constraints is the problem of approximating a (possibly dense) matrix as a product of two or more sparse factors. This is playing an important role in many domains and applications such as dictionary learning and signal processing [19, 17, 16], linear operator acceleration [11, 10], deep learning [2], to mention only a few.

In this work, we consider a particular instance of the matrix factorization problem with sparsity constraints, the case in which just two factors are considered, which have a prescribed support. In details, given a matrix $A \in \mathbb{R}^{m\times n}$, we look for two sparse factors X, Y such that:

$$\begin{aligned} \underset{X\in\mathbb{R}^{m\times r}, Y\in\mathbb{R}^{n\times r}}{\text{Minimize}} \quad & L(X,Y) = \|A - XY^\top\|^2 \\ \text{Subject to:} \quad & \texttt{supp}(X) \subseteq I \text{ and } \texttt{supp}(Y) \subseteq J \end{aligned} \tag{1}$$

where $\|\cdot\|$ is the Frobenius norm, $I \subseteq [\![m]\!] \times [\![r]\!]$, $J \subseteq [\![n]\!] \times [\![r]\!]$[1] are given support constraints, i.e., $\texttt{supp}(X) \subseteq I$ implies that $\forall (i,j) \notin I, X_{ij} = 0$. We call Problem (1) *sparse matrix factorization with fixed support*.

The main aim of this work is to investigate the theoretical properties of Problem (1). To the best of our knowledge the analysis of matrix factorization problems with fixed supports has never been addressed in the literature. This analysis is however interesting, for at least two reasons.

First of all, there are many practical applications in which the solution of this problem is required. Indeed, there are matrices that can be written as the product of factors whose support is known in advance. This is the case for instance of many fast transforms such as the Discrete Fourier Transform (DFT) or the the Hadamard Transform (HT), in which the fixed supports of the factors have the butterfly structure [10, 2].

[1] $[\![m]\!] := \{1, \dots, m\}$

Preprint. Under review.

Moreover, Problem (1) can be seen as a subproblem of a more general matrix factorization problem with structured sparsity constraints:

$$\begin{aligned} \underset{X\in\mathbb{R}^{m\times r},Y\in\mathbb{R}^{n\times r}}{\text{Minimize}} \quad & L(X,Y) = \|A - XY^\top\|^2 \\ \text{Subject to:} \quad & X \in \Sigma_X \text{ and } Y \in \Sigma_Y \end{aligned} \tag{2}$$

where $\Sigma_X \subseteq \mathbb{R}^{m\times r}$, $\Sigma_Y \subseteq \mathbb{R}^{n\times r}$ are some sets of structured sparse matrices. Relevant examples of such sets are for instance the sets of matrices with:

- at most k non-zero coefficients $\Sigma_k^{\texttt{total}} = \{X \in \mathbb{R}^{m\times r} \mid \|X\|_0 \leq k\}$;
- at most k non-zero coefficients per column $\Sigma_k^{\texttt{col}} = \{X \in \mathbb{R}^{m\times r} \mid \|X_{\bullet,i}\|_0 \leq k, \forall i = 1,\dots,r\}$;
- at most k non-zero coefficients per row $\Sigma_k^{\texttt{row}} = \{X \in \mathbb{R}^{m\times r} \mid \|X_{i,\bullet}\|_0 \leq k, \forall i = 1,\dots,m\}$;

where for a vector or matrix X, $\|X\|_0$ counts the number of nonzero entries in X.

Any heuristic algorithm for the solution of (2) will eventually need to deal with a subproblem of the form (1), one way or another. Indeed, matrix factorization with sparsity constraints somehow generalizes the sparse recovery problem [4], in which we want to recover a sparse vector $x \in \mathbb{R}^n$ from the knowledge of its measurement vector (possibly corrupted by noise) $y = Ax \in \mathbb{R}^m$ with known measurement matrix $A \in \mathbb{R}^{m\times n}$. Mimicking the decomposition of the classical sparse recovery problem into a support recovery step and a coefficient recovery step, Problem (2) can also be split into two subproblems:

1) Determine the support of the factors X and Y, i.e. the set of indices $\texttt{supp}(X)$, $\texttt{supp}(Y)$ whose coefficients are different from zero. For instance, if $\Sigma_X = \Sigma_Y = \Sigma_k^{\texttt{total}}$, we need to identify the position of (at most) k non-zero coefficients of X and Y.
2) Determine the value of the coefficients in the supports of X and Y.

The solution of a problem in the form of (1) will be needed both for *one-step* algorithms that jointly estimate the supports and coefficients, and for the *two-step* algorithms that solve the two problems successively. Also, as it happens in sparse linear regresssion, many common post-processing methods consist in "debiasing" the solution by a two-step approach [2].

Our aim is to then study the theoretical properties of Problem (1) and in particular to assess its difficulty. Assessing the difficulty of this subproblem is crucial to have a good understanding also of the difficulty of the full problem (2).

In particular, we consider three complementary aspects related to Problem (1).

First, we show the NP-hardness of Problem (1). While this result contrasts with the theory established for coefficient recovery with a fixed support in the classical sparse recovery problem (that can be trivially addressed by least squares), it is in line with the known hardness of related matrix factorization with additional constraints or different losses. Indeed, famous variants of matrix factorization such as non-negative matrix factorization (NMF) [21, 18], weighted low rank [5] and matrix completion [5] were all proved to be NP-hard. We prove the NP-hardness by reduction from the Low Rank Matrix Completion problem with noise. To our knowledge this proof is new and cannot be trivially deduced from any existing result on the more classical full support case (i.e., the case in which $I = [\![m]\!] \times [\![r]\!], J = [\![n]\!] \times [\![r]\!]$, which is equivalent to low rank matrix approximation [3]).

Second, we show that despite the hardness of Problem (1) in the general case, many pairs of support constraints (I, J) make the problem solvable by an effective direct algorithm based on the singular value decomposition (SVD). The investigation of those supports is also covered in this work and a dedicated polynomial algorithm is proposed to deal with this family of supports. This includes for example the full support case. Our analysis of tractable instances of Problem (1) actually includes and substantially generalizes the analysis of the instances that can be classically handled with the SVD decomposition. In fact, the presence of the constraints on the support makes it impossible to directly use the SVD to solve the problem, because coefficients outside the support have to be zero. However, the presented family of support constraints allows for an iterative decomposition of the problem into "blocks" that can be exploited to build up a solution of the corresponding instances of Problem (2) using blockwise SVDs.

The third contribution of this paper is the study of the landscape of function L of Problem (1), notably we investigate the existence of *spurious local minima* and *spurious local valleys*, which will be collectively referred to as *spurious objects*. They will be formally introduced in Section 4, but intuitively these objects may represent a challenge for the convergence of local optimization methods.

The landscape of the loss functions for neural networks in general, and for linear neural networks in particular, has been a popular subject of study recently. In particular, great attention has been devoted to the investigation of the properties of critical points and global optima of the training problem with quadratic loss [7, 23, 12]. These works have direct link to ours since matrix factorization (without any constraint) can be seen as a specific case of neural network (with two layers, no bias and linear activation function). Notably it has been proved [23] that for linear neural networks, every local minimum is a global minimum and if the network is shallow (i.e., there is only one hidden layer), critical points are either global minima or strict saddle points (i.e., their Hessian have at least one negative eigenvalue). However, there is still a *tricky* type of landscape that could represent a challenge for local optimization methods and has not been covered until recently: spurious local valleys [13, 22].

To the best of our knowledge, existing analyses of spurious local valleys are proposed for matrix factorization problems without support constraints, cf. [23, 22, 7], while the study of the landscape of Problem (1) remains untouched in the literature and our work can be considered as a generalization of such previous results.

To summarize, our main contributions in this paper are:

1. We prove that Problem (1) is NP-hard in Theorem 2.3.
2. We introduce families of support constraints (I, J) making Problem (1) tractable (Theorem 3.1 and Theorem 3.3).
3. We show that the landscape of Problem (1) corresponding to the support pairs (I, J) in these families are free of spurious local valleys, regardless of the factorized matrix A (Theorem 4.1, Theorem 4.2). We also investigate the presence of spurious local minima for such families (Theorem 4.1, Theorem 4.3).
4. These results might suggest a conjecture, that holds true for the full support case: that an instance of Problem (1) is tractable if and only if their corresponding landscape is benign, i.e. free of spurious objects. We give a counter-example to this conjecture (Example 4.2) and show experimentally that first-order methods for fixed support matrix factorization problem can fail despite a benign landscape and that a good initialization is really important.
5. We propose an algorithm to solve these tractable instances of Problem (1), and discuss the advantages of our approach over state-of-the-art first order optimization methods.

In the paper we report only the main results, all the proofs can be found in the supplementary material.

1.1 Notations

For $n \in \mathbb{N}$, the set $\{1, \dots, n\}$ is denoted by $[\![n]\!]$. The notations $\mathbf{0}$ and $\mathbf{1}$ stand for the matrices whose coefficients are all zeros and ones respectively. The identity matrix of size $n \times n$ is denoted by $\mathbf{I}_n$. Given a matrix $A \in \mathbb{R}^{m \times n}$ and $T \subseteq [\![n]\!]$, $A_{\bullet,T} \in \mathbb{R}^{m \times |T|}$ is the submatrix of A restrained to the columns indexed in T. If $T = \{k\}$ is a singleton, $A_{\bullet,T}$ is simplified as $A_{\bullet,k}$ (the k^{th} column of A). For $(i, j) \in [\![m]\!] \times [\![n]\!]$, $A_{i,j}$ is the coefficient of A at index (i, j).

A support constraint I on a matrix $X \in \mathbb{R}^{m \times r}$ can be interpreted either as a subset $I \subseteq [\![m]\!] \times [\![r]\!]$ or as its indicator matrix $1_I \in \{0, 1\}^{m \times r}$ defined as: $(1_I)_{i,j} = 1$ if $(i, j) \in I$, 0 otherwise. Both representations will be used interchangeably and the meaning should be clear from the context.

The notation $\mathtt{supp}(A)$ is used for both vector and matrix: if $A \in \mathbb{R}^m$ is a vector, then $\mathtt{supp}(A) = \{i \mid A_i \neq 0\} \subseteq [\![m]\!]$; if $A \in \mathbb{R}^{m \times n}$ is a matrix, then $\mathtt{supp}(A) = \{(i, j) \mid A_{i,j} \neq 0\} \subseteq [\![m]\!] \times [\![n]\!]$. Given two matrices $A, B \in \mathbb{R}^{m \times n}$, the Hadamard product $A \odot B$ between A and B is defined as $(A \odot B)_{i,j} = A_{i,j} B_{i,j}, \forall (i, j) \in [\![m]\!] \times [\![n]\!]$. Since a support constraint I of a matrix X can be thought of as a binary matrix of the same size, we define $X \odot I := X \odot 1_I$ analogously (it is a matrix whose coefficients in I are unchanged while the others are set to zero).

2 Matrix factorization with fixed support is NP-hard

To show that Problem (1) is NP-hard we use the classical technique to prove NP-hardness: reduction. Our choice of reducible problem is Low Rank Matrix Completion with noise [5].

Definition 2.1 (Low rank matrix completion with noise [5])**.** *Let $W \in \{0,1\}^{m\times n}$ be a binary matrix. Given $A \in \mathbb{R}^{m\times n}$ and $s \in \mathbb{N}$, the matrix completion problem (MCP) is:*

$$\underset{X\in\mathbb{R}^{m\times s}, Y\in\mathbb{R}^{n\times s}}{\text{Minimize}} \|A - XY^\top\|_W^2 = \|(A - XY^\top)\odot W\|^2. \tag{3}$$

This problem is proved to be NP-hard when $s = 1$ [5] by its reducibility from Maximum-Edge Biclique Problem, which is NP-complete [15]. This is expressed in the following theorem:

Theorem 2.1 (NP-hardness of matrix completion with noise [5])**.** *Given a binary weighing matrix $W \in \{0,1\}^{m\times n}$ and $A \in [0,1]^{m\times n}$, the rank-one matrix completion problem is:*

$$\underset{x\in\mathbb{R}^m, y\in\mathbb{R}^n}{\text{Minimize}} \|A - xy^\top\|_W^2. \tag{4}$$

Denote p^* the infimum of Problem (4) *and $\epsilon = 2^{-12}(mn)^{-7}$. It is NP-hard to find an approximate solution to* (4) *with objective function accuracy less than ϵ, i.e. with objective value $p \leq p^* + \epsilon$.*

To prove that the problem of fixed support factorization is NP-hard, it is sufficient to construct an instance of Problem (1) which is equivalent to the MCP problem (3) with $s = 1$. The following lemma gives a reduction from MCP to fixed support matrix factorization.

Lemma 2.2. *For any binary matrix $W \in \{0,1\}^{m\times n}$, there exist an integer r and two sets I and J such that for all $A \in \mathbb{R}^{m\times n}$, Problem* (4) *and Problem* (1) *share the same infimum. I and J can be constructed in polynomial time. Moreover, if one problem has a known solution of accuracy ϵ, we can find a solution with the same accuracy for the other in polynomial time.*

Sketch of the proof. Up to a transposition, we can assume without loss of generality that $m \geq n$. Let $r = n + 1 = \min(m, n) + 1$. We define $I \in \{0,1\}^{m\times(n+1)}$ and $J \in \{0,1\}^{n\times(n+1)}$ as follows:

$$I_{i,j} = \begin{cases} 1 - W_{i,j} & \text{if } j \neq n \\ 1 & \text{if } j = n+1 \end{cases}, J_{i,j} = \begin{cases} 1 & \text{if } j = i \text{ or } j = n+1 \\ 0 & \text{otherwise} \end{cases}$$

This construction can clearly be made in polynomial time. We show in Appendix A that the two problems share the same infimum. □

Using Lemma 2.2, we obtain a result of NP-hardness for the problem of fixed support matrix factorization as follows.

Theorem 2.3. *When $A \in [0,1]^{m\times n}$, it is NP-hard to solve $\inf_{X,Y,\mathtt{supp}(X)\subseteq I,\mathtt{supp}(Y)\subseteq J} \|A - XY^\top\|^2$ with arbitrary index sets I, J and objective function accuracy less than $2^{-12}(mn)^{-7}$.*

We point out that, while the result is interesting on its own, for some applications, such as those arising in machine learning, the accuracy bound $O((mn)^{-7})$ may not be really appealing. We thus keep as an interesting open research direction to determine if some precision threshold exists that make the general problem easy.

3 Tractable instances of matrix factorization with fixed support

Even though matrix factorization with fixed support is generally NP-hard, when we consider the full support case $I = [\![m]\!] \times [\![r]\!], J = [\![n]\!] \times [\![r]\!]$ (i.e., no coefficients of X,Y are set to zero, they are all optimized), it is equivalent to *low rank matrix approximation* (LRMA) [3], which can be solved using the Singular Value Decomposition (SVD) [6] [2]. This section is devoted to enlarging the family of supports for which Problem (1) can be solved by an effective direct algorithm. We start with an important definition:

[2] Exact SVD is not polynomially tractable, yet it can be practically computed to machine precision in $O(mn^2)$ [8], see also [20, Lecture 31, page 236]. It is thus convenient to think of LRMA as polynomially solvable.

Definition 3.1 (Support of rank-one contribution)**.** *Given two support constraints $I \in \{0,1\}^{m\times r}$ and $J \in \{0,1\}^{n\times r}$ in the fixed support matrix factorization problem and $k \in [\![r]\!]$, we define the k^{th} rank-one contribution support $\mathcal{S}_k(I,J)$ (or in short, $\mathcal{S}_k$) as:*

$$\mathcal{S}_k(I,J) = I_{\bullet,k}J_{\bullet,k}^\top, \tag{5}$$

This can be seen either as: a tensor product: $\mathcal{S}_k \in \{0,1\}^{m\times n}$ is a binary matrix; or a Cartesian product: $\mathcal{S}_k$ is a set of matrix indices defined as $\texttt{supp}(I_{\bullet,k}) \times \texttt{supp}(J_{\bullet,k})$.

Given a pair of support constraints I, J, if $\texttt{supp}(X) \subseteq I, \texttt{supp}(Y) \subseteq J$, we have:

$$\texttt{supp}(X_{\bullet,k}Y_{\bullet,k}^\top) \subseteq \mathcal{S}_k,\ \forall k \in [\![r]\!].$$

Since $XY^\top = \sum_{k=1}^r X_{\bullet,k}Y_{\bullet,k}^\top$ the notion of contribution support $\mathcal{S}_k$ captures the constraint on the support of the k^{th} *rank-one contribution*, $X_{\bullet,k}Y_{\bullet,k}^\top$, of the matrix product $XY^\top$.

We can partition $[\![r]\!]$ in terms of equivalence classes of rank-one supports:

Definition 3.2 (Equivalence classes of rank-one supports, representative rank-one supports)**.** *Given $I \in \{0,1\}^{m\times r}$, $J \in \{0,1\}^{n\times r}$, define an equivalence relation on $[\![r]\!]$ as: $i \sim j$ if and only if $\mathcal{S}_i = \mathcal{S}_j$ (or equivalently $(I_{\bullet,i}, J_{\bullet,i}) = (I_{\bullet,j}, J_{\bullet,j})$). This yields a partition of $[\![r]\!]$ into equivalence classes.*

Denote $\mathcal{P}$ the collection of equivalence classes. For each class $P \in \mathcal{P}$ denote $\mathcal{S}_P$ a representative rank-one support, $R_P \subseteq [\![m]\!]$ and $C_P \subseteq [\![n]\!]$ the supports of rows and columns in $\mathcal{S}_P$, respectively. For every $k \in P$ we have $\mathcal{S}_k = \mathcal{S}_P$ and $I_{\bullet,k} = R_P$, $J_{\bullet,k} = C_P$.

For every $\mathcal{P}' \subseteq \mathcal{P}$ denote $\mathcal{S}_{\mathcal{P}'} = \cup_{P\in\mathcal{P}'}\mathcal{S}_P \subseteq [\![m]\!] \times [\![n]\!]$ and $\bar{\mathcal{S}}_{\mathcal{P}'} = ([\![m]\!] \times [\![n]\!])\backslash\mathcal{S}_{\mathcal{P}'}$.

A first simple sufficient condition ensuring the tractability of an instance of Problem (1) is as follows.

Theorem 3.1. *Consider $I \in \{0,1\}^{m\times r}$, $J \in \{0,1\}^{n\times r}$, and $\mathcal{P}$ the collection of equivalence classes of Definition 3.2. If the representative rank-one supports are pairwise disjoint, i.e., $\mathcal{S}_P \cap \mathcal{S}_{P'} = \emptyset$ for each distinct $P, P' \in \mathcal{P}$, then matrix factorization with fixed support is tractable for any $A \in \mathbb{R}^{m\times n}$.*

The proof of this theorem shows that, if the condition in Theorem 3.1 is satisfied, the function $L(X,Y)$ can be decomposed into a sum of functions which are instances of the LRMA problem. In this case we can reduce Problem (1) to the problem of finding low rank approximations of submatrices of the target matrix A and thus solve such instances in polynomial time, as we show in Algorithm 1. Given the target matrix $A \in \mathbb{R}^{m\times n}$ and the supports constraints $I \in \{0,1\}^{m\times r}$, $J \in \{0,1\}^{n\times r}$ that satisfy the condition in Theorem 3.1, Algorithm 1 returns two factors X, Y that solve Problem (1).

Algorithm 1 SVD for fixed support matrix factorization (under assumptions of Theorem 3.1)

procedure SVD_FSMF($A \in \mathbb{R}^{m\times n}, I \in \{0,1\}^{m\times r}, J \in \{0,1\}^{n\times r}$)
  Initialize $X = \mathbf{0}, Y = \mathbf{0}$.
  Partition $[\![r]\!]$ into $\mathcal{P}$, the set of equivalent classes of rank-one supports (cf. Definition 3.2).
  for $P \in \mathcal{P}$ **do**
    Compute the SVD of A_{R_P,C_P} to find a pair $(X^\star, Y^\star)$, $X^\star \in \mathbb{R}^{|R_P|\times|P|}$, $Y^\star \in \mathbb{R}^{|C_P|\times|P|}$ with minimum $\|A_{R_P,C_P} - X^\star(Y^\star)^\top\|^2$.
    Assign $X_{R_P,P} = X^\star, Y_{C_P,P} = Y^\star$.
  end for
  return (X,Y)
end procedure

Theorem 3.1 requires all the rank-one contribution supports of different classes to be disjoint. Nevertheless, this condition is quite restrictive. The next result allows partial intersection between two representative rank-one contribution supports.

Definition 3.3 (Complete equivalence classes of rank-one supports - CEC)**.** *$P \in \mathcal{P}$ is a* complete equivalence class *(or* CEC*) if $|P| \geq \min\{|C_P|, |R_P|\}$. Denote $\mathcal{P}^\star \subseteq \mathcal{P}$ the family of all complete equivalence classes, $T = \cup_{P\in\mathcal{P}^\star}P \subseteq [\![r]\!]$, $\bar{T} = [\![r]\!]\backslash T$, and the shorthand $\mathcal{S}_T = \mathcal{S}_{\mathcal{P}^\star}$.*

The interest of complete equivalence classes is that their expressivity is powerful enough to represent any matrix whose support is included in $\mathcal{S}_T$, as illustrated by the following lemma:

Lemma 3.2. *Given $I \in \{0,1\}^{m\times r}$, $J \in \{0,1\}^{n\times r}$, consider T, $\mathcal{S}_T$ as in Definition 3.3. For any matrix $A \in \mathbb{R}^{m\times n}$ such that $\mathtt{supp}(A) \subseteq \mathcal{S}_T$, there exist $X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{n\times r}$ such that $A = XY^\top$ and $\mathtt{supp}(X) \subseteq (\llbracket m \rrbracket \times T) \cap I$, $\mathtt{supp}(Y) \subseteq (\llbracket n \rrbracket \times T) \cap J$.*

Algorithm 2 shows a procedure to find X, Y satisfying Lemma 3.2.

Algorithm 2 Fixed support matrix factorization under the assumptions of Lemma 3.2

procedure FILL_CEC($A \in \mathbb{R}^{m\times n}$, $I \in \{0,1\}^{m\times r}, J \in \{0,1\}^{n\times r}, T \subseteq \llbracket r \rrbracket$)
  Partition T into $\mathcal{P}^\star$, the family of all CECs.
  Initialize $X = \mathbf{0}, Y = \mathbf{0}$.
  for $P \in \mathcal{P}^\star$ **do**
    Let $A' = A - XY^\top$
    if $|P| \geq |C_P|$ **then**
      Choose a matrix $X' \in \mathbb{R}^{|R_P|\times|P|}$ full row rank.
      Assign $X_{R_P,P} = X', Y_{C_P,P} = (X'(X'X'^\top)^{-1}A'_{R_P,C_P})^\top$.
    else
      Choose a matrix $Y' \in \mathbb{R}^{|C_P|\times|P|}$ full row rank.
      Assign $X_{R_P,P} = A'_{R_P,C_P}(Y'Y'^\top)^{-1}Y', Y_{C_P,P} = Y'$.
    end if
    return (X, Y)
  end for
end procedure

The next definition introduces the key properties that the indices k which are not in any CEC need to satisfy in order to make the matrix factorization with fixed support overall tractable.

Definition 3.4 (Rectangular support outside CECs of rank-one supports)**.** *Given $I \in \{0,1\}^{m\times r}$, $J \in \{0,1\}^{n\times r}$, consider T and $\mathcal{S}_T$ as in Definition 3.3 and $\bar{T} = \llbracket r \rrbracket \setminus T$. For $k \in \bar{T}$ define the support outside CECs of the k^{th} rank-one support as:*

$$\mathcal{S}'_k = \mathcal{S}_k \setminus \mathcal{S}_T.$$

If $\mathcal{S}'_k = R_k \times C_k$ for some $R_k \subseteq \llbracket m \rrbracket, C_k \subseteq \llbracket n \rrbracket$, (or equivalently $\mathcal{S}'_k$ is of rank at most one), we say the support outside CECs of the k^{th} rank-one support $\mathcal{S}'_k$ is rectangular.

To state our tractability result, we further categorize the indices in I and J as follows:

Definition 3.5 (Taxonomy of indices of I and J)**.** *With the notations of Definition 3.4, assume that $\mathcal{S}'_k$ is rectangular for all $k \in \bar{T}$. We decompose the indices of I (resp J) into three sets as follows:*

	Classification for I	Classification for J
1	$I_T = \{(i,k) \mid k \in T, i \in \llbracket m \rrbracket\} \cap I$	$J_T = \{(j,k) \mid k \in T, j \in \llbracket n \rrbracket\} \cap J$
2	$I^1_{\bar{T}} = \{(i,k) \mid k \notin T, i \in R_k\} \cap I$	$J^1_{\bar{T}} = \{(j,k) \mid k \notin T, j \in C_k\} \cap J$
3	$I^2_{\bar{T}} = \{(i,k) \mid k \notin T, i \notin R_k\} \cap I$	$J^2_{\bar{T}} = \{(j,k) \mid k \notin T, j \notin C_k\} \cap J$

The following theorem states a more general result than that in Theorem 3.1.

Theorem 3.3. *Consider $I \in \{0,1\}^{m\times r}$, $J \in \{0,1\}^{n\times r}$. Assume that for all $k \in \bar{T}$, $\mathcal{S}'_k$ is rectangular and that for all $k \neq l, k, l \in \bar{T}$, we have $\mathcal{S}'_k = \mathcal{S}'_l$ or $\mathcal{S}'_k \cap \mathcal{S}'_l = \emptyset$. Then, $(I^1_{\bar{T}}, J^1_{\bar{T}})$ satisfy the assumptions of Theorem 3.1. Moreover, for any matrix $A \in \mathbb{R}^{m\times n}$, two instances of Problem* (1) *with data (A, I, J) and $(A \odot 1_{\bar{\mathcal{S}}_T}, I^1_{\bar{T}}, J^1_{\bar{T}})$ respectively, share the same infimum. Given an optimal solution of one instance, we can construct the optimal solution of the other in polynomial time.*

Theorem 3.3 implies that instead of solving the problem with support constraints (I, J), we can deal with other support constraints satisfying Theorem 3.1, which allow for an efficient solution by Algorithm 1. In particular, Theorem 3.1 is a special case of Theorem 3.3 when all the equivalent classes (including CECs) have disjoint representative rank-one supports.

An algorithm for instances satisfying Theorem 3.3 is given in Algorithm 3 (cf. Corollary B.1 in Appendix B).

Interestingly, the condition of Theorem 3.3 also ensures nice properties to Problem (1): if a solution is locally optimal, it is also globally optimal, despite the non-convexity of the objective function, cf. proof of Theorem 3.3 in (B.3). Such condition is verified for certain interesting support constraints in practice. In [9], we show such an example. We propose a hierarchical extension of our method, designed to handle multi-layer matrix factorization (the case in which the matrix is approximated as the product of more than two factors) and demonstrate the superior performance of Algorithm 3 in comparison to other first-order optimization approaches in terms both of computational time and accuracy.

Algorithm 3 SVD for fixed support matrix factorization (under assumptions of Theorem 3.3)

procedure SVD_FSMF2($A \in \mathbb{R}^{m\times n}, I \in \{0,1\}^{m\times r}, J \in \{0,1\}^{n\times r}$)
  Partition the indices of I, J into $I_T, I^1_{\bar{T}}, I^2_{\bar{T}}$ (and $J_T, J^1_{\bar{T}}, J^2_{\bar{T}}$) (Definition 3.4).
  $(X^1_{\bar{T}}, Y^1_{\bar{T}}) =$ SVD_FSMF($A \odot 1_{\bar{\mathcal{S}_T}}, I^1_{\bar{T}}, J^1_{\bar{T}}$) ($T, \mathcal{S}_T$ in Definition 3.3).
  $(X_T, Y_T) =$ FILL_CEC($A \odot 1_{\mathcal{S}_T}, I, J, T$).
  return $(X_T + X^1_{\bar{T}}, Y_T + Y^1_{\bar{T}})$
end procedure

4 Landscape of matrix factorization with fixed support

In this section, we first recall the definition of *spurious local valleys* and *spurious local minima*, which are undesirable objects in the landscape of a function preventing local optimization methods to converge to globally optimal solutions. Previous works [22, 23, 7] showed that the landscape of the optimization problem associated to low rank approximation is free of such *spurious objects*, which potentially gives the intuition for its tractability.

We prove that similar results hold for the family of tractable support constraints for Problem (1) that we introduced in Theorem 3.3. These results might suggest a natural conjecture: an instance of Problem (1) is tractable if and only if the landscape is benign. However, this is not true. We show an example that contradicts this conjecture: we show an instance of Problem (1) that can be solved efficiently, despite the fact that its corresponding landscape contains spurious objects. We will see in the next section that the opposite direction is not so evident either: we propose a numerical illustration of the fact that even when the landscape is benign, the solution of Problem (1) may not be so straightforward with standard iterative methods.

4.1 Spurious local minima and spurious local valleys

We start by recalling the classical definitions of global and local minima of a real-valued function.

Definition 4.1 (Spurious local minimum [23, 14])**.** *Consider $L : \mathbb{R}^d \to \mathbb{R}$. A vector $x^* \in \mathbb{R}^d$ is:*

*a **global minimum** (of L) if $L(x^*) \leq L(x), \forall x$.*

*a **local minimum** if there is a neighborhood $\mathcal{N}$ of x^* such that $L(x^*) \leq L(x), \forall x \in \mathcal{N}$.*

*a **strict local minimum** if there is a neighborhood $\mathcal{N}$ of x^* such that $L(x^*) < L(x), \forall x \in \mathcal{N}, x \neq x^*$.*

*a **spurious local minimum** if x^* is a local minimum but it is not a global minimum.*

The presence of spurious local minima is undesirable because local optimization methods can get stuck in one of them and never reach the global optimum. However, this is not the only undesirable landscape in an optimization problem: spurious local valleys, as defined next, are also challenging.

Definition 4.2 (Sublevel Set [1])**.** *Consider $L : \mathbb{R}^d \to \mathbb{R}$. For every $\alpha \in \mathbb{R}$, the α-level set of L is the set $E_\alpha = \{x \in \mathbb{R}^d \mid L(x) \leq \alpha\}$.*

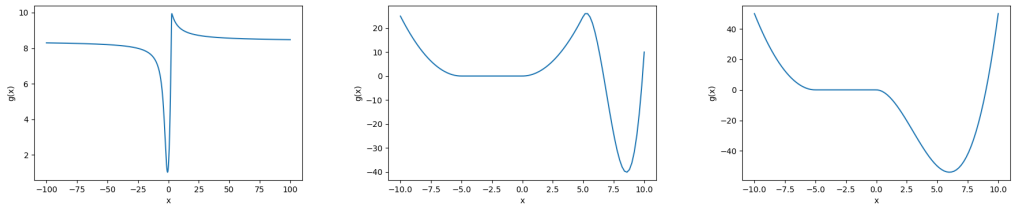
Definition 4.3 (Path-Connected Set and Path-Connected Component)**.** *A subset $S \subseteq \mathbb{R}^d$ is path-connected if for every $x, y \in S$, there is a continuous function $r : [0,1] \to S$ such that $r(0) = x, r(1) = y$. A path-connected component of $E \subseteq \mathbb{R}^d$ is a maximal path-connected subset: $S \subseteq E$ is path-connected, and if $S' \subseteq E$ is path-connected with $S \subseteq S'$ then $S = S'$.*

Definition 4.4 (Spurious Local Valley [22, 13]). *Consider $L : \mathbb{R}^d \to \mathbb{R}$ and a set $S \subset \mathbb{R}^d$.*

*S is a **local valley** of L if it is a non-empty path-connected component of some sublevel set.*

*S is a **spurious local valley** of L if it is a local valley of L and does not contain a global minimum.*

The notion of spurious local valley is inspired by the definition of a *strict* spurious local minimum (a strict local minimum that is also spurious). If x^* is a strict spurious local minimum, then $\{x^*\}$ is a spurious local valley. However, the notion of spurious local valley has a wider meaning than just a neighborhood of a strict spurious local minimum. Figure 1 illustrates some other scenarios. As shown



(a) (b) (c)

Figure 1: Examples of functions with spurious objects.

on Figure 1a, the segment (approximately) $[10, +\infty)$ creates a spurious local valley, and this function has only one local (and global) minimizer, at zero. In Figure 1b, there are spurious local minima that are not strict, but form a spurious local valley anyway. It is worth noticing that the concept of a spurious local valley does *not* cover that of a spurious local minimum. Functions can have spurious (non-strict) local minima even if they do not possess any spurious local valley (Figure 1c).

Therefore, in this paper, we treat the existence of spurious local valleys and spurious local minima independently. The common point is that if the landscape possesses either of them, local optimization methods need to have proper initialization to have guarantees of convergence to a global minimum.

4.2 Landscape of matrix factorization with support constraints

We start with the first result on the landscape in the simple setting of Theorem 3.1.

Theorem 4.1. *Under the assumption of Theorem 3.1, the function $L(X, Y) = \|A - XY^\top\|^2$ with $\mathtt{supp}(X) \subseteq I, \mathtt{supp}(Y) \subseteq J$ does not admit any spurious local valley for any matrix A. In addition, any critical point of L which is not a global minimizer is a strict saddle point.*

Since a strict saddle point cannot be a local minimum, L does not have spurious local minima either. The proof of Theorem 4.1 in Appendix C is based on results [22, 23] on the landscape of low rank matrix factorization. On the other hand, results on the landscape for the setting in Theorem 3.3 are less straightforward. We introduce first the result on the non-existence of spurious local valleys.

Theorem 4.2. *Under the assumptions of Theorem 3.3, the landscape of the function $L(X, Y) = \|A - XY^\top\|^2$ with $\mathtt{supp}(X) \subseteq I, \mathtt{supp}(Y) \subseteq J$ has no spurious local valley for any matrix A.*

The next natural question is whether spurious local minima exist in the setting of Theorem 3.3. While in the setting of Theorem 3.1, all critical points which are not global minima are saddle points, the setting of Theorem 3.3 allows second order critical points, which are not global minima.

Example 4.1. *Consider the following pair of support contraints I, J and factorized matrix A:*

$$I = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \; J = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \; A = \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}.$$

With the notations of Definition 3.3, $T = \{1\}$. This choice of I and J satisfies the setting of Theorem 3.3. The infimum of $L(X, Y) = \|A - XY^\top\|^2$ is zero, and attained, for example at $X^* = I_2, Y^* = A$. Consider the following critical point (X_0, Y_0) satisfying the support constraints I, J:

$$X_0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \; Y_0 = \begin{pmatrix} 0 & 10 \\ 0 & 0 \end{pmatrix}, \; X_0 Y_0^\top = \begin{pmatrix} 10 & 0 \\ 0 & 0 \end{pmatrix} \neq A$$

The Hessian of the function L at (X_0, Y_0) is positive semi-definite.

This example shows that if we want to prove the non-existence of spurious local minima in the new setting, one cannot rely only on the Hessian. This is quite challenging since the computation of the second order derivatives is already tedious. Nevertheless, with proper additional assumptions, we can still say something about spurious local minima in the new setting.

Theorem 4.3. *Consider (X, Y) such that $\mathtt{supp}(X) \subseteq I, \mathtt{supp}(Y) \subseteq J$. With the assumptions of Theorem 3.3 and notations of Definition 3.3: if $X_{R_P,P}$ or $Y_{C_P,P}$ has full row rank for each $P \in \mathcal{P}^\star$, then (X, Y) is not a spurious local minimum of Problem* (1)*. Otherwise there is a feasible path, along which $L(\cdot, \cdot)$ is constant, that joins (X, Y) to some (X', Y') which is not a spurious local minimum.*

4.3 Absence of correlation between tractability and benign landscape

So far, we have witnessed that the instances of Problem (1) satisfying the assumptions of Theorem 3.3 are not only efficiently solvable using Algorithm 3: they also have a landscape with no spurious local valleys. Although Theorem 4.3 does not exclude completely the existence of spurious local minima, together with Theorem 4.1, we eliminate a large number of such points. The question of interest is: Is there a link between such benign landscape and the tractability of the problem? Even if the natural answer seems to be yes, as it is the case for the full support case, we prove that this conjecture is not true. We first provide a counter example showing that tractability does not imply a benign landscape. Then, in Section 5 we provide numerical illustration of the fact that even with a benign landscape the convergence of the gradient descent method may not be straightforward.

First, we provide a sufficient condition for the existence of a spurious local valley in matrix factorization problem with fixed support.

Theorem 4.4. *Given two support constraints I, J of Problem* (1)*, if there exist i_1, i_2, j_1, j_2 such that $i_1 \neq i_2, j_1 \neq j_2$ and (i_1, j_1) belongs to at least 2 rank-one contribution supports, one of which is $\mathcal{S}_k$, and if $(i_1, j_2), (i_2, j_1), (i_2, j_2)$ belong only to $\mathcal{S}_k$, then:*

1) There exists A such that: $L(X, Y) = \|A - XY^\top\|^2$ has a spurious local valley.

2) There exists A such that: $L(X, Y) = \|A - XY^\top\|^2$ has a spurious local minimum.

In both cases, A can be chosen such that $A_{i_2,j_2} \neq 0$.

The property $A_{i_2,j_2} \neq 0$ allows to build a counter-example to the conjecture mentioned above:

Example 4.2. *Consider an instance of Problem* (1) *with the following I, J:*

$$I = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \; J = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

This pair I, J satisfies the assumptions of Theorem 4.4 with $i_1 = 1, i_2 = 2, j_1 = 1, j_2 = 2$. Thus, with well chosen $A \in \mathbb{R}^{2\times 2}, A = (A_{i,j}), 1 \leq i, j \leq 2$ such that $A_{2,2} \neq 0$, the landscape admits spurious objects. On the other hand, the problem is tractable for every $A \in \mathbb{R}^{2\times 2}$ with $A_{2,2} \neq 0$. Indeed, $\inf_{\mathtt{supp}(X)\subseteq I, \mathtt{supp}(Y)\subseteq J} L(X, Y) = 0$ with optimal factors analytically given by:

$$X = \begin{pmatrix} 1 & A_{1,2}/A_{2,2} \\ 0 & 1 \end{pmatrix}, Y = \begin{pmatrix} A_{1,1} - A_{1,2}A_{2,1}/A_{2,2} & A_{2,1} \\ 0 & A_{2,2} \end{pmatrix}, XY^\top = \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix}$$

When $A_{2,2} = 0$, the infimum of $L(X, Y)$ might not be achievable, see Remark A.1 in Appendix A.

The existence of spurious local valleys shown in Theorem 4.4 highlights the importance of initialization: if an initial point is already inside a spurious valley, first-order methods cannot escape this suboptimal area. An optimist may wonder if there nevertheless exist a smart initialization that avoids all spurious local valleys. As proved in Appendix C.6, the answer is positive.

Theorem 4.5. *Given any I, J, A such that the infimum of Problem* (1) *is attained, the initialization $(X, \mathbf{0}), \mathtt{supp}(X) \subseteq I$ (or symmetrically $(\mathbf{0}, Y), \mathtt{supp}(Y) \subseteq J$) is not in any spurious local valley.*

Yet, such an initialization does not guarantee that first-order methods converge to a global minimum. In the proof of this result we indeed show that there exists a continuous path joining this smart initialization to an optimal solution without increasing the loss function. Thus, such initialization is completely outside any spurious local valley. Nevertheless, first-order methods are not bound to follow our constructive continuous path. In the next section we further elaborate on the importance of the starting guess for local optimization methods.

5 Numerical illustration: landscape's properties and convergence of the gradient descent

As shown in Section 4.2, Problem (1) has a good landscape under the assumptions of Theorem 3.3. This might suggest that, from a random initialization, popular optimization methods such as gradient descent might easily be able to return the globally optimal solution. Nevertheless, the situation is more tricky. Actually, the effectiveness of those methods in this specific case has never been shown in practice. Thus, this section shows the empirical performance of gradient descent in tackling the problem of matrix factorization with fixed support.

Consider the following minimalistic instance of Problem (1):

$$A = (0, 1) \qquad I = (1) \qquad J = (1, 1)^\top$$

This instance satisfies the assumptions of Theorem 3.1, thus its landscape is free of spurious objects by Theorem 4.1. The infimum of this instance is zero, attained by solutions of the form $X^* = (a), Y^* = (0, b)$ with $ab = 1$.

We perform gradient descent for this instance. We denote $X = (x), Y = (y_1, y_2)$ and we define $g(Y) := g(y_1, y_2) = \min_X L(X, Y) = \min_x (xy_1)^2 + (1 - xy_2)^2$. Empirical experiments show that the application of gradient descent to $L(X, Y)$ is very well approximated by the application of gradient descent to $g(y_1, y_2)$. We consider then this procedure, that allows us for instance to have a 3D visualization as in Figure 2 (this is not possible for the original problem that has 3 parameters in total). Figure 2b (the loss surface of function $g(y_1, y_2)$) also shows visual proof of the fact that the landscape has no spurious local object, as proved in Theorem 4.1.

With fixed y_1, y_2, g is a simple quadratic function w.r.t x. Solving the quadratic minimization problem, we have $g(y_1, y_2) = y_1^2/(y_1^2 + y_2^2)$ and it is attained with $x = y_2/(y_1^2 + y_2^2)$. We consider two initializations: $X_0 = (0.02), Y_0 = (20, 10)$ and $X_1 = (0.2), Y_1 = (2, 1)$, which both satisfy the condition $x = y_2/(y_1^2 + y_2^2)$. The learning rate α is chosen by backtracking line search, satisfying the Armijo condition [14][3].

From Figure 2a it is clear that the performance of the gradient descent is deeply affected by the choice of the initial guess, despite the absence of spurious objects in the landscape. Indeed, Figure 2c presents the surface of the gradient of $g(y_1, y_2)$ and shows that the sequence generated starting from (X_0, Y_0) (blue line on the right) resides completely inside an area with very small gradient (of order 10^{-5}), despite the fact this area is not close to any optimal solution, and thus the method has a lot of difficulties to converge to the optimum. In contrast, (X_1, Y_1) lies in an area with large gradient and its corresponding sequence of solutions achieves optimality much faster.

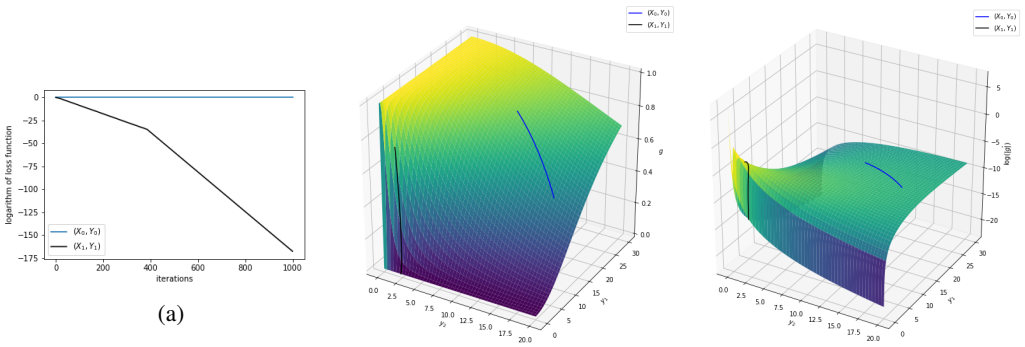


Figure 2: (a) Evolution of the logarithm of $L(X, Y)$ with two different initializations. (b) The surface of $g(y_1, y_2) = \min_X L(X, Y)$ (c) The surface of $\log \|\nabla g(y_1, y_2)\|$. Trajectories of gradient descent from (X_0, Y_0) after $\mathbf{10^5}$ iterations (blue) and from (X_1, Y_1) after $\mathbf{10^3}$ iterations (black).

[3]For the problem $\min_x f(x)$ Armijo condition requires α to satisfy $f(x - \alpha\nabla f(x)) \leq f(x) - \alpha c\|\nabla f(x)\|^2$, we set $c = 10^{-4}$.

The example shows that the effectiveness of gradient descent for Problem (1) heavily depends on initialization, which is not evident to choose. In contrast, our Algorithm 3.3 does not require to tune any hyper-parameter.

6 Conclusion

In this paper, we studied the problem of two-layer matrix factorization with fixed support. We showed that sparse matrix factorization with prior knowledge of the support is still NP-hard in general. Nevertheless, certain structured supports allow for an efficient solution algorithm. Furthermore, we also showed the non-existence of spurious objects in the landscape of function $L(X, Y)$ of Problem (1) with these support constraints. Although it would have seemed natural to assume an equivalence between tractability and benign landscape of Problem (1), we also show a counter-example that contradicts this conjecture. That shows that there is still room for improvement of the current tools (spurious objects) to characterize the tractability of an instance. We have also shown numerically the advantages of our approach compared to state-of-the-art first-order optimization methods. We refer the reader to [9] for a deeper comparison. We propose there an extension of the approach to fixed-support multilayer sparse factorization and show the superiority of our method in terms of accuracy and speed.

References

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

[2] T. Dao, A. Gu, M. Eichhorn, A. Rudra, and C. Re. Learning fast algorithms for linear transforms using butterfly factorizations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 1517–1527, 2019.

[3] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[4] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, 2013.

[5] N. Gillis and F. Glineur. Low-rank matrix approximation with weights or missing data is NP-hard. *SIAM Journal on Matrix Analysis and Applications*, 32, 2010.

[6] G. Golub and W. Kahan. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis*, 2(2):205–224, 1965.

[7] K. Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems 29*, pages 586–594. 2016.

[8] N. Kishore Kumar and J. Shneider. Literature survey on low rank approximation of matrices. *ArXiv preprint 1606.06511*, 2016.

[9] Q.T Le, L. Zheng, E. Riccietti, and R. Gribonval. Fast learning of fast transforms, with guarantees. Technical report, 2021.

[10] L. Le Magoarou and R. Gribonval. Chasing butterflies: In search of efficient dictionaries. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3287–3291, 2015.

[11] L. Le Magoarou and R. Gribonval. Flexible Multi-layer Sparse Approximations of Matrices and Applications. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):688–700, 2016.

[12] X. Li, J. Lu, R. Arora, J. Haupt, H. Liu, Z. Wang, and T. Zhao. Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization. *IEEE Transactions on Information Theory*, 65(6):3489–3514, 2019.

[13] Q. Nguyen. On connected sublevel sets in deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 4790–4799, 2019.

[14] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, second edition, 2006.

[15] R. Peeters. The maximum edge biclique problem is NP-complete. *Discrete Appl Math*, 131, 2000.

[16] R. Rubinstein, A. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98:1045 – 1057, 07 2010.

[17] R. Rubinstein, M. Zibulevsky, and M. Elad. Double sparsity: Learning sparse dictionaries for sparse signal approximation. *IEEE Transactions on Signal Processing*, 58:1553 – 1564, 2010.

[18] Y. Shitov. A short proof that NMF is NP-hard. *Arxiv preprint 1605.04000*, 2016.

[19] I. Tošić and P. Frossard. Dictionary learning. *IEEE Signal Processing Magazine*, 28(2):27–38, 2011.

[20] L. N. Trefethen and D. Bau III. *Numerical linear algebra*, volume 50. SIAM, 1997.

[21] S. A. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2010.

[22] L. Venturi, A. S. Bandeira, and J. Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20(133):1–34, 2019.

[23] Z. Zhu, D. Soudry, Y. C. Eldar, and M. Wakin. The global optimization geometry of shallow linear neural networks. *Journal of Mathematical Imaging and Vision*, 62:279–292, 2019.

Supplementary material

Supplementary material for the paper: "Spurious Valleys, Spurious Minima and NP-hardness of Sparse Matrix Factorization With Fixed Support". This appendix is organized as follows:

- Appendix A: Proofs for Section 2: Matrix factorization with fixed support is NP-hard
- Appendix B: Proofs for Section 3: Tractable instances of Matrix factorization with fixed support
- Appendix C: Proofs for Section 4: Landscape of Matrix Factorization with Fixed Support
- Appendix D: Proofs for other intermediate and minor technical results.

Aside from standard notations, we introduce additional necessary ones in the main proofs:

Additional notations for proofs

Given a matrix $A \in \mathbb{R}^{m\times n}$, if $S \subseteq [\![m]\!], T \subseteq [\![n]\!]$, then $A_{S,T} \in \mathbb{R}^{|S|\times|T|}$ is the submatrix of A restrained to rows and columns indexed in S and T respectively. Given $T \subseteq [\![n]\!]$, $A_T \in \mathbb{R}^{m\times n}$ is the matrix that has the same columns as A for indexes in T and is zero elsewhere (while the notation $A_{\bullet,T} \in \mathbb{R}^{m\times|T|}$ introduced in the main text is a submatrix of A). If $I \in \{0,1\}^{m\times n}$ is the matrix support constraint, I_T is equivalent to a new support constraint $I \cap \{(i,j) \mid j \in T\}$. Notice that this notation for the support constraints I and J will not cause any confusion with Definition 3.5.

Operator norm of a matrix A is denoted by $|||A|||$, while $\|A\|$ is the Frobenius norm.

A Proofs for Section 2

A.1 Proof of Lemma 2.2

Up to a transposition, we can assume WLOG that $m \geq n$. We will show that with $r = n+1 = \min(m,n)+1$, we can find two supports I and J satisfying the conclusion of Lemma 2.2.

To create an instance of fixed support matrix factorization (i.e., two supports I, J) that is *equivalent* to Problem (4), we define $I \in \{0,1\}^{m\times(n+1)}$ and $J \in \{0,1\}^{n\times(n+1)}$ as follows:

$$I_{i,j} = \begin{cases} 1 - W_{i,j} & \text{if } j \neq n \\ 1 & \text{if } j = n+1 \end{cases}$$
$$J_{i,j} = \begin{cases} 1 & \text{if } j = i \text{ or } j = n+1 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Figure 3 illustrates an example of support constraints built from W.

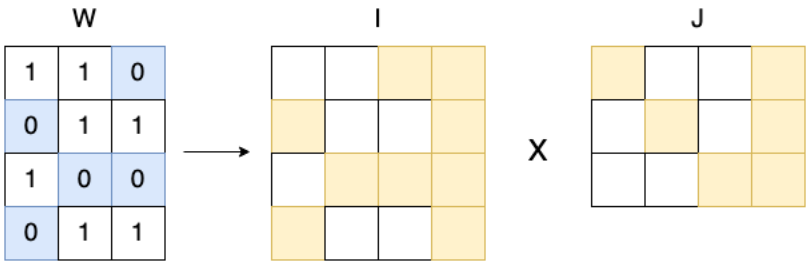


Figure 3: Factor supports I and J constructed from the weighted matrix $W \in \{0,1\}^{4\times3}$. Colored squares in I and J are positions in the supports.

We consider the following problem:

$$\underset{\text{supp}(X)\subseteq I, \text{supp}(Y)\subseteq J}{\text{Minimize}} \|A - XY^\top\|^2$$

where I and J are defined in Equation (6). This construction (of I and J) can clearly be made in polynomial time.

Consider the coefficients $(XY^\top)_{i,j}$:

1) If $W_{i,j} = 0$: $(XY^\top)_{i,j} = \sum_{k=1}^{n+1} X_{i,k}Y_{j,k} = X_{i,j}Y_{j,j} + X_{i,n+1}Y_{j,n+1}$ (except for $k = n+1$, only $Y_{j,j}$ can be different from zero due to our choice of J).
2) If $W_{i,j} = 1$: $(XY^\top)_{i,j} = \sum_{k=1}^{n+1} X_{i,k}Y_{j,k} = X_{i,n+1}Y_{j,n+1}$ (same reason as in the previous case, in addition to the fact that $I_{i,j} = 1 - W_{i,j} = 0$).

Therefore, the following equation holds:

$$(XY^\top) \odot W = (X_{\bullet,n+1}Y_{n+1,\bullet}^\top) \odot W \tag{7}$$

We will prove that Problem (1) and Problem (4) share the same infimum[4]. Let $\mu_1 = \inf_{x\in\mathbb{R}^m, y\in\mathbb{R}^n} \|A - xy^\top\|_W^2$ and $\mu_2 = \inf_{\mathtt{supp}(X)\subseteq I, \mathtt{supp}(Y)\subseteq J} \|A - XY^\top\|^2$. It is clear that $\mu_i \geq 0 > -\infty, i = 1, 2$. Our objective is to prove $\mu_1 \leq \mu_2$ and $\mu_2 \leq \mu_1$.

1) Proof of $\mu_1 \leq \mu_2$: By definition of an infimum, for all $\mu > \mu_1$, there exist x, y such that $\|A - xy^\top\|_W^2 \leq \mu$. We can choose X and Y (with $\mathtt{supp}(X) \subseteq I, \mathtt{supp}(Y) \subseteq J$) as follows: we take the last columns of X and Y equal to x and y ($X_{\bullet,n+1} = x, Y_{\bullet,n+1} = y$). For the *remaining* columns of X and Y, we choose:

$$\begin{aligned} X_{i,j} &= A_{i,j} - x_i y_j \quad && \text{if } I_{i,j} = 1, j \leq n \\ Y_{i,j} &= 1 && \text{if } J_{i,j} = 1, j \leq n \end{aligned}$$

This choice of X and Y will make $\|A - XY^\top\|^2 = \|A - xy^\top\|_W^2 \leq \mu$. Indeed, for all (i, j) such that $W_{i,j} = 0$, we have:

$$\begin{aligned} (A - XY^\top)_{i,j} &= A_{i,j} - X_{i,j}Y_{j,j} - X_{i,n+1}Y_{j,n+1} \\ &= A_{i,j} - (A_{i,j} - x_i y_j) - x_i y_j \\ &= 0 \end{aligned}$$

Therefore, it is clear that: $(A - XY^\top) \odot (\mathbf{1} - W) = \mathbf{0}$.

$$\begin{aligned} \|A - XY^\top\|^2 &= \|(A - XY^\top) \odot W\|^2 + \|(A - XY^\top) \odot (\mathbf{1} - W)\|^2 \\ &= \|(A - XY^\top) \odot W\|^2 \\ &\overset{(7)}{=} \|(A - X_{\bullet,n+1}Y_{\bullet,n+1}^\top) \odot W\|^2 \\ &= \|(A - xy^\top) \odot W\|^2 \\ &= \|A - xy^\top\|_W^2 \end{aligned}$$

Therefore, $\mu_2 \leq \mu_1$.

2) Proof of $\mu_1 \leq \mu_2$: Inversely, for all $\mu > \mu_2$, there exists X, Y satisfying $\mathtt{supp}(X) \subseteq I, \mathtt{supp}(Y) \subseteq J$ such that $\|A - XY^\top\|^2 \leq \mu$. We choose $x = X_{\bullet,n+1}, y = Y_{\bullet,n+1}$. It is immediate that:

$$\begin{aligned} \|A - xy^\top\|_W^2 &= \|(A - xy^\top) \odot W\|^2 \\ &= \|(A - X_{\bullet,n+1}Y_{\bullet,n+1}^\top) \odot W\|^2 \\ &\overset{(7)}{=} \|(A - XY^\top) \odot W\|^2 \\ &\leq \|(A - XY^\top) \odot W\|^2 + \|(A - XY^\top) \odot (\mathbf{1} - W)\|^2 \\ &= \|A - XY^\top\|^2 \end{aligned}$$

Thus, $\|A - xy^\top\|_W^2 \leq \|A - XY^\top\|^2 \leq \mu$. We have $\mu_1 \leq \mu_2$.

[4]We focus on the infimum instead of minimum since there are cases where the infimum is not attained, as shown in Remark A.1

This shows that $\mu_1 = \mu_2$. Moreover, the proofs of $\mu_1 \leq \mu_2$ and $\mu_2 \leq \mu_1$ also show the procedures to obtain an optimal solution of one problem with a given accuracy ϵ provided that we know an optimal solution of the other with the same accuracy.

Remark A.1. *In the proof of Lemma 2.2, we focus on the infimum instead of minimum since there are cases where the infimum is not attained. Indeed, consider the following matrix and binary matrix in Problem* (1)*:*

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\ I = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},\ J = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}. \tag{8}$$

The infimum of this problem is zero, which can be shown by choosing:

$$X = \begin{pmatrix} -k & k \\ 0 & \frac{1}{k} \end{pmatrix}, Y = \begin{pmatrix} k & k \\ 0 & \frac{1}{k} \end{pmatrix} \text{ so that } XY^\top = \begin{pmatrix} 0 & 1 \\ 1 & \frac{1}{k^2} \end{pmatrix}.$$

In the limit, when k goes to infinity, we have:

$$\lim_{k\to\infty} \|A - XY^\top\|^2 = \lim_{k\to\infty} \frac{1}{k^2} = 0.$$

Yet, there does not exist any couple (x, y) such that $\|A - XY\|^2 = 0$. Indeed, any such couple would need to satisfy:

$$\begin{cases} X_{1,2}Y_{2,2} & = 1 \\ X_{2,2}Y_{1,2} & = 1 \\ X_{2,2}Y_{2,2} & = 0 \end{cases}$$

However, the third equation implies that either $X_{2,2} = 0$ or $Y_{2,2} = 0$, which makes either $X_{2,2}Y_{1,2} = 0$ or $X_{1,2}Y_{2,2} = 0$. This leads to a contradiction.

In fact, I and J are constructed from the following weight binary matrix W (the construction is similar to one in the proof of Lemma 2.2).

$$W = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \tag{9}$$

Problem (4) *with A, W defined in Equations* (8) *and* (9) *has unattainable infimum as well.*

A.2 Proof of Theorem 2.3

The proof uses Lemma 2.2. Given any instance of Problem (4) (i.e., a matrix $A \in [0,1]^{m\times n}, W \in \{0,1\}^{m\times n}$), we can produce an instance of Problem (1) (the same matrix A and $I \in \{0,1\}^{m\times r}, J \in \{0,1\}^{n\times r}$) such that both problems have the same infimum. Moreover, for any given objective function accuracy, we can use the procedure of Lemma 2.2 to make sure the solutions of both problems share the same accuracy.

Last but not least, all the procedures are polynomial. This defines a polynomial reduction from Problem (4) to Problem (1). That shows the NP-hardness of Problem (1).

B Proofs for Section 3

B.1 Proof of Theorem 3.1

In this proof, for each equivalent class $P \in \mathcal{P}$ (Definition 3.2) we use the notations $X_P \in \mathbb{R}^{m\times r}, Y_P \in \mathbb{R}^{n\times r}$ (introduced in **Additional notations for proofs**) since $P \subset [\![r]\!]$. We also use R_P, C_P (Definition 3.2). We notice that for each equivalent class $P \in \mathcal{P}$, we have:

$$(X_P Y_P^\top)_{R_P,C_P} = X_{R_P,P} Y_{C_P,P}^\top \tag{10}$$

and the product $XY^\top$ can be decomposed as:

$$XY^\top = \sum_{P\in\mathcal{P}} X_P Y_P^\top. \tag{11}$$

Due to the hypothesis of this theorem, with $P, P' \in \mathcal{P}, P' \neq P$, we further have:

$$X_{P'} Y_{P'}^\top \odot \mathcal{S}_P = \mathbf{0} \tag{12}$$

The objective function $L(X, Y)$ is:

$$\begin{aligned}
\|A - XY^\top\|^2 &\overset{(11)}{=} \|A - \sum_{P' \in \mathcal{P}} X_{P'} Y_{P'}^\top\|^2 \\
&= \left(\sum_{P \in \mathcal{P}} \|(A - \sum_{P' \in \mathcal{P}} X_{P'} Y_{P'}^\top) \odot \mathcal{S}_P\|^2 \right) + \|(A - \sum_{P' \in \mathcal{P}} X_{P'} Y_{P'}^\top) \odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2 \\
&\overset{(12)}{=} \left(\sum_{P \in \mathcal{P}} \|(A - X_P Y_P^\top) \odot \mathcal{S}_P\|^2 \right) + \|A \odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2 \\
&= \left(\sum_{P \in \mathcal{P}} \|A_{R_P, C_P} - (X_P Y_P^\top)_{R_P, C_P}\|^2 \right) + \|A \odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2 \\
&\overset{(10)}{=} \left(\sum_{P \in \mathcal{P}} \|A_{R_P, C_P} - X_{R_P, P} Y_{C_P, P}^\top\|^2 \right) + \|A \odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2
\end{aligned} \tag{13}$$

Therefore, if we ignore the constant $\|A \odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2$, the function $L(X, Y)$ is decomposed into a sum of functions $\|A_{R_P, C_P} - X_{R_P, P} Y_{C_P, P}^\top\|^2$, which are instances of a low rank approximation problem. Since all the optimized parameters are $\{(X_{R_P, P}, Y_{C_P, P})\}_{P \in \mathcal{P}}$, the optimal solution of L is $\{(X^\star_{R_P, P}, Y^\star_{C_P, P})\}_{P \in \mathcal{P}}$, where $(X^\star_{R_P, P}, Y^\star_{C_P, P})$ is the optimal solution of the function $\|A_{R_P, C_P} - X_{R_P, P} Y_{C_P, P}^\top\|^2$. Since $(X^\star_{R_P, P}, Y^\star_{C_P, P})$ can be calculated by SVD, the problem can be solved efficiently.

B.2 Proof of Lemma 3.2

Before proving Lemma 3.2, we prove an intermediate lemma. This lemma can be thought as a special case of Lemma 3.2, where there is only one complete equivalent class (CEC).

Lemma B.1. *Consider $I \in \{0, 1\}^{m \times r}$, $J \in \{0, 1\}^{n \times r}$, and $P \in \mathcal{P}$, $\mathcal{S}_P$ as in Definition 3.2. Assume that P is complete as in Definition 3.3. Then, for any matrix $A \in \mathbb{R}^{m \times n}$ such that $\mathtt{supp}(A) \subseteq \mathcal{S}_P$, one can construct two matrices $X \in \mathbb{R}^{m \times r}, Y \in \mathbb{R}^{n \times r}$ such that $A = XY^\top$ and $\mathtt{supp}(X) \subseteq I_P, \mathtt{supp}(Y) \subseteq J_P$ (the notations I_P, J_P are introduced in **Additional notations for proofs**).*

Proof. WLOG, up to permuting rows and columns, we can assume $P = [\![|P|]\!], R_P = [\![|R_P|]\!]$ and $C_P = [\![|C_P|]\!]$ (R_P and C_P are defined in Definition 3.2). To have $\mathtt{supp}(X) \subseteq I_P$ and $\mathtt{supp}(Y) \subseteq J_P$, X and Y must have the following form:

$$X = \begin{pmatrix} X' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, Y = \begin{pmatrix} Y' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

where $X' \in \mathbb{R}^{|R_P| \times |P|}, Y' \in \mathbb{R}^{|C_P| \times |P|}$.

Since $\mathtt{supp}(A) \subseteq \mathcal{S}_P = [\![|R_P|]\!] \times [\![|C_P|]\!]$, then the matrix A must have the form:

$$A = \begin{pmatrix} A' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

where $A' \in \mathbb{R}^{|R_P| \times |C_P|}$.

To conclude the proof, it is thus sufficient to find two matrices X', Y' such that $X'Y'^\top = A'$. Indeed, if such X' and Y' can be constructed, we have:

$$XY^\top = \begin{pmatrix} X'Y'^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} A' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = A$$

Due to the definition of CEC, we have $|P| \geq \min(|R_P|, |C_P|)$. WLOG, up to transposing the problem, we assume $|P| \geq |R_P|$. Let X' be a matrix with full row rank, $Y' = A'^\top (X' X'^\top)^{-1} X'$ (both are well defined due to $|P| \geq |R_P|$). It is evident that:

$$X'Y'^\top = X'X'^\top (X'X'^\top)^{-1} A' = A'$$

which is our desired property for X' and Y'. □

Proof of lemma 3.2. With the introduction of the notation I_T and J_T, the support constraints of X, Y can be simply re-written as: $\texttt{supp}(X) \subseteq I_T, \texttt{supp}(Y) \subseteq J_T$.

We prove by induction on the size of a subset $\mathcal{P}' \subseteq \mathcal{P}^\star$ that: for each matrix $A \in \mathbb{R}^{m\times n}$ such that $\texttt{supp}(A) \subseteq \mathcal{S}_{\mathcal{P}'}$, there exist $X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{n\times r}$ such that $A = XY^\top$ and $\texttt{supp}(X) \subseteq I_{T'}$, $\texttt{supp}(Y) \subseteq J_{T'}$ where $T' = \cup_{P\in\mathcal{P}'} P$. Applying the result to $\mathcal{P}' = \mathcal{P}^\star$ will give the conclusion since $T = \cup_{P\in\mathcal{P}^\star} P$ and $\mathcal{S}_T = \mathcal{S}_{\mathcal{P}^\star}$.

By Lemma B.1 the result is true if $|\mathcal{P}'| = 1$. Assume the result is true if $|\mathcal{P}'| = p$, where $1 \leq p < |\mathcal{P}^\star|$, and consider the case where $|\mathcal{P}'| = p + 1$.

Pick an arbitrary class $P \in \mathcal{P}'$ and partition $\mathcal{P}'$ into $\mathcal{P}'' = \mathcal{P}'\backslash\{P\}$ and $\{P\}$. Consider a matrix A such that $\texttt{supp}(A) \subseteq \mathcal{S}_{\mathcal{P}'}$ and let $B = A \odot \mathcal{S}_P$. Since $\texttt{supp}(B) \subset \mathcal{S}_P$, by Lemma B.1 there are X_P, Y_P such that $B = X_P Y_P^\top$, $\texttt{supp}(X_P) \subset I_P$, $\texttt{supp}(Y_P) \subset J_P$.

Define $A' = A - B$. Since $\texttt{supp}(A) \subseteq \mathcal{S}_{\mathcal{P}'}$ and $\texttt{supp}(B) \subseteq \mathcal{S}_P \subseteq \mathcal{S}_{\mathcal{P}'}$ we have $\texttt{supp}(A') \subseteq \mathcal{S}_{\mathcal{P}'}$. Moreover by definition $A'_{\mathcal{S}_P} = 0$, hence $\texttt{supp}(A') \subseteq \mathcal{S}_{\mathcal{P}''}$. By the induction hypothesis on p there are $X_{\mathcal{P}''}$, $Y_{\mathcal{P}''}$ such that $A' = X_{\mathcal{P}''} Y_{\mathcal{P}''}^\top$, $\texttt{supp}(X_{\mathcal{P}''}) \subset I_{T''}$, $\texttt{supp}(Y_{\mathcal{P}''}) \subset J_{T''}$ where $T'' = \cup_{P'\in\mathcal{P}''} P'$.

Defining $X = X_P + X_{\mathcal{P}''}$ and $Y = Y_P + Y_{\mathcal{P}''}$ we obtain $\texttt{supp}(X) \subset I_{T''} \cup I_P = I_{T'}$, $\texttt{supp}(Y) \subset J_{T''} \cup J_P = J_{T'}$. Moreover, we have:

$$XY^\top = X_P Y_P^\top + X_{\mathcal{P}''} Y_{\mathcal{P}''}^\top = B + A' = A$$

□

B.3 Proof of Theorem 3.3

First, we decompose the factors X and Y using the taxonomy of indices from Definition 3.5.

Definition B.1. *Given I_T, J_T and $I^i_{\bar{T}}, J^i_{\bar{T}}, i = 1, 2$ as in Definition 3.5, consider (X, Y) a feasible solution of matrix factorization with fixed support (i.e., $\texttt{supp}(X) \in I, \texttt{supp}(Y) \in J$), we denote:*

1) $X_T = X \odot I_T, X^i_{\bar{T}} = X \odot I^i_{\bar{T}}$, *for* $i = 1, 2$.

2) $Y_T = Y \odot I_T, Y^i_{\bar{T}} = Y \odot I^i_{\bar{T}}$, *for* $i = 1, 2$.

with $\odot$ the Hadamard product between a matrix and a support constraint (introduced in Section 1.1).

The following is a technical result.

Lemma B.2. *Given $I \in \{0, 1\}^{m\times r}$, $J \in \{0, 1\}^{n\times r}$, consider $I, J, T, \mathcal{S}_T, \mathcal{S}_{\mathcal{P}}$ as in Definition 3.2, $X_T, X^i_{\bar{T}}, Y_T, Y^i_{\bar{T}}$ as in Definition 3.4 and assume that for all $k \in \bar{T}$, $\mathcal{S}'_k$ is rectangular.*

It holds:

$$XY^\top \odot 1_{\mathcal{S}_T} = XY^\top - (X^1_{\bar{T}})(Y^1_{\bar{T}})^\top = X_T Y_T^\top + \sum_{(i,j)\neq(1,1)} (X^i_{\bar{T}})(Y^j_{\bar{T}})^\top, \tag{14}$$

$$XY^\top \odot 1_{\mathcal{S}_{\mathcal{P}}\backslash\mathcal{S}_T} = (X^1_{\bar{T}})(Y^1_{\bar{T}})^\top, \tag{15}$$

$$XY^\top \odot 1_{\bar{\mathcal{S}}_{\mathcal{P}}} = \mathbf{0}. \tag{16}$$

Proof. Since $I_T, I^1_{\bar{T}}, I^2_{\bar{T}}$ form a partition of I (and similarly $J_T, J^1_{\bar{T}}, J^2_{\bar{T}}$ form a partition of J), one can write X and Y as:

$$X = X_T + X^1_{\bar{T}} + X^2_{\bar{T}}, \qquad Y = Y_T + Y^1_{\bar{T}} + Y^2_{\bar{T}}.$$

Since $\texttt{supp}(X_T) \subseteq I_T, \texttt{supp}(X^i_{\bar{T}}) \subseteq I_{\bar{T}}, \texttt{supp}(Y_T) \subseteq J_T, \texttt{supp}(Y^i_{\bar{T}}) \subseteq J_{\bar{T}}, i = 1, 2$, the product $XY^\top$ can be decomposed as:

$$XY^\top = X_T Y_T^\top + \sum_{1\leq i,j\leq 2} (X^i_{\bar{T}})(Y^j_{\bar{T}})^\top. \tag{17}$$

To prove all the three points of the lemma, it is sufficient to show that:

a) $\texttt{supp}((X^1_{\bar{T}})(Y^1_{\bar{T}})^\top) \subseteq \mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T$.

b) All the remaining components in Equation (17) (namely $X_T Y_T^\top$ and $(X^i_{\bar{T}})(Y^j_{\bar{T}})^\top, (i,j) \neq (1,1)$) have their supports included in $\mathcal{S}_T$.

With these properties, the proof of the three points of the lemma is immediate since:

1) $XY^\top \odot 1_{\mathcal{S}_T} = (X_T Y_T^\top + \sum_{1\leq i,j\leq 2}(X^i_{\bar{T}})(Y^j_{\bar{T}})^\top) \odot 1_{\mathcal{S}_T} = X_T Y_T^\top + \sum_{(i,j)\neq(1,1)}(X^i_{\bar{T}})(Y^j_{\bar{T}})^\top$.

2) $XY^\top \odot 1_{\mathcal{S}_\mathcal{P}\setminus\mathcal{S}_T} = (X_T Y_T^\top + \sum_{1\leq i,j\leq 2}(X^i_{\bar{T}})(Y^j_{\bar{T}})^\top) \odot 1_{\mathcal{S}_\mathcal{P}\setminus\mathcal{S}_T} = (X^1_{\bar{T}})(Y^1_{\bar{T}})^\top$.

3) $XY^\top \odot 1_{\bar{\mathcal{S}}_\mathcal{P}} = (X_T Y_T^\top + \sum_{1\leq i,j\leq 2}(X^i_{\bar{T}})(Y^j_{\bar{T}})^\top) \odot 1_{\bar{\mathcal{S}}_\mathcal{P}} = \mathbf{0}$.

To be able to see why a) and b) are true, we first remark that $\texttt{supp}(X_T Y_T^\top) \subseteq \mathcal{S}_T$ (definition of $\mathcal{S}_T$ in Definition 3.2). In addition, considering a matrix index (i,j), we have:

$$((X^1_{\bar{T}})(Y^2_{\bar{T}})^\top)_{i,j} = \sum_{k=1}^{r}(X^1_{\bar{T}})_{i,k}(Y^2_{\bar{T}})_{j,k} = \sum_{k|(j,k)\in J^2_{\bar{T}}} X_{i,k}Y_{j,k} \tag{18}$$

Due to the hypothesis of rectangular support outside CEC (Definition 3.4), we have: $\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T = \cup_{\ell\in\bar{T}} R_\ell \times C_\ell$. If $(i,j) \in \mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T$, Equation (18) shows that $((X^1_{\bar{T}})(Y^2_{\bar{T}})^\top)_{i,j} = 0$ since there is no such $k \in J^2_{\bar{T}}$ due to the definition of $J^2_{\bar{T}}$). Moreover, $\texttt{supp}((X^1_{\bar{T}})(Y^2_{\bar{T}})^\top) \subseteq \mathcal{S}_\mathcal{P}$ (since $\texttt{supp}(X^1_{\bar{T}}) \subseteq I, \texttt{supp}(Y^2_{\bar{T}}) \subseteq J$). Thus, it shows that $\texttt{supp}((X^1_{\bar{T}})(Y^2_{\bar{T}})^\top) \subseteq \mathcal{S}_\mathcal{P} \setminus (\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T) = \mathcal{S}_T$.

We can proceed similarly to prove $\texttt{supp}((X^2_{\bar{T}})(Y^2_{\bar{T}})^\top), \texttt{supp}((X^2_{\bar{T}})(Y^1_{\bar{T}})^\top) \subseteq \mathcal{S}_T$.

Finally, we rewrite Equation (18) for $(i,j) \in \texttt{supp}((X^1_{\bar{T}})(Y^1_{\bar{T}})^\top)$:

$$((X^1_{\bar{T}})(Y^1_{\bar{T}})^\top)_{i,j} = \sum_{\substack{k|(j,k)\in J^1_{\bar{T}} \\ (i,k)\in I^1_{\bar{T}}}} X_{i,k}Y_{j,k}$$

Therefore, $(i,j) \in \cup_{\ell\in\bar{T}} R_\ell \times C_\ell$ (by definition of $I^1_{\bar{T}}, J^1_{\bar{T}}$). This shows that $\texttt{supp}((X^1_{\bar{T}})(Y^1_{\bar{T}})^\top) \subseteq \mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T$. □

Here, we present the proof of Theorem 3.3.

Proof of Theorem 3.3. Consider $X_T, Y_T, X^i_{\bar{T}}, Y^i_{\bar{T}}, i = 1,2$ defined as in Definition B.1. Let μ_1 and μ_2 be the infimum value of Problem (1) with the first instance (A, I, J) and with the second instance $(A', I^1_{\bar{T}}, J^1_{\bar{T}})$ (with $A' = A \odot 1_{\bar{\mathcal{S}}_T}$) respectively.

First, we remark that $I^1_{\bar{T}}$ and $J^1_{\bar{T}}$ satisfy the assumptions of Theorem 3.1. Indeed, it holds $\mathcal{S}_k(I^1_{\bar{T}}, J^1_{\bar{T}}) = \mathcal{S}_k(I,J) \setminus \mathcal{S}_T = \mathcal{S}'_k$ by construction. For any two indices $k, l \in \bar{T}$, the representative rank-one supports are either equal ($\mathcal{S}'_k = \mathcal{S}'_l$) or disjoint ($\mathcal{S}'_k \cap \mathcal{S}'_l = \emptyset$) by assumption. That shows why $I^1_{\bar{T}}$ and $J^1_{\bar{T}}$ satisfy the assumptions of Theorem 3.1.

Next, we prove that $\mu_1 = \mu_2$. Since $(\mathcal{S}_T, \mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T, \bar{\mathcal{S}}_\mathcal{P})$ form a partition of $[\![m]\!] \times [\![n]\!]$, we have $1_C \odot 1_D = \mathbf{0}, C \neq D, C, D \in \{\mathcal{S}_T, \mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T, \bar{\mathcal{S}}_\mathcal{P}\}$. From the definition of A' it holds $A' \odot 1_{\bar{\mathcal{S}}_\mathcal{P}} = A \odot 1_{\bar{\mathcal{S}}_\mathcal{P}}$ and $A' \odot 1_{\mathcal{S}_T} = \mathbf{0}$. Moreover, from Lemma B.2 it holds $(X^1_{\bar{T}})(Y^1_{\bar{T}})^\top \odot 1_{\mathcal{S}_T \cup \bar{\mathcal{S}}_\mathcal{P}} = XY^\top \odot 1_{\mathcal{S}_\mathcal{P}\setminus\mathcal{S}_T} \odot 1_{\mathcal{S}_T\cup\bar{\mathcal{S}}_\mathcal{P}} = \mathbf{0}$.

We then obtain:

$$
\begin{aligned}
&\|A' - X^1_{\bar{T}}(Y^1_{\bar{T}})^\top\|^2 \\
&= \|(A' - X^1_{\bar{T}}(Y^1_{\bar{T}})^\top) \odot 1_{\mathcal{S}_T}\|^2 + \|(A' - X^1_{\bar{T}}(Y^1_{\bar{T}})^\top) \odot 1_{\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T}\|^2 + \|(A' - X^1_{\bar{T}}(Y^1_{\bar{T}})^\top) \odot 1_{\bar{\mathcal{S}}_\mathcal{P}}\|^2 \\
&= \|(A' - (X^1_{\bar{T}})(Y^1_{\bar{T}})^\top) \odot 1_{\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T}\|^2 + \|A' \odot 1_{\bar{\mathcal{S}}_\mathcal{P}}\|^2 \\
&\overset{(15)}{=} \|(A - XY^\top) \odot 1_{\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T}\|^2 + \|A \odot 1_{\bar{\mathcal{S}}_\mathcal{P}}\|^2 \\
&\overset{(14)}{=} \|(A - XY^\top) \odot 1_{\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T}\|^2 + \|(A - XY^\top) \odot 1_{\bar{\mathcal{S}}_\mathcal{P}}\|^2 \\
&\leq \|(A - XY^\top) \odot 1_{\mathcal{S}_T}\|^2 + \|(A - XY^\top) \odot 1_{\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T}\|^2 + \|(A - XY^\top) \odot 1_{\bar{\mathcal{S}}_\mathcal{P}}\|^2 \\
&= \|(A - XY^\top)\|^2
\end{aligned}
\tag{19}
$$

Therefore, for any solution (X, Y) satisfying $\texttt{supp}(X) \subseteq I, \texttt{supp}(Y) \subseteq J$, we can choose $X' = X^1_{\bar{T}}, Y' = Y^1_{\bar{T}}$ (having $\texttt{supp}(X') \subseteq I^1_{\bar{T}}, \texttt{supp}(Y') \subseteq J^1_{\bar{T}}$) such that $\|A - XY^\top\| \geq \|A' - X'Y'^\top\|$. This shows $\mu_1 \geq \mu_2$.

On the other hand, given any solution (X', Y') satisfying $\texttt{supp}(X') \subseteq I^1_{\bar{T}}, \texttt{supp}(Y') \subseteq J^1_{\bar{T}}$, we can construct a solution (X, Y) in which $\texttt{supp}(X) \subseteq I, \texttt{supp}(Y) \subseteq J$ such that the equality $\|A - XY^\top\|^2 = \|A' - X'Y'^\top\|^2$ holds. To do that, we construct (X, Y) to have an equality in the second last line of Equation (19). Since $X = X_T + X^1_{\bar{T}} + X^2_{\bar{T}}, Y = Y_T + Y^1_{\bar{T}} + Y^2_{\bar{T}}$, we choose:

1) $X^1_{\bar{T}} = X', Y^1_{\bar{T}} = Y'$,

2) $X^2_{\bar{T}}, Y^2_{\bar{T}}$ arbitrarily,

3) X_T and Y_T such that:

$$X_T Y_T^\top = (A - \sum_{(i,j) \neq (1,1)} (X^i_{\bar{T}})(Y^j_{\bar{T}})^\top) \odot 1_{\mathcal{S}_T}$$

which is possible due to Lemma 3.2.

By Lemma B.2, with this choice we have:

$$
\begin{aligned}
(A - XY^\top) \odot 1_{\mathcal{S}_T} &= (A \odot 1_{\mathcal{S}_T}) - (XY^\top \odot 1_{\mathcal{S}_T}) \\
&= A \odot 1_{\mathcal{S}_T} - \sum_{(i,j) \neq (1,1)} (X^i_{\bar{T}})(Y^j_{\bar{T}})^\top - X_T Y_T^\top \\
&= (A - \sum_{(i,j) \neq (1,1)} (X^i_{\bar{T}})(Y^j_{\bar{T}})^\top) \odot 1_{\mathcal{S}_T} - X_T Y_T^\top = \mathbf{0}
\end{aligned}
\tag{20}
$$

Therefore $\|A - XY^\top\|^2 = \|A' - X'Y'^\top\|^2$ and so $\mu_2 \geq \mu_1$. We have successfully proved that $\mu_1 = \mu_2$. In addition, given (X, Y) an optimal solution of Problem (1) with instance (A, I, J), we have shown how to construct an optimal solution (X', Y') with instance $(A \odot 1_{\bar{\mathcal{S}}_T}, I^1_{\bar{T}}, J^1_{\bar{T}})$ and vice versa. That completes our proof. □

The following Corollary is a direct consequence of the proof of Theorem 3.3.

Corollary B.1. *With the same assumptions and notations as in Theorem 3.3, a feasible point (X, Y) (i.e., such that $\texttt{supp}(X) \subseteq I, \texttt{supp}(Y) \subseteq J$) is an optimal solution of Problem* (1) *if and only if:*

1) $(X \odot I^1_{\bar{T}}, Y \odot J^1_{\bar{T}})$ is an optimal solution of Problem (1) *with data $(A \odot 1_{\bar{\mathcal{S}}_T}, I^1_{\bar{T}}, J^1_{\bar{T}})$.*

2) The following equation holds: $(A - XY^\top) \odot 1_{\mathcal{S}_T} = \mathbf{0}$

In the proof of Theorem 3.3, one can choose $X_{\bar{T}}^2, Y_{\bar{T}}^2$ arbitrarily. If we choose $X_{\bar{T}}^2 = \mathbf{0}, Y_{\bar{T}}^2 = \mathbf{0}$, thanks to Equation 20, X_T and Y_T has to satisfy:

$$\begin{aligned} X_T Y_T^\top &= (A - \sum_{(i,j)\neq(1,1)} (X_{\bar{T}}^i)(Y_{\bar{T}}^j)^\top) \odot 1_{\mathcal{S}_T} \\ &= (A - (X_{\bar{T}}^1)(Y_{\bar{T}}^1)^\top) \odot 1_{\mathcal{S}_T} \\ &= A \odot 1_{\mathcal{S}_T} \end{aligned}$$

Last equation is due to Lemma B.2, Equation 15 implying $\texttt{supp}((X_{\bar{T}}^1)(Y_{\bar{T}}^1)^\top) \subseteq \mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T$.

C Proofs for Section 4

This section contains the proofs of the most technical results of the paper. Our results rely on the classical ones on the landscape of the function $L(X, Y) = \|A - XY^\top\|^2$ where X, Y have no support constraints (low rank matrix approximation case). Those results will be introduced below.

C.1 Previous results on the landscape

For the non-existence of spurious local minima in the classical case, previous works [7, 23] used the fact that the Hessian is not positive semi-definite (PSD) to prove that a critical point (but not a global minimizer) is a saddle point. To prove the non-existence of spurious local valleys, the following lemma was employed in previous works:

Lemma C.1 (Sufficient condition for the non-existence of any spurious local valley [22, Lemma 2])**.** *Consider a continuous function $L : \mathbb{R}^d \to \mathbb{R}$. Assume that, for any initial parameter $\tilde{x} \in \mathbb{R}^d$, there exists a continuous path $f : t \in [0, 1] \to \mathbb{R}^d$ such that:*

a) $f(0) = \tilde{x}$.

b) $f(1) \in \arg\min_{x\in\mathbb{R}^d} L(x)$.

c) The function $L \circ f : t \in [0, 1] \to \mathbb{R}$ is non-increasing.

Then there is no spurious local valley in the landscape of function L.

The proof can be found in [22]. The main idea of the proof is: given any initial point, if one can find a continuous path connecting the initial point to a global minimizer and the loss function is non-increasing on the path, then there does not exist any spurious local valley.

Conversely, we generalize an idea from [22] into the following lemma, which gives a sufficient condition for the existence of a spurious local valley:

Lemma C.2 (Sufficient condition for the existence of a spurious local valley)**.** *Consider a continuous function $L : \mathbb{R}^d \to \mathbb{R}$ whose global minimum is attained. Assume we know three subsets $S_1, S_2, S_3 \subset \mathbb{R}^d$ such that:*

1) The global minima of L are in S_1.

2) Every continuous path from S_3 to S_1 passes through S_2.

3) $\inf_{x\in S_2} L(x) > \inf_{x\in S_3} L(x) > \inf_{x\in S_1} L(x)$.

Then L has a spurious local valley.

Proof. Denote $\Sigma = \{x \mid L(x) = \inf_{x\in\mathbb{R}^d} L(\theta)\}$ the set of global minimizers of L. Σ is not empty due to the assumption that the global minimum is attained.

Since $\inf_{x\in S_2} L(x) > \inf_{x\in S_3} L(x)$, there exists $\tau \in S_3$ such that $L(\tau) < \inf_{x\in S_2} L(x)$. Consider Φ the path-connected component of the sublevel set $\{x \mid L(x) \leq L(\tau)\}$ that contains τ. It is sufficient to prove that $\Phi \cap \Sigma = \emptyset$.

Indeed, by contradiction, let's assume that there exists $\tau' \in \Phi \cap \Sigma$. By definition, there exists a continuous function $f : [0, 1] \to \Phi$ such that $f(0) = \tau \in S_3, f(1) = \tau' \in S_1$. Due to the assumption

that every continuous path from S_3 to S_1 has to pass through a point in S_2, there must exist $t \in (0,1)$ such that $f(t) \in S_2 \cap \Phi$. Therefore, $L(f(t)) \leq L(\tau)$ (since $f(t) \in \Phi$) and $L(f(t)) > L(\tau)$ (since $f(t) \in S_2$), which is a contradiction. □

To finish this section, we formally recall previous results which are related to Problem (1) and will be used in our subsequent proofs. Previous works focused on low rank matrix approximation (or *full support matrix factorization*) [5]. The questions of the existence of spurious local valleys and spurious local minima were addressed for full support matrix factorization and deep linear neural networks [22, 13, 23, 7]. We present only results related to our problem of interest.

Theorem C.3 (No spurious local valleys in linear networks [22])**.** *Consider linear neural networks of any depth $K \geq 1$ and of any layer widths $p_k \geq 1$ and any input - output dimension $n, m \geq 1$ with the following form:*

$$\Phi(b, \theta) = W_K \dots W_1 b$$

where $\theta = (W_i)_{i=1}^K$, and $b \in \mathbb{R}^n$ is a training input sample. With the squared loss function, there is no spurious local valley. More specifically, the function $L(\theta) = \|\Phi(B, \theta) - A\|^2$ satisfies the condition of Lemma C.1 for any matrices $A \in \mathbb{R}^{m \times N}$ and $B \in \mathbb{R}^{n \times N}$ (A and B are the whole sets of training output and input respectively).

Theorem C.4 (No spurious local minima in shallow linear networks [23])**.** *Let $B \in \mathbb{R}^{d_0 \times N}, A \in \mathbb{R}^{d_2 \times N}$ be input and output training examples. Consider the optimization problem:*

$$\underset{\substack{X \in \mathbb{R}^{d_0 \times d_1}, \\ Y \in \mathbb{R}^{d_1 \times d_2}}}{\text{Minimize}} L(X, Y) = \|XYB - A\|^2$$

If B is full row rank, then function f has no spurious local minimum. More specifically, a critical point of f which is not global minimizer is a saddle point (i.e., its Hessian is not positive semi-definite).

Both theorems are valid for a particular case of matrix factorization with fixed support: full support matrix factorization. Indeed, given a factorized matrix $A \in \mathbb{R}^{m \times n}$, in Theorem C.3, if $K = 2, B = \mathbf{I}_n (n = N)$, then the considered function is $L = \|A - W_2 W_1\|^2$. This is Problem (1) without support constraints I and J (and without a transpose on W_1, which does not change the nature of the problem). Theorem C.3 guarantees that L satisfies the conditions of Lemma C.1, thus has no spurious local valley.

Similarly, in Theorem C.4, if $B = \mathbf{I}_{d_0} (d_0 = N)$ (therefore, B is full row rank), we return to the same situation of Theorem C.3. In general, Theorem C.4 claims that the landscape of the full support matrix factorization problem has the strict saddle point property and thus, does not have spurious local minima.

However, once we turn to Problem (1) with arbitrary I and J, such benign landscape is not guaranteed anymore, as we will show in Example 4.2. Our work can be considered as a generalization of previous results [23, 22, 7].

C.2 Proof of Theorem 4.1

Recall that under the assumption of Theorem 3.1, all the variables to be optimized are $\{(X_{R_P,P}, Y_{C_P,P})\}_{P \in \mathcal{P}}$ ($P, \mathcal{P}$ are defined in Definition 3.2).

We assume $\mathcal{P} = \{P_1, P_2, \dots, P_\ell\}, P_i \subseteq [\![r]\!], i \leq \ell$. From Equation (13), we have:

$$\|A - XY^\top\|^2 = \left(\sum_{P \in \mathcal{P}} \|A_{R_P,C_P} - X_{R_P,P} Y_{C_P,P}^\top\|^2 \right) + \|A \odot \bar{\mathcal{S}_\mathcal{P}}\|^2 \tag{21}$$

Therefore, the function $L(X, Y)$ is a sum of functions $L_P(X_{R_P,P}, Y_{C_P,P}) := \|A_{R_P,C_P} - X_{R_P,P} Y_{C_P,P}^\top\|^2$, which do not share parameters and are instances of the full support matrix factorization problem. The global minimizers of L are $\{(X^\star_{R_P,P}, Y^\star_{C_P,P})\}_{P \in \mathcal{P}}$, where $(X^\star_{R_P,P}, Y^\star_{C_P,P})$ is any global minimizer of $\|A_{R_P,C_P} - X_{R_P,P} Y_{C_P,P}^\top\|^2$.

[5] Since previous works also considered the case $r \geq m, n$, low rank approximation might be misleading sometimes. That is why we occasionally use the name full support matrix factorization to emphasize this fact.

1) **Non-existence of any spurious local valley**: By Theorem C.3, from any initial solution $(X^0_{R_P,P}, Y^0_{C_P,P})$, there exists a continuous function f_P that satisfies the conditions in Lemma C.1, which are:
   i) $f_P(0) = (X^0_{R_P,P}, Y^0_{C_P,P})$.
   ii) $f_P(1) = (X^\star_{R_P,P}, Y^\star_{C_P,P})$.
   iii) $L_P \circ f_P : [0,1] \to \mathbb{R}$ is non-increasing.

   Consider the continuous function $f = (f_{P_1}, \ldots, f_{P_\ell})$. It satisfies the assumptions of Lemma C.1, which shows the non-existence of any spurious local valley.

2) **Non-existence of any spurious local minimum**: Due to the decomposition in Equation (21), the gradient and Hessian of $L(X,Y)$ have the following form:

$$\frac{\partial L}{\partial X_{R_P,P}} = \frac{\partial L_P}{\partial X_{R_P,P}}, \qquad \frac{\partial L}{\partial Y_{C_P,P}} = \frac{\partial L_P}{\partial Y_{C_P,P}}, \ \forall P \in \mathcal{P}$$

$$H(L(X,Y)) = \begin{pmatrix} H(L_{P_1}(X_{R_{P_1},P_1}, Y_{C_{P_1},P_1})) & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & H(L_{P_\ell}(X_{R_{P_\ell},P_\ell}, Y_{C_{P_\ell},P_\ell})) \end{pmatrix}$$

   If (X,Y) is a critical point of $L(X,Y)$, $(X_{R_P,P}, Y_{C_P P})$ has to be a critical point of the function L_P for all $P \in \mathcal{P}$. Provided that (X,Y) is not a global minimizer of $L(X,Y)$, there exists $P \in \mathcal{P}$ such that $(X_{R_P,P}, Y_{C_P,P})$ is not a global minimizer of L_P. By Theorem C.4, $H(L_P)_{|(X_{R_P,P}, Y_{C_P,P})}$ is not positive semi-definite. Hence, $H(L)_{|(X,Y)}$ is not positive semi-definite either, which implies that (X,Y) it is not a spurious local minimum as well.

C.3 Proof of Theorem 4.2

Sketch of the proof of Theorem 4.2. The proof of the non-existence of any spurious local valley is based on Lemma C.1 (i.e., from any initial solution (X^0, Y^0), there is a continuous non-increasing path that leads to the optimal solution). The continuous path is constructed by concatenating:

1) A continuous path f_1, with non-increasing cost $L \circ f_1$, that connects (X^0, Y^0) to some pair (X^1, Y^1) satisfying:
$$(A - X^1(Y^1)^\top) \odot 1_{\mathcal{S}_T} = \mathbf{0}.$$
   The function f_1 is built in Lemma C.8.

2) A continuous path f_2, with non-increasing cost $L \circ f_2$, that connects (X^1, Y^1) to a global minimizer (X^*, Y^*). The function f_2 is built in Lemma C.9. □

Before jumping to the construction of f_1, f_2, we justify in Lemma C.5 an assumption that will simplify the proof: we assume that for all $P \in \mathcal{P}^\star$, either $X^0_{R_P,P}$ or $Y^0_{C_P,P}$ has full row rank.

Lemma C.5. *Given $I \in \{0,1\}^{m \times r}$, $J \in \{0,1\}^{n \times r}$ consider T and $\mathcal{S}_T$ as in Definition 3.2 and a feasible point (X,Y). There exists a continuous function $f(t) = (X_f(t), Y_f(t)), t \in [0,1]$ such that*

I) $(X_f(0), Y_f(0)) = (X,Y)$;

II) $\mathtt{supp}(X_f(t)) \subseteq I, \mathtt{supp}(Y_f(t)) \subseteq J, \forall t \in [0,1]$;

III) $X_f(t)(Y_f(t))^\top = XY^\top, \forall t \in [0,1]$;

IV) for each $P \in \mathcal{P}^\star$, $(X_f(1))_{R_P,P}$ or $(Y_f(1))_{C_P,P}$ has full row rank.

The proof relies on two intermediate results that we state first. The idea of Lemma C.6 can be found in [22]. Since it is not formally proved as a lemma or theorem, we reprove it here for self-containedness.

Lemma C.6. *Let $X \in \mathbb{R}^{R \times p}, Y \in \mathbb{R}^{C \times p}, \min(R,C) \leq p$. Then there exists a continuous function $f(t) = (X_f(t), Y_f(t))$ on $[0,1]$ such that:*

- $(X_f(0), Y_f(0)) = (X,Y)$.

- $XY^\top = X_f(t)(Y_f(t))^\top, \forall t \in [0,1]$.
- $X_f(1)$ *or* $Y_f(1)$ *has full row rank.*

The proof of Lemma C.6 is postponed to Section D.1.

Corollary C.1. *Given $I \in \{0,1\}^{m\times r}$, $J \in \{0,1\}^{n\times r}$, a feasible point (X,Y), and $P \in \mathcal{P}^\star$ as in Definition 3.2, there is a continuous function $f(t) = (X_f(t), Y_f(t)), t \in [0,1]$ such that*

- $(X_f(0), Y_f(0)) = (X_P, Y_P)$*;*
- $\mathtt{supp}(X_f(t)) \subseteq I_P, \mathtt{supp}(Y_f(t)) \subseteq J_P$*.*
- $X_f(t)(Y_f(t))^\top = X_P(Y_P)^\top, \forall t \in [0,1]$*;*
- $(X_f(1))_{R_P,P}$ *or* $(Y_f(1))_{C_P,P}$ *has full row rank.*

Proof of Corollary C.1. WLOG, up to permuting columns, we can assume $P = [\![|P|]\!], R := R_P = [\![|R_P|]\!]$ and $C := C_P = [\![|C_P|]\!]$ (R_P and C_P are defined in Definition 3.2), and the second condition will be satisfied if we build functions $X_f(t)$ and $Y_f(t)$ of the form:

$$X_f(t) = \begin{pmatrix} X'_f(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, Y_f(t) = \begin{pmatrix} Y'_f(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

with $X'_f(t) \in \mathbb{R}^{R\times P}$, $Y'_f(t) \in \mathbb{R}^{C\times P}$. Since P is a CEC, we have $p \geq \min(R,C)$ hence we can use Lemma C.6 to build $X'_f(t), Y'_f(t)$ such that the other conditions are satisfied. □

Proof of Lemma C.5. First, we decompose X and Y as:

$$X = X_{\bar{T}} + \sum_{P\in\mathcal{P}^\star} X_P, \qquad Y = Y_{\bar{T}} + \sum_{P\in\mathcal{P}^\star} Y_P$$

and observe that, since $\bar{T}$ and $P \in \mathcal{P}^\star$ form a partition of $[\![r]\!]$, the product $XY^\top$ can be written as:

$$XY^\top = X_{\bar{T}}Y_{\bar{T}}^\top + \sum_{P\in\mathcal{P}^\star} X_P Y_P^\top.$$

For each $P \in \mathcal{P}^\star$ we use Corollary C.1 to build continuous functions $(X_f^P(t), Y_f^P(t))$ such that:

1. $(X_f^P(0), Y_f^P(0)) = (X_P, Y_P)$.
2. $\mathtt{supp}(X_f^P(t)) \subseteq I_P, \mathtt{supp}(Y_f^P(t)) \subseteq J_P, \forall t \in [0,1]$.
3. $X_f^P(t)(Y_f^P(t))^\top = X_P Y_P^\top, \forall t \in [0,1]$.
4. $(X_f^P(1))_{R_P,P}$ or $(Y_f^P(1))_{C_P,P}$ has full row rank.

and we define $(X_f(t), Y_f(t))$ as:

$$X_f(t) = X_{\bar{T}} + \sum_{P\in\mathcal{P}^\star} X_f^P(t), \qquad Y(t) = Y_{\bar{T}} + \sum_{P\in\mathcal{P}^\star} Y_f^P(t)$$

To conclude, it is immediate to check that $(X_f(t), Y_f(t))$ satisfies all the conditions $I) - IV)$. □

We are now almost equiped to proceed to the construction of the two continuous paths f_1 and f_2 announced in the sketch of the proof of Theorem 4.2. The last ingredient is the following technical lemma, which proof is postponed to Section D.2

Lemma C.7. *Given $I \in \{0,1\}^{m\times r}$, $J \in \{0,1\}^{n\times r}$, consider T, $\mathcal{S}_T$ as in Definition 3.2. For any feasible point (X,Y) in which $\forall P \in \mathcal{P}^\star, X_{R_P,P}$ or $Y_{C_P,P}$ has full row rank and any continuous function $g : [0,1] \to \mathbb{R}^{m\times n}$ satisfying $\mathtt{supp}(g(t)) \subseteq \mathcal{S}_T$ and $g(0) = X_T Y_T$, there exists a continuous function $f : [0,1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} : f(t) = (X_f(t), Y_f(t))$ (X_f and Y_f are also functions) such that:*

1) $\texttt{supp}(X_f(t)) \subseteq I_T, \texttt{supp}(Y_f(t)) \subseteq J_T$.

2) $X_f(0) = X_T, Y_f(0) = Y_T$.

3) $g(t) = X_f(t)Y_f(t)^\top, \forall t \in [0,1]$.

4) $\|X_f(z) - X_f(t)\|^2 + \|Y_f(z) - Y_f(t)\|^2 \leq \mathcal{C}\|g(z) - g(t)\|^2$.

where $\mathcal{C} = \max_{P \in \mathcal{P}^\star} \left(\max \left(\left\| \left\| X^\dagger_{R_P,P} \right\| \right\|^2, \left\| \left\| Y^\dagger_{C_P,P} \right\| \right\|^2 \right) \right)$.

The continuous path f_1 is built in the following lemma.

Lemma C.8. *Under the assumption of Theorem 3.3, for any feasible point (X, Y) in which $\forall P \in \mathcal{P}^\star$, $X_{R_P,P}$ or $Y_{C_P,P}$ has full row rank, there exists a continuous function $f : [0,1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$: $f(t) = (X_f(t), Y_f(t))$ such that:*

1) $\texttt{supp}(X_f(t)) \subseteq I, \texttt{supp}(Y_f(t)) \subseteq J$.

2) $X_f(0) = X, Y_f(0) = Y$.

3) $L(X_f(t), Y_f(t)) = \|A - X_f(t)Y_f(t)^\top\|^2$ *is non-increasing.*

4) $(A - X_f(1)(Y_f(1))^\top) \odot 1_{\mathcal{S}_T} = \mathbf{0}$.

Proof. We use the notations $X_T = X \odot I_T, X_{\bar{T}} = X \odot I_{\bar{T}}$ ($Y_T, Y_{\bar{T}}$ are defined analogously). Denoting $Z = XY^\top$, we construct f such that $X_f(t)Y_f(t)^\top = B(t)$, where $B(t)$ is defined as:

$$B(t) = Z \odot 1_{\bar{\mathcal{S}}_T} + (At + Z(1-t)) \odot 1_{\mathcal{S}_T}$$

Indeed, such a function f makes $L(X_f(t), Y_f(t))$ non-increasing:

$$\begin{aligned} \|A - X_f(t)Y_f(t)^\top\|^2 &= \|A - B(t)\|^2 \\ &= \|(A - Z) \odot 1_{\bar{\mathcal{S}}_T}\|^2 + (1-t)^2\|(A - Z) \odot 1_{\mathcal{S}_T}\|^2 \end{aligned} \tag{22}$$

Thus, the rest of the proof is devoted to show that such a function f exists by using Lemma C.7. Consider the function $g(t) = B(t) - X_{\bar{T}}(Y_{\bar{T}})^\top$. We have that $g(t)$ is continuous, $g(0) = B(0) - X_{\bar{T}}(Y_{\bar{T}})^\top = Z - X_{\bar{T}}(Y_{\bar{T}})^\top = X_T(Y_T)^\top$ and:

$$\begin{aligned} g(t) \odot 1_{\bar{\mathcal{S}}_T} &= (B(t) - X_{\bar{T}}(Y_{\bar{T}})^\top) \odot 1_{\bar{\mathcal{S}}_T} \\ &= (Z - X_{\bar{T}}(Y_{\bar{T}})^\top) \odot 1_{\bar{\mathcal{S}}_T} \\ &= (X_T Y_T^\top) \odot 1_{\bar{\mathcal{S}}_T} = \mathbf{0} \end{aligned}$$

which shows $\texttt{supp}(g(t)) \subseteq \mathcal{S}_T$. In addition to the hypotheses of full row rank of either $X_{R_P,P}$ or $Y_{C_P,P}$, by invoking Lemma C.7, there exists a function $(X^C_f(t), Y^C_f(t))$ such that:

1) $\texttt{supp}(X^C_f(t)) \subseteq I_T, \texttt{supp}(Y^C_f(t)) \subseteq J_T$.

2) $X^C_f(0) = X_T, Y^C_f(0) = Y_T$.

3) $g(t) = X^C_f(t)(Y^C_f(t))^\top, \forall t \in [0,1]$.

We can define our desired function $(X_f(t), Y_f(t))$ as:

$$X_f(t) = X_{\bar{T}} + X^C_f(t), \qquad Y = Y_{\bar{T}} + Y^C_f(t)$$

and show that it satisfies all the conditions. The first condition is easily verified due to the constraint on the support of $X^C_f(t), Y^C_f(t)$ and X, Y. The second condition is satisfied since:

$$\begin{aligned} X_f(0) &= X_{\bar{T}} + X^C_f(0) = X_{\bar{T}} + X_T = X \\ Y_f(0) &= Y_{\bar{T}} + Y^C_f(0) = Y_{\bar{T}} + Y_T = Y \end{aligned}$$

The third condition results from Equation (22) and the following equation:

$$X_f(t)(Y_f(t))^\top = X_{\bar{T}}Y_{\bar{T}}^\top + X_f^C(t)(Y_f^C(t))^\top = X_{\bar{T}}Y_{\bar{T}}^\top + g(t) = B(t)$$

The last condition is due to the definition of $B(t)$:

$$\begin{aligned}(A - X_f(1)(Y_f(1))^\top) \odot 1_{\mathcal{S}_T} &= (A - B(1)) \odot 1_{\mathcal{S}_T} \\ &= (A - Z \odot 1_{\bar{\mathcal{S}}_T} - A \odot 1_{\mathcal{S}_T}) \odot 1_{\mathcal{S}_T} = \mathbf{0}\end{aligned}$$ □

The second continuous path, f_2, is built in the following lemma.

Lemma C.9. *Under the assumption of Theorem 3.3, for any feasible point (X, Y) in which $\forall P \in \mathcal{P}^\star$, $X_{R_P,P}$ or $Y_{C_P,P}$ has full row rank and verify: $(A - XY^\top) \odot 1_{\mathcal{S}_T} = \mathbf{0}$, there exists a continuous function $f : [0, 1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} : f(t) = (X_f(t), Y_f(t))$ such that:*

1) $\texttt{supp}(X_f(t)) \subseteq I, \texttt{supp}(Y_f(t)) \subseteq J$.

2) $X_f(0) = X^1, Y_f(0) = Y^1$.

3) $L(X(t), Y(t)) = \|A - X(t)Y(t)^\top\|^2$ *is non-increasing.*

4) $(X_f(1), Y_f(1))$ *is an optimal solution of* $L(X, Y)$.

Proof. Consider $X_T, X^i_{\bar{T}}, Y_T, Y^i_{\bar{T}}, i = 1, 2$ as in Definition B.1. We redefine $A' = A \odot 1_{\bar{\mathcal{S}}_T}, I' = I^1_{\bar{T}}, J' = J^1_{\bar{T}}$ as introduced in Theorem 3.3.

In light of Corollary B.1, an optimal solution (X^*, Y^*) has the following form:

1) $(X^*)^1_{\bar{T}} = X^* \odot I^1_{\bar{T}}, (Y^*)^1_{\bar{T}} = Y^* \odot J^1_{\bar{T}}$ is an optimal solution of the instance of Problem (1) with (A', I', J').

2) $(X^*)^2_{\bar{T}} = X^* \odot I^2_{\bar{T}}, (Y^*)^2_{\bar{T}} = Y^* \odot J^2_{\bar{T}}$ can be arbitrary.

3) $(X^*)_T = X^* \odot I_T, (Y^*)_T = Y^* \odot J_T$ satisfy:

$$X^*_T(Y^*_T)^\top = (A - \sum_{(i,j)\neq(1,1)} (X^*)^i_{\bar{T}}((Y^*)^j_{\bar{T}})^\top) \odot 1_{\mathcal{S}_T}$$

Since (I', J') has its support constraints satisfying the assumptions of Theorem 3.1 as shown in Theorem 3.3, by Theorem 4.1, there exists a function $(X^{\bar{T}}_f(t), Y^{\bar{T}}_f(t))$ such that:

1) $\texttt{supp}(X^{\bar{T}}_f(t)) \subseteq I^1_{\bar{T}}, \texttt{supp}(Y^{\bar{T}}_f(t)) \subseteq J^1_{\bar{T}}$.

2) $X^{\bar{T}}_f(0) = X^1_{\bar{T}}, Y^{\bar{T}}_f(0) = Y^1_{\bar{T}}$.

3) $L'(X^{\bar{T}}_f(t), Y^{\bar{T}}_f(t)) = \|A' - X^{\bar{T}}_f(t)Y^{\bar{T}}_f(t)^\top\|^2$ is non-increasing.

4) $(X^{\bar{T}}_f(1), Y^{\bar{T}}_f(1))$ is an optimal solution of the instance of Problem (1) with (A', I', J').

Consider the function $g(t) = \left(A - (X^{\bar{T}}_f(t) + X^2_{\bar{T}})(Y^{\bar{T}}_f(t) + Y^2_{\bar{T}})^\top\right) \odot 1_{\mathcal{S}_T}$. This construction makes $g(0) = X_T Y_T^\top$. Indeed,

$$\begin{aligned}g(0) &= \left(A - (X^{\bar{T}}_f(0) + X^2_{\bar{T}})(Y^{\bar{T}}_f(0) + Y^2_{\bar{T}})^\top\right) \odot 1_{\mathcal{S}_T} \\ &= \left(A - (X^1_{\bar{T}} + X^2_{\bar{T}})(Y^1_{\bar{T}} + Y^2_{\bar{T}})^\top\right) \odot 1_{\mathcal{S}_T} \\ &= A \odot 1_{\mathcal{S}_T} - (\sum_{1\leq i,j\leq 2} X^i_{\bar{T}}(Y^j_{\bar{T}})^\top) \odot 1_{\mathcal{S}_T} \\ &\overset{(1)}{=} (XY^\top - \sum_{1\leq i,j\leq 2} X^i_{\bar{T}}(Y^j_{\bar{T}})^\top) \odot 1_{\mathcal{S}_T} \\ &\overset{(2)}{=} X_T Y_T^\top\end{aligned}$$

where (1) holds by the hypothesis of the lemma $(A - XY^\top) \odot 1_{\mathcal{S}_T} = \mathbf{0}$, and (2) holds by Equation (17) and the fact that $\texttt{supp}(X_T Y_T^\top) \subseteq \mathcal{S}_T$. With the assumption that $\forall P \in \mathcal{P}^\star$, $X_{R_P,P}$ or $Y_{C_P,P}$ has full row rank of this lemma, that $g(t)$ continuous, $\texttt{supp}(g(t)) \subseteq \mathcal{S}_T$ and $g(0) = X_T Y_T^\top$, there exist functions $(X_f^C(t), Y_f^C(t))$ satisfying the assumptions of Lemma C.7:

1) $\texttt{supp}(X_f^C(t)) \subseteq I_T, \texttt{supp}(Y_f^C(t)) \subseteq J_T$.

2) $X_f^C(0) = X_T, Y_f^C(0) = Y_T$.

3) $g(t) = X_f^C(t) Y_f^C(t)^\top, \forall t \in [0,1]$.

Finally, one can define the function $X_f(t), Y_f(t)$ as:

$$X_f(t) = X_f^{\bar{T}}(t) + X_f^C(t) + X_{\bar{T}}^2, \qquad Y_f(t) = Y_f^{\bar{T}}(t) + Y_f^C(t) + Y_{\bar{T}}^2$$

Here is the verification of conditions: The first condition is valid due to the supports of $X_f^P(t), Y_f^P(t), P \in \{\bar{T}, C\}$ and $X_{\bar{T}}^2, Y_{\bar{T}}^2$. The second condition is satisfied since:

$$X_f(0) = X_f^{\bar{T}}(0) + X_f^C(0) + X_{\bar{T}}^2 = X_{\bar{T}}^1 + X_T + X_{\bar{T}}^2 = X$$
$$Y_f(0) = Y_f^{\bar{T}}(0) + Y_f^C(0) + Y_{\bar{T}}^2 = Y_{\bar{T}}^1 + Y_T + Y_{\bar{T}}^2 = Y$$

The third condition is valid since:

$$\begin{aligned}
&\|A - X_f(t) Y_f(t)^\top\|^2 = \|A - X_f^C(t)(Y_f^C(t))^\top - (X_f^{\bar{T}}(t) + X_{\bar{T}}^2)(Y_f^{\bar{T}}(t) + Y_{\bar{T}}^2)^\top\|^2 \\
&= \|g(t) - X_f^C(t) Y_f^C(t)^\top\|^2 + \|(A - X_f^{\bar{T}}(t)(Y_f^{\bar{T}}(t))^\top) \odot 1_{\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T}\|^2 + \|A \odot 1_{\bar{\mathcal{S}}_\mathcal{P}}\|^2 \\
&= \|(A' - X_f^{\bar{T}}(t)(Y_f^{\bar{T}}(t))^\top) \odot 1_{\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T}\|^2 + \|A' \odot 1_{\bar{\mathcal{S}}_\mathcal{P}}\|^2 \\
&\overset{(19)}{=} \|A' - X_f^{\bar{T}}(t)(Y_f^{\bar{T}}(t))^\top\|^2
\end{aligned}$$

Since $\|A' - X_f^{\bar{T}}(t)(Y_f^{\bar{T}}(t))^\top\|^2$ is non-increasing, the function $\|A - X_f(t) Y_f(t)^\top\|^2$ is non-increasing as well. Last but not least, $(X_f(1), Y_f(1))$ is indeed a global minimizer since it satisfies all the optimality condition in Corollary B.1 by definition of $X_f^P(t), Y_f^P(t), P \in \{\bar{T}, C\}$. □

We now have all the elements to present the proof of the main result.

Proof of Theorem 4.2. Given any initial point (X^0, Y^0), we make the assumption that for all $P \in \mathcal{P}^\star$, either $X_{R_P,P}^0$ or $Y_{C_P,P}^0$ has full row rank. Indeed, if there exists $P \in \mathcal{P}^\star$ that does not have this property, we can employ Lemma C.5 to follow a continuous path along which the product of $XY^\top = X^0 (Y^0)^\top$ does not change (thus, the function $L(X, Y)$ is constant) and arrive at a point satisfying this additional assumption.

With this additional assumption, one can employ Lemma C.8 to build a continuous path $f_1(t) = (X_1(t), Y_1(t))$, such that $t \mapsto L(X_1(t), Y_1(t))$ is non-increasing, that connects (X^0, Y^0) to a point (X^1, Y^1) satisfying:

$$(A - X^1 (Y^1)^\top) \odot 1_{\mathcal{S}_T} = \mathbf{0}.$$

Again, one can assume that $\forall P \in \mathcal{P}^\star$, $X_{R_P,P}^1$ or $Y_{C_P,P}^1$ has full row rank (one can invoke Lemma C.5 one more time). Therefore, (X^1, Y^1) satisfies the conditions of Lemma C.9 . Hence, there exists a continuous path $f_2(t) = (X_2(t), Y_2(t))$ that makes $L(X_2(t), Y_2(t))$ non-increasing and that connects (X^1, Y^1) to (X^*, Y^*), a global minimizer.

Finally, since the concatenation of f_1 and f_2 satisfies the assumptions of Lemma C.1, we can conclude that there is no spurious local valley in the landscape of $\|A - XY^\top\|^2$. □

C.4 Proof of Theorem 4.3

The following corollary is necessary for the proof of Theorem 4.3 (it will be proved in Section D.3).

Corollary C.2. *Given $I \in \{0,1\}^{m\times r}$, $J \in \{0,1\}^{n\times r}$, consider T, $\mathcal{S}_T$ as in Definition 3.2. Given any feasible point (X,Y) in which $\forall P \in \mathcal{P}^\star$, $X_{R_P,P}$ or $Y_{C_P,P}$ has full row rank.*

For any B satisfying $\texttt{supp}(B) \subseteq \mathcal{S}_T$, there exists (X^*, Y^*) such that:

1) $\texttt{supp}(X^*) \subseteq I_T, \texttt{supp}(Y^*) \subseteq J_T$.

2) $X^*(Y^*)^\top = B$.

3) $\|X_T - X^*\|^2 + \|Y_T - Y^*\|^2 \leq \mathcal{C}\|X_T Y_T^\top - B\|^2$.

where $\mathcal{C} = \max\limits_{P\in\mathcal{P}^\star}\left(\max\left(\left|\left|\left|X^\dagger_{R_P,P}\right|\right|\right|^2, \left|\left|\left|Y^\dagger_{C_P,P}\right|\right|\right|^2\right)\right)$.

Sketch of the proof of Theorem 4.3. To prove this theorem, we proceed through two main steps:

1) First, we show that any local minimum satisfies:
$$(A - XY^\top) \odot 1_{\mathcal{S}_T} = \mathbf{0} \tag{23}$$
2) Second, we show that if a point (X,Y) satisfies the full-rank condition of the theorem and Equation (23), then it cannot be a spurious local minimum.

Finally, if (X,Y) does not satisfy the full-rank conditions, Lemma C.5 yields a path along which L is constant that joins (X,Y) to some (X',Y') which satisfies the full-rank conditions. □

Proof. As mentioned in the sketch of the proof, given any (X,Y) that does not satisfy the full-rank conditions, Lemma C.5 shows the existence of a path f along which L is constant and f connects (X,Y) to some (X',Y') which satisfies the full-rank conditions. Therefore, this proof will be entirely devoted to show that if the full-rank conditions are satisfied with a feasible solution (X,Y), then (X,Y) cannot be a spurious local minimum. This fact will be shown by the two following steps:

FIRST STEP: Consider the function $L(X,Y)$, we have:
$$\begin{aligned} L(X,Y) &= \|A - XY^\top\|^2 \\ &= \|A - \sum_{P'\in\mathcal{P}^\star} X_{P'}Y_{P'}^\top - X_{\bar{T}}Y_{\bar{T}}^\top\|^2 \end{aligned}$$

If (X,Y) is truly a local minimum, then (X_P, Y_P) is also the local minimum of the following function:
$$L'(X_P,Y_P) = \|(A - \sum_{P'\neq P} X_{P'}Y_{P'}^\top - X_{\bar{T}}Y_{\bar{T}}^\top) - X_PY_P^\top\|^2$$

where L' is simply L but we optimize only with respect to (X_P,Y_P) and we keep fixed all the remaining coefficients. In other words, (X_P,Y_P) is a local minimum of the problem:
$$\begin{aligned} \underset{X'\in\mathbb{R}^{m\times r}, Y'\in\mathbb{R}^{n\times r}}{\text{Minimize}} \quad & L'(X',Y') = \|B - X'Y'^\top\|^2 \\ \text{Subject to:} \quad & \texttt{supp}(X') \subseteq I_P \\ & \texttt{supp}(Y') \subseteq J_P \end{aligned}$$

where $B = A - \sum_{P'\neq P} X_{P'}Y_{P'}^\top - X_{\bar{T}}Y_{\bar{T}}$. Since all columns of I_P (resp. of J_P) are identical, all rank-one contribution supports are totally overlapping. Thus, all local minima are global minima (Theorem 4.1). Global minima are attained when:
$$X_PY_P^\top = B \odot 1_{\mathcal{S}_P}$$

due to the expressivity of a CEC in Lemma B.1. Thus, for any $P \in \mathcal{P}^\star$, $\forall (i,j) \in \mathcal{S}_P$, we have:

$$0 = (B - X_P Y_P^\top)_{i,j} = (A - \sum_{P' \in \mathcal{P}^\star} X_{P'} Y_{P'}^\top - X_{\bar{T}} Y_{\bar{T}}^\top)_{i,j} = (A - XY^\top)_{i,j}$$

which implies Equation (23).

SECOND STEP: In this step, we assume that Equation (23) holds. Consider $X_T, X_{\bar{T}}^i, Y_T, Y_{\bar{T}}^i, i = 1, 2$ as in Definition 3.5. Let $A' = A \odot 1_{\bar{\mathcal{S}}_T}, I' = I_{\bar{T}}^1, J' = J_{\bar{T}}^1$.

We distinguish two cases. In the first case, $(X_{\bar{T}}^1, Y_{\bar{T}}^1)$ is an optimal solution of the instance of Problem (1) with (A', I', J'). Then, by Corollary B.1, (X, Y) is an optimal solution of Problem (1), hence it cannot be a spurious local minimum. We now focus on the second case, where $(X_{\bar{T}}^1, Y_{\bar{T}}^1)$ is *not* the optimal solution of the instance of Problem (1) with (A', I', J'). We show that in this case, in any neighborhood of (X, Y), there exists a point (X', Y') such that $\texttt{supp}(X') \subseteq I$, $\texttt{supp}(Y') \subseteq J'$ and $L(X, Y) > L(X', Y')$. Thus (X, Y) cannot be a local minimum.

Since $I_{\bar{T}}^1, J_{\bar{T}}^1$ satisfy the assumptions of Theorem 3.1, Problem (1) has no spurious local minima (Theorem 4.1). As $(X_{\bar{T}}^1, Y_{\bar{T}}^1)$ is not an optimal solution, it cannot be a local minimum either, i.e., by definition, in any neighborhood of $(X_{\bar{T}}^1, Y_{\bar{T}}^1)$, there exists a feasible solution $\tilde{X}, \tilde{Y}$, $\texttt{supp}(\tilde{X}) \subseteq I'$, such that $\texttt{supp}(\tilde{Y}) \subseteq J'$ and

$$\|A' - X_{\bar{T}}^1 (Y_{\bar{T}}^1)^\top\|^2 > \|A' - \tilde{X}\tilde{Y}^\top\|^2 \tag{24}$$

By Equation (19), we have:

$$\begin{aligned} \|A' - (X_{\bar{T}}^1)(Y_{\bar{T}}^1)^\top\|^2 &= \|(A' - (X_{\bar{T}}^1)(Y_{\bar{T}}^1)^\top) \odot 1_{\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T}\|^2 + \|A' \odot 1_{\bar{\mathcal{S}}_\mathcal{P}}\|^2 \\ &= \|(A - (X_{\bar{T}}^1)(Y_{\bar{T}}^1)^\top) \odot 1_{\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T}\|^2 + \|A \odot 1_{\bar{\mathcal{S}}_\mathcal{P}}\|^2 \end{aligned} \tag{25}$$

Similarly, we have:

$$\|A' - \tilde{X}\tilde{Y}^\top\|^2 = \|(A - \tilde{X}\tilde{Y}^\top) \odot 1_{\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T}\|^2 + \|A \odot 1_{\bar{\mathcal{S}}_\mathcal{P}}\|^2 \tag{26}$$

Thanks to Equations(24), (25) and (26), we have:

$$\|(A - (X_{\bar{T}}^1)(Y_{\bar{T}}^1)^\top) \odot 1_{\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T}\|^2 > \|(A - \tilde{X}\tilde{Y}^\top) \odot 1_{\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T}\|^2 \tag{27}$$

Consider the matrix:

$$B := \left(A - (\tilde{X} + X_{\bar{T}}^2)(\tilde{Y} + Y_{\bar{T}}^2)^\top\right) \odot 1_{\mathcal{S}_T}$$

Since $\texttt{supp}(B) \subseteq \mathcal{S}_T$ and $\forall P \in \mathcal{P}^\star$, $X_{R_P,P}$ or $Y_{C_P,P}$ has full row rank by assumption, by Corollary C.2, there exists $(X^\star, Y^\star)$ such that:

1) $\texttt{supp}(X^\star) \subseteq I_T, \texttt{supp}(Y^\star) \subseteq J_T$.
2) $X^\star (Y^\star)^\top = B$.
3) $\|X_T - X^\star\|^2 + \|Y_T - Y^\star\|^2 \leq \mathcal{C} \|X_T Y_T^\top - B\|^2$.

where $\mathcal{C} = \max_{P \in \mathcal{P}^\star} \left(\max \left(\left\| X_{R_P,P}^\dagger \right\|^2, \left\| Y_{C_P,P}^\dagger \right\|^2 \right) \right)$. We define the point$(X', Y')$ as:

$$X' = X'_T + (X')_{\bar{T}}^1 + X_{\bar{T}}^2, \qquad Y' = Y'_T + (Y')_{\bar{T}}^1 + Y_{\bar{T}}^2$$

with $X'_T := X^\star, Y'_T := Y^\star, (X')_{\bar{T}}^1 := \tilde{X}, (Y')_{\bar{T}}^1 := \tilde{Y}$. The point (X', Y') still satisfies Equation (23). Indeed,

$$\begin{aligned} (A - X'(Y')^\top) \odot 1_{\mathcal{S}_T} &= \left(A - X'_T (Y'_T)^\top - ((X')_{\bar{T}}^1 + X_{\bar{T}}^2)((Y')_{\bar{T}}^1 + Y_{\bar{T}}^2)^\top\right) \odot 1_{\mathcal{S}_T} \\ &= (B - X'_T (Y'_T)^\top) \odot 1_{\mathcal{S}_T} = \mathbf{0}. \end{aligned} \tag{28}$$

It is clear that (X', Y') satisfies $\texttt{supp}(X') \subseteq I$, $\texttt{supp}(Y') \subseteq J$ due to the support of its components (X'_T, Y'_T), $((X')^i_{\bar{T}}, (Y')^i_{\bar{T}})$, $i = 1, 2$. Moreover, we have:

$$\begin{aligned}
&\|A - X'(Y')^\top\|^2 \\
&= \|(A - X'(Y')^\top) \odot 1_{\mathcal{S}_T}\|^2 + \|(A - X'(Y')^\top) \odot 1_{\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T}\|^2 + \|(A - X'(Y')^\top) \odot 1_{\bar{\mathcal{S}}_\mathcal{P}}\|^2 \\
&\overset{(28)}{=} 0 + \|(A - (X')^1_{\bar{T}}((Y')^1_{\bar{T}})^\top) \odot 1_{\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T}\|^2 + \|A \odot 1_{\bar{\mathcal{S}}_\mathcal{P}}\|^2 \\
&\overset{(27)}{<} 0 + \|(A - X^1_{\bar{T}}(Y^1_{\bar{T}})^\top) \odot 1_{\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T}\|^2 + \|A \odot 1_{\bar{\mathcal{S}}_\mathcal{P}}\|^2 \\
&= \|A - XY^\top\|^2.
\end{aligned}$$

Lastly, we show that (X', Y') can be chosen arbitrarily close to (X, Y) by choosing $(\tilde{X}, \tilde{Y})$ close enough to $(X^1_{\bar{T}}, Y^1_{\bar{T}})$. For this, denoting $\epsilon := \|X^1_{\bar{T}} - \tilde{X}\|^2 + \|Y^1_{\bar{T}} - \tilde{Y}\|^2$, we first compute

$$\begin{aligned}
&\|X - X'\|^2 + \|Y - Y'\|^2 \\
&= \|X_T - X'_T\|^2 + \|Y_T - Y'_T\|^2 + \|X^1_{\bar{T}} - (X')^1_{\bar{T}}\|^2 + \|Y^1_{\bar{T}} - (Y')^1_{\bar{T}}\|^2 \\
&\leq \mathcal{C}\|X_T Y_T^\top - B\|^2 + \epsilon
\end{aligned}$$

Moreover, due to Equation (23), we have:

$$\begin{aligned}
\mathbf{0} &= (A - XY^\top) \odot 1_{\mathcal{S}_T} \\
&\overset{(17)}{=} (A - X_T Y_T^\top - \sum_{1 \leq i,j \leq 2} (X^i_{\bar{T}})(Y^j_{\bar{T}})^\top) \odot 1_{\mathcal{S}_T} \\
&= (A - \sum_{1 \leq i,j \leq 2} (X^i_{\bar{T}})(Y^j_{\bar{T}})^\top) \odot 1_{\mathcal{S}_T} - X_T Y_T^\top
\end{aligned}$$

Therefore, $X_T Y_T^\top = [A - (X^1_{\bar{T}} + X^2_{\bar{T}})(Y^1_{\bar{T}} + Y^2_{\bar{T}})^\top] \odot 1_{\mathcal{S}_T}$. We have:

$$\begin{aligned}
\|X_T Y_T^\top - B\|^2 &= \|[A - (X^1_{\bar{T}} + X^2_{\bar{T}})(Y^1_{\bar{T}} + Y^2_{\bar{T}})^\top] \odot 1_{\mathcal{S}_T} - B\|^2 \\
&= \|[(\tilde{X} + X^2_{\bar{T}})(\tilde{Y} + Y^2_{\bar{T}})^\top - (X^1_{\bar{T}} + X^2_{\bar{T}})(Y^1_{\bar{T}} + Y^2_{\bar{T}})^\top] \odot 1_{\mathcal{S}_T}\|^2 \\
&\leq \|(\tilde{X} + X^2_{\bar{T}})(\tilde{Y} + Y^2_{\bar{T}})^\top - (X^1_{\bar{T}} + X^2_{\bar{T}})(Y^1_{\bar{T}} + Y^2_{\bar{T}})^\top\|^2
\end{aligned}$$

When $\epsilon \to 0$, we have $\|(\tilde{X} + X^2_{\bar{T}})(\tilde{Y} + Y^2_{\bar{T}})^\top - (X^1_{\bar{T}} + X^2_{\bar{T}})(Y^1_{\bar{T}} + Y^2_{\bar{T}})^\top\| \to 0$. Therefore, with ϵ small enough, one have $\|X - X'\|^2 + \|Y - Y'\|^2$ can be arbitrarily small. This concludes the proof. □

C.5 Proof of Theorem 4.4

Let $l \neq k$ be another rank-one contribution support $\mathcal{S}_l$ that contains (i_1, j_1). WLOG, we can assume $i_1 = j_1 = 1, i_2 = j_2 = 2$ and $k = 1, l = 2$.

1) We define the matrix A as:

$$A_{i,j} = \begin{cases} 1 & \text{if } (i,j) \in \{(1,2),(2,1),(2,2)\} \\ 0 & \text{otherwise} \end{cases}$$

Thus, $A_{i_2,j_2} = A_{2,2} \neq 0$. We have $\inf_{X,Y} \|A - XY^\top\| = 0$ and this can be obtained setting $X^*_{1,1} = X^*_{2,1} = 1, Y^*_{1,1} = Y^*_{1,2} = 1, X^*_{1,2} = 1, Y^*_{2,1} = -1$ and the other coefficients equal to zero:

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}, X^* = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}, Y^* = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}.$$

Moreover, the infimum of $\|A - XY^\top\|$ can be attained if and only if $X_{2,1}Y_{1,1} = X_{1,1}Y_{2,1} = X_{2,1}Y_{2,1} = 1$ (since $(1,2),(2,1),(2,2)$ are only in $\mathcal{S}_1$). This implies $X_{1,1}Y_{1,1} =$

$(X_{2,1}Y_{1,1})(X_{1,1}Y_{2,1})/(X_{2,1}Y_{2,1}) = 1$. Therefore, $\sigma_1 := \sum_{p\neq 1} X_{1,p}Y_{1,p} = -X_{1,1}Y_{1,1} = -1$ at the minimum solution.

To show the existence of a spurious local valley, it is sufficient to choose S_1, S_2, S_3 as follows:

i) $S_1 = \{(X,Y) \mid \sigma_1 = -1\}$.
ii) $S_2 = \{(X,Y) \mid \sigma_1 = 1\}$.
iii) $S_3 = \{(X,Y) \mid \sigma_1 = 5\}$.

Similarly to what we have seen in the previous section, we are interested in the function:

$$\begin{aligned} g(\alpha) &= \inf_{\substack{\text{supp}(X)\subseteq I, \text{supp}(Y)\subseteq J, \\ \sigma_1 = \alpha}} \|A - XY\|^2 \\ &= \inf_{\substack{\text{supp}(X)\subseteq I, \text{supp}(Y)\subseteq J, \\ \sigma_1 = \alpha}} \sum_{i,j} (A_{i,j} - \sum_p X_{i,p}Y_{j,p})^2 \\ &\geq \inf_{\substack{\text{supp}(X)\subseteq I, \text{supp}(Y)\subseteq J, \\ \sigma_1 = \alpha}} \sum_{\substack{i\in\{1,2\}, \\ j\in\{1,2\}}} (A_{i,j} - \sum_p X_{i,p}Y_{j,p})^2 \\ &= \inf_{X_{1,1},X_{2,1},Y_{1,1},Y_{2,1}} (-\alpha - X_{1,1}Y_{1,1})^2 + (1 - X_{1,1}Y_{2,1})^2 \\ &\qquad + (1 - X_{2,1}Y_{1,1})^2 + (1 - X_{2,1}Y_{2,1})^2. \end{aligned}$$

The last quantity is the best rank-one approximation of the following 2×2 matrix:

$$A' = \begin{pmatrix} -\alpha & 1 \\ 1 & 1 \end{pmatrix}$$

which is given by:

$$\frac{2(\alpha+1)^2}{(\alpha^2+3) + \sqrt{(\alpha^2+3)^2 - 4(\alpha+1)^2}}.$$

Moreover, this infimum can be attained if $X_{i_1,k}, X_{i_2,k}, Y_{j_1,k}, Y_{j_2,k}$ are the first eigenvectors of A' and the other remaining coefficients are zero. Therefore, we have:

$$g(\alpha) = \frac{2(\alpha+1)^2}{(\alpha^2+3) + \sqrt{(\alpha^2+3)^2 - 4(\alpha+1)^2}}.$$

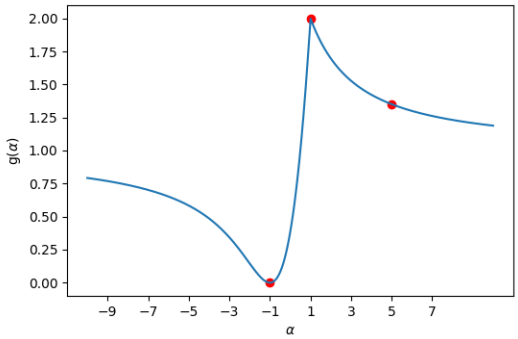


Figure 4: Graph of function $g(\alpha)$ and three considered σ_1.

We verify that the our choice satisfies all the conditions of Lemma C.2.

1) The minimum value of L is zero. It is only attained with $\sigma_1 = -1$ as shown. Thus, the global minimum belongs to S_1.
2) For any continuous function $r : [0,1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} : t \to (X(t), Y(t))$ we have $\sigma_1(t) = \sum_{p\neq 1} X(t)_{1,p}Y(t)_{1,p}$ is also continuous. If $(X(0), Y(0)) \in S_3, (X(1), Y(1)) \in S_1$ (i.e., $\sigma_1(0) = 5, \sigma_1(1) = -1$), then by the Mean Value Theorem, there must exist $t \in (0,1)$ such that $\sigma_1(t) = 1$, which means $(X(t), Y(t)) \in S_2$.

3) Since $g(1) > g(5) > g(-1)$ (Figure 4), we have $\inf_{\theta\in S_2} L(\theta) > \inf_{\theta\in S_3} L(\theta) > \inf_{\theta\in S_1} L(\theta)$.

The proof is concluded with the application of Lemma C.2.

2) To show that a spurious local minimum exists, we choose $A \in \mathbb{R}^{m\times n}$ as follows:

$$A_{i,j} = \begin{cases} a & \text{if } i = 1, j = 1 \\ b & \text{if } i = 2, j = 2 \\ 0 & \text{otherwise} \end{cases}$$

where $a > b > 0$. Thus, $A_{i_2,j_2} = A_{2,2} \neq 0$. It is again evident that $\inf_{X,Y} f(X,Y) = \inf_{X,Y} \|A - XY\|^2 = 0$. It can be attained by setting $X^*_{2,1} = X^*_{1,2} = 1, Y^*_{2,1} = b, Y^*_{1,2} = a$ and all the other coefficients equal to zero:

$$A = \begin{pmatrix} a & 0 & 0 & \dots & 0 \\ 0 & b & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}, X^* = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}, Y^* = \begin{pmatrix} 0 & a & 0 & \dots & 0 \\ b & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

Now, we will consider the following (X', Y') such that $X'_{1,1}Y'_{1,1} = a$ and all the other coefficients equal to zero:

$$X' = \begin{pmatrix} X'_{1,1} & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}, \; Y' = \begin{pmatrix} Y'_{1,1} & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}.$$

We will prove that (X', Y') is a spurious local minimum. As it holds $L(X', Y') = b^2 > 0$ it cannot be a global minimum. We will then show that (X', Y') is indeed a local minimum. We have:

$$\begin{aligned} &\|A - XY\|^2 = \sum_{i,j}(A_{i,j} - \sum_{p=1}^{r} X_{i,p}Y_{j,p})^2 \\ &\geq (A_{2,1} - \sum_{p=1}^{r} X_{2,p}Y_{1,p})^2 + (A_{1,2} - \sum_{p=1}^{r} X_{1,p}Y_{2,p})^2 + (A_{2,2} - \sum_{p=1}^{r} X_{2,p}Y_{2,p})^2 \\ &= (X_{2,1}Y_{1,1})^2 + (X_{1,1}Y_{2,1})^2 + (b - X_{2,1}Y_{2,1})^2 \\ &\geq 2(X_{1,1}Y_{1,1})|X_{2,1}Y_{2,1}| + (X_{2,1}Y_{2,1})^2 - 2bX_{2,1}Y_{2,1} + b^2 \quad \text{(Cauchy-Schwarz inequality)} \\ &\geq 2(X_{1,1}Y_{1,1} - b)|X_{2,1}Y_{2,1}| + b^2. \end{aligned}$$

Thus, there exists a neighborhood of (X', Y') such that $X_{1,1}Y_{1,1} - b > 0$ for all (X, Y) in that neighbourhood, since $X'_{1,1}Y'_{1,2} = a > b$. Therefore, $\|A - XY\|^2 \geq b^2 = L(X', Y') > 0$ in that neighborhood. This concludes the proof.

C.6 Proof for Theorem 4.5

Consider a minimizer (X^*, Y^*) of Problem (1) with $\texttt{supp}(X^*) \subseteq I, \texttt{supp}(Y^*) \subseteq J$. Such a solution exists due to the assumption of our theorem.

We only prove the result the initialization $(X, \mathbf{0}), \texttt{supp}(X) \subseteq I$. The case of the initialization $(\mathbf{0}, Y)$, $\texttt{supp}(Y) \subseteq J$ can be dealt with similarly.

To prove the theorem, it is sufficient to construct a continuous function $f(t) = (X_f(t), Y_f(t)) : [0, 1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$ such that:

1) $f(0) = (X, \mathbf{0})$.
2) $f(1) = (X^*, Y^*)$.

3) $L(f(t))$ is non-increasing w.r.t t.

Indeed, if such a function f exists, the sublevel set corresponding to $L(X, \mathbf{0})$ has both $(X, \mathbf{0})$ and (X^*, Y^*) in the same path-connected components (since $L(f(t))$ is non-increasing).

The function f is a concatenation of two functions $f_1 : [0, 1/2] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$, $f_2 : [1/2, 1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$, defined as follows:

1) $f_1(t) = ((1 - 2t)X + 2tX^*, \mathbf{0})$.
2) $f_2(t) = (X^*, (2t - 1)Y^*)$.

It is obvious that $f(0) = f_1(0) = (X, \mathbf{0})$ and $f(1) = f_2(1) = (X^*, Y^*)$. Moreover f is continuous since $f_1(1/2) = f_2(1/2) = (X^*, \mathbf{0})$. The fact that $L \circ f$ is non-increasing can be verified as:

1) Consider $L(f_1(t)) = \|A - ((1 - t)X + tX^*)\mathbf{0}^\top\|^2 = \|A\|^2$ is constant for $t \in [0, 1/2]$.
2) Consider $G(t) := L(f_2(t)) = \|A - (2t - 1)X^*Y^*\|^2$. This function is convex w.r.t t. Moreover, it attains a global minimum at $t = 1$ (since we assume that (X^*, Y^*) is a global minimizer of Problem 1). As a result, $t \mapsto G(t)$ is non-increasing on $[1/2, 1]$.

The function $L \circ f$ is thus, non-increasing on $[0, 1]$.

Remark C.1. *Theorem 4.5 confirms that the initialization at $(X, \mathbf{0})$ and $(\mathbf{0}, Y)$ always keeps the solution out of any spurious local valley for Problem* (1) *(as long as the infimum is attained). Nevertheless, whether this initial point allows first-order methods to converge to global minimum is not guaranteed. This phenomenon is illustrated with a scalar function in Figure 1b with the maximal point separating two valleys. Intuitively, the points $(X, \mathbf{0})$ and $(\mathbf{0}, Y)$ stand high enough (like the top of the mountain) to avoid being trapped in a local valley.*

D Proofs for other intermediate lemmas

D.1 Proof of Lemma C.6

Proof of Lemma C.6. WLOG, we assume that $m \leq r$. If X has full row rank, then one can choose constant function $f(t) = (X, Y)$ to satisfy the conditions of the lemma. Therefore, we can focus on the case where $\texttt{rank}(X) = q < m$. WLOG, we can assume that the first q columns of X $(X_1, \ldots, X_q)$ are linearly independent. The remaining columns of X can be expressed as:

$$X_k = \sum_{i=1}^{q} \alpha_i^k X_i, \forall q < k \leq r$$

We define Y' by their columns as follow:

$$Y_i' = \begin{cases} Y_i + \sum_{k=q+1}^{r} \alpha_i^k Y_k & \text{if } i \leq q \\ 0 & \text{otherwise} \end{cases}$$

By construction, we have $XY^\top = XY'^\top$. We define the function $f_1 : [0, 1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$ as:

$$f_1(t) = (X, (1 - t)Y + tY')$$

This function will not change the value of f since we have:

$$X((1 - t)Y^\top + tY'^\top) = (1 - t)XY^\top + tXY'^\top = XY^\top.$$

Let X' be a matrix whose first q columns are identical to that of X and $\texttt{rank}(X') = m$. The second function f_2 defined as:

$$f_2(t) = ((1 - t)X + tX', Y')$$

also has the product $XY^\top$ unchanged (since first q columns of $X(t)$ are constant and last $r - q$ rows of Y' are zero). Moreover, $f_2(0) = (X', Y')$ where X' has full row rank. Therefore, the concatenation of two functions f_1 and f_2 (and shrink t by a factor of 2) are the desired function f. □

D.2 Proof of Lemma C.7

To prove this lemma, we introduce and prove two other lemmas first.

Lemma D.1. *Let $X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{n\times r}, \min(m,n) \leq r$ and assume that X or Y has full row rank. Given any continuous function $g : [0,1] \to \mathbb{R}^{m\times n}$ in which $g(0) = XY^\top$, there exists a continuous function $f : [0,1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} : f(t) = (X_f(t), Y_f(t))$ (X_f, Y_f are also functions) such that:*

1) $X_f(0) = X, Y_f(0) = Y$.

2) $g(t) = X_f(t)Y_f(t)^\top, \forall t \in [0,1]$.

3) $\|X_f(z) - X_f(t)\|^2 + \|Y_f(z) - Y_f(t)\|^2 \leq \mathcal{C}\|g(z) - g(t)\|^2$

where $\mathcal{C} = \max_{P\in\mathcal{P}^\star} \left(\max\left(\left|\left|\left|X^\dagger\right|\right|\right|^2, \left|\left|\left|Y^\dagger\right|\right|\right|^2\right)\right)$.

Proof. WLOG, we can assume that X has full row rank. The desired function can be defined as:

$$\begin{aligned} X_f(t) &= X \\ Y_f(t) &= Y + (g(t) - g(0))^\top (XX^\top)^{-1} X = Y + (X^\dagger(g(t) - g(0)))^\top \end{aligned} \tag{29}$$

where $X^\dagger = X^\top(XX^\top)^{-1}$ the pseudo-inverse of X. The function Y_f is well-defined due to the assumption of X being full row rank. It is immediate for the first two constraints. The third one is satisfied since:

$$\begin{aligned} \|X_f(t_1) - X_f(t_2)\|^2 + \|Y_f(t_1) - Y_f(t_2)\|^2 &= \|Y_f(t_1) - Y_f(t_2)\|^2 \\ &= \|(X^\dagger(g(t_1) - g(t_2)))^\top\|^2 \\ &\leq \left|\left|\left|X^\dagger\right|\right|\right|^2 \|g(t_1) - g(t_2)\|^2 \\ &\leq \mathcal{C}\|g(t_1) - g(t_2)\|^2 \end{aligned}$$

□

Lemma D.2. *Consider $I \in \{0,1\}^{m\times r}$, $J \in \{0,1\}^{r\times n}$ and $P \in \mathcal{P}$, $\mathcal{S}_P$ as in Definition 3.2. Assume that P is complete. For any feasible point (X,Y) whose $X_{R_P,P}$ or $Y_{C_P,P}$ (R_P, C_P are defined Definition 3.2) has full rank and any continuous function $g : [0,1] \to \mathbb{R}^{m\times n}$ satisfying $\mathtt{supp}(g(t)) \subset \mathcal{S}_P$ and $g(0) = X_P Y_P^\top$, there exists a continuous function $f : [0,1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$: $f(t) = (X_f(t), Y_f(t))$ such that:*

1) $\mathtt{supp}(X_f(t)) \subseteq I_P, \mathtt{supp}(Y_f(t)) \subseteq J_P$.

2) $X_f(0) = X_P, Y_f(0) = Y_P$.

3) $g(t) = X_f(t)Y_f(t)^\top, \forall t \in [0,1]$.

4) $\|X_f(z) - X_f(t)\|^2 + \|Y_f(z) - Y_f(t)\|^2 \leq \mathcal{C}\|g(z) - g(t)\|^2$.

where $\mathcal{C} = \max_{P\in\mathcal{P}^\star} \left(\max\left(\left|\left|\left|X^\dagger_{R_P,P}\right|\right|\right|^2, \left|\left|\left|Y^\dagger_{C_P,P}\right|\right|\right|^2\right)\right)$.

Proof. WLOG, we assume that $P = [\![|P|]\!], R_P = [\![|R_P|]\!], C_P = [\![|C_P|]\!]$. WLOG, we can assume $|P| \geq |R_P|$ and $X_{R_P,P}$ is full row rank (due to the hypothesis and the fact that P is complete).

A continuous function $f(t)$ that satisfies the first constrain must have the form:

$$X_f(t) = \begin{pmatrix} X'(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, Y_f(t) = \begin{pmatrix} Y'(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

where $X' : [0,1] \to \mathbb{R}^{|R_P|\times|P|}$ and $Y' : [0,1] \to \mathbb{R}^{|C_P|\times|P|}$ are continuous functions.

Moreover, if $g : [0,1] \to \mathbb{R}^{m\times n}$ satisfying $\texttt{supp}(g(t)) \subseteq \mathcal{S}_T$, then g has to have the form:

$$g(t) = \begin{pmatrix} A'(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

where $A'(t) : [0,1] \to \mathbb{R}^{|R_P|\times|C_P|}$ is a continuous function. Since $g(0) = X_P Y_P^\top$, $A'(0) = (X_{R_P,P})(Y_{C_P,P})^\top$. Thus, it is sufficient to find $X'(t)$ and $Y'(t)$ such that:

i) $X'(0) = X_{R_P,P}, Y'(0) = Y_{C_P,P}$ (to satisfy the second condition).

ii) $A'(t) = X'(t)Y'(t)$ to satisfy the third condition because:

$$\begin{aligned} X_f(t)Y_f(t)^\top &= \begin{pmatrix} X'(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\begin{pmatrix} Y'(t)^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\ &= \begin{pmatrix} X'(t)Y'(t)^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\ &= \begin{pmatrix} A'(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = g(t) \end{aligned}$$

iii) $\|X'(z) - X'(t)\|^2 + \|Y'(z) - Y'(t)\|^2 \leq \mathcal{C}\|A'(z) - A'(t)\|^2 = \mathcal{C}\|g(z) - g(t)\|^2$ to satisfy the last condition since:

$$\|X_f(z) - X_f(t)\|^2 + \|Y_f(z) - Y_f(t)\|^2 = \|X'(z) - X'(t)\|^2 + \|Y'(z) - Y'(t)\|^2$$

Such functions exist thanks Lemma D.1 (since we assume $X_{R_P,P}$ has full rank). □

Proof of Lemma C.7. We prove by induction on the size of subset $\mathcal{P}' \subseteq \mathcal{P}^\star$ that: let $T' = \cup_{P\in\mathcal{P}'} P$, for any continuous function $g : [0,1] \to \mathbb{R}^{m\times n}$ satisfying $\texttt{supp}(g(t)) \subset \mathcal{S}_{\mathcal{P}'}$ and $g(0) = X_{T'}Y_{T'}^\top$, there exists a continuous function $f_{\mathcal{P}'} : [0,1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} : f_{\mathcal{P}'} = (X_f^{\mathcal{P}'}(t), Y_f^{\mathcal{P}'}(t))$ such that:

1) $\texttt{supp}(X_f^{\mathcal{P}'}(t)) \subseteq I_{T'}, \texttt{supp}(Y_f^{\mathcal{P}'}(t)) \subseteq J_{T'}$.

2) $X_f^{\mathcal{P}'}(0) = X_{T'}, Y_f^{\mathcal{P}'}(0) = Y_{T'}$.

3) $g(t) = X_f^{\mathcal{P}'}(t)Y_f^{\mathcal{P}'}(t)^\top, \forall t \in [0,1]$.

4) $\|X_f^{\mathcal{P}'}(z) - X_f^{\mathcal{P}'}(t)\|^2 + \|Y_f^{\mathcal{P}'}(z) - Y_f^{\mathcal{P}'}(t)\|^2 \leq \mathcal{C}'\|g(z) - g(t)\|^2$.

where $\mathcal{C}' = \max_{P\in\mathcal{P}'}\left(\max\left(\left\|\left\|X_{R_P,P}^\dagger\right\|\right\|^2, \left\|\left\|Y_{C_P,P}^\dagger\right\|\right\|^2\right)\right)$. Applying this result with $\mathcal{P}' = \mathcal{P}^\star$ will yield the proof of this lemma.

By Lemma D.2 the result is true if $|\mathcal{P}'| = 1$ (due to the hypothesis $\forall P \in \mathcal{P}^\star, X_{R_P,P}$ or $Y_{C_P,P}$ has full row rank). Assume the result is true if $|\mathcal{P}'| = p$, where $1 \leq p < |\mathcal{P}^\star|$, and consider the case where $|\mathcal{P}'| = p + 1$.

Consider $P \in \mathcal{P}'$ and partition $\mathcal{P}'$ into $\mathcal{P}'' = \mathcal{P}' \setminus \{P\}$ and P. Let $T'' = T' \setminus P$, consider

$$h_1(t) = (g(t) - X_P Y_P^\top) \odot 1_{\mathcal{S}_{T''}}, \qquad h_2(t) = X_P Y_P^\top \odot 1_{\mathcal{S}_{T''}} + g(t) \odot 1_{\mathcal{S}_P \setminus \mathcal{S}_{T''}}$$

We verify that the function $h_1(t)$ satisfying the hypotheses to use induction step: h_1 continuous, $\texttt{supp}(h_1(t)) \subseteq \mathcal{S}_{T''}$ and finally $h_1(0) = (g(0) - X_P Y_P^\top) \odot 1_{\mathcal{S}_{T''}} = X_{T''}Y_{T''}^\top \odot 1_{\mathcal{S}_{T''}} = X_{T''}Y_{T''}^\top$. Therefore, there exists a function $(X_f^1(t), Y_f^1(t))$ such that:

1) $\texttt{supp}(X_f^1(t)) \subseteq I_{T''}, \texttt{supp}(Y_f^1(t)) \subseteq J_{T''}$.

2) $X_f^1(0) = X_{T''}, Y_f^1(0) = Y_{T''}$.

3) $h_1(t) = X_f^1(t)Y_f^1(t)^\top, \forall t \in [0,1]$.

4) $\|X_f^1(z) - X_f^1(t)\|^2 + \|Y_f^1(z) - Y_f^1(t)\|^2 \leq \mathcal{C}''\|h_1(z) - h_1(t)\|^2$.

where $\mathcal{C}'' = \max_{P' \in \mathcal{P}''} \left(\max \left(\left\| X^\dagger_{R_{P'},P'} \right\|^2, \left\| Y^\dagger_{C_{P'},P'} \right\|^2 \right) \right)$.

On the other hand, $h_2(t)$ satisfies the assumptions of Lemma D.2: $h_2(t)$ is continuous and $\texttt{supp}(h_2(t)) = \texttt{supp}(X_P Y_P^\top \odot 1_{\mathcal{S}_{T''}} + g(t) \odot 1_{\mathcal{S}_P \setminus \mathcal{S}_{T''}}) \subseteq \texttt{supp}(X_P Y_P^\top) \cup (\mathcal{S}_P \setminus \mathcal{S}_{T''}) = \mathcal{S}_P$.

In addition, since $g(0) \odot 1_{\mathcal{S}_P \setminus \mathcal{S}_{T''}} = (X_{T'} Y_{T'}^\top) \odot 1_{\mathcal{S}_P \setminus \mathcal{S}_{T''}} = (X_{T''} Y_{T''}^\top + X_P Y_P^\top) \odot 1_{\mathcal{S}_P \setminus \mathcal{S}_{T''}} = X_P Y_P^\top \odot 1_{\mathcal{S}_P \setminus \mathcal{S}_{T''}}$, we have $h_2(0) = X_P Y_P^\top \odot 1_{\mathcal{S}_{T''}} + g(t) \odot 1_{\mathcal{S}_P \setminus \mathcal{S}_{T''}} = X_P Y_P^\top \odot (1_{\mathcal{S}_{T''}} + 1_{\mathcal{S}_P \setminus \mathcal{S}_{T''}}) = X_P Y_P^\top$. Therefore, there exists a function $(X_f^2(t), Y_f^2(t))$ such that:

1) $\texttt{supp}(X_f^2(t)) \subseteq I_P, \texttt{supp}(Y_f^2(t)) \subseteq J_P$.

2) $X_f^2(0) = X_P, Y_f^2(0) = Y_P$.

3) $h_2(t) = X_f^2(t) Y_f^2(t)^\top, \forall t \in [0,1]$.

4) $\|X_f^2(z) - X_f^2(t)\|^2 + \|Y_f^2(z) - Y_f^2(t)\|^2 \leq \max \left(\left\| X^\dagger_{R_P,P} \right\|^2, \left\| Y^\dagger_{C_P,P} \right\|^2 \right) \|h_2(z) - h_2(t)\|^2$.

Lastly, the functions $(X_f^{\mathcal{P}'}(t), Y_f^{\mathcal{P}'}(t))$ can be constructed as:

$$X_f^{\mathcal{P}'}(t) = X_f^1(t) + X_f^2(t), \qquad Y_f^{\mathcal{P}'}(t) = Y_f^1(t) + Y_f^2(t)$$

We verify the validity of this construction. The first condition is clear thanks to the construction of $X_f^i(t), Y_f^i(t), i = 1, 2$. The second condition is satisfied as well since:

$$\begin{aligned} X_f^{\mathcal{P}'}(0) &= X_f^1(0) + X_f^2(0) = X_{T''} + X_P = X_{T'} \\ Y_f^{\mathcal{P}'}(0) &= Y_f^1(0) + Y_f^2(0) = Y_{T''} + Y_P = Y_{T'} \end{aligned}$$

The third condition is satisfied as well due to the definition of $h_i(t), i = 1, 2$:

$$\begin{aligned} X_f^{\mathcal{P}'}(t) Y_f^{\mathcal{P}'}(t)^\top &= X_f^1(t) Y_f^1(t)^\top + X_f^2(t) Y_f^2(t)^\top \\ &= h_1(t) + h_2(t) \\ &= (g(t) - X_P Y_P^\top) \odot 1_{\mathcal{S}_{T''}} + X_P Y_P^\top \odot 1_{\mathcal{S}_{T''}} + g(t) \odot 1_{\mathcal{S}_P \setminus \mathcal{S}_{T''}} \\ &= g(t) \odot (1_{\mathcal{S}_{T''}} + 1_{\mathcal{S}_P \setminus \mathcal{S}_{T''}}) = g(t) \end{aligned}$$

Finally, the fourth condition holds since:

$$\begin{aligned} &\|X_f(z) - X_f(t)\|^2 + \|Y_f(z) - Y_f(t)\|^2 \\ &= \|X_f^1(z) - X_f^1(t)\|^2 + \|Y_f^1(z) - Y_f^1(t)\|^2 + \|X_f^2(z) - X_f^2(t)\|^2 + \|Y_f^2(z) - Y_f^2(t)\|^2 \\ &\leq \mathcal{C}''\|h_1(z) - h_1(t)\|^2 + \max \left(\left\| X^\dagger_{R_P,P} \right\|^2, \left\| Y^\dagger_{C_P,P} \right\|^2 \right) \|h_2(z) - h_2(t)\|^2 \\ &\leq \mathcal{C}'(\|h_1(z) - h_1(t)\|^2 + \|h_2(z) - h_2(t)\|^2) \\ &= \mathcal{C}'(\|(g(z) - g(t)) \odot 1_{\mathcal{S}_{T''}}\|^2 + \|(g(z) - g(t)) \odot 1_{\mathcal{S}_P \setminus \mathcal{S}_{T''}}\|^2) \\ &= \mathcal{C}'\|g(z) - g(t)\|^2 \end{aligned}$$

□

D.3 Proof of Corollary C.2

Corollary C.2 is a direct result of Lemma C.7.

Proof. Consider the function $g(t) = (1-t)X_T Y_T^\top + tB$. We have $g(t)$ is continuous, $g(0) = X_T Y_T^\top$ and $\texttt{supp}(g(t)) \subseteq \texttt{supp}(X_T Y_T^\top) \cup \subset (B) = \mathcal{S}_T$. Together with the hypothesis that $\forall P \in \mathcal{P}^\star, X_{R_P,P}$ or $Y_{C_P,P}$ has full row rank, there exists a function $(X_f(t), Y_f(t))$ such that:

1) $\texttt{supp}(X_f(t)) \subseteq I_T, \texttt{supp}(Y_f(t)) \subseteq J_T$.
2) $X_f(0) = X_T, Y_f(0) = Y_T$.
3) $g(t) = X_f(t) Y_f(t)^\top, \forall t \in [0,1]$.
4) $\|X_f(z) - X_f(t)\|^2 + \|Y_f(z) - Y_f(t)\|^2 \leq \mathcal{C}\|g(z) - g(t)\|^2$.

by using Lemma C.7. We can choose $X^* = X_f(1), Y^* = Y_f(1)$. The first condition is immediately satisfied. The second condition holds since: $X^*(Y^*)^\top = X_f(1)Y_f(1)^\top = g(1) = B$. The last condition results from:

$$\begin{aligned}\|X_T - X^*\|^2 + \|Y_T - Y^*\|^2 &= \|X_f(1) - X_f(0)\|^2 + \|Y_f(1) - Y_f(0)\|^2 \\ &\leq \mathcal{C}\|g(0) - g(1)\|^2 \\ &= \mathcal{C}\|X_T Y_T^\top - B\|^2\end{aligned}$$

This concludes the proof. □