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

Quoc-Tung Le, Elisa Riccietti, Rémi Gribonval

SPURIOUS VALLEYS, NP-HARDNESS, AND TRACTABILITY OF SPARSE MATRIX FACTORIZATION WITH FIXED SUPPORT

QUOC-TUNG LE*, ELISA RICCIETTI*, AND REMI GRIBONVAL*

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.

Key words. Sparse Matrix Factorization, Fixed Support, NP-hardness, Landscape

AMS subject classifications. 15A23, 90C26

1. Introduction.

Matrix factorization with sparsity constraints is the problem of approximating a (possibly dense) matrix as the product of two or more sparse factors. It is playing an important role in many domains and applications such as dictionary learning and signal processing [39, 35, 34], linear operator acceleration [27, 26, 5], deep learning [10, 11, 6], to mention only a few. Given a matrix Z, sparse matrix factorization can be expressed as the optimization problem:

$$(1.1)\qquad \begin{aligned} &\underset{X^1,\ldots,X^N}{\text{Minimize}}\ \|Z - X^1 \ldots X^N\|_F^2 \\ &\text{subject to: constraints on } \mathtt{supp}(X_i),\ \forall 1 \leq i \leq N \end{aligned}$$

where $\mathtt{supp}(X) := \{(i,j) \mid X_{i,j} \neq 0\}$ is the set of indices whose entries are nonzero.

For example, one can employ generic sparsity constraints such as $|\mathtt{supp}(X_i)| \leq k_i, 1 \leq i \leq N$ where k_i controls the sparsity of each factor. More structured types of sparsity (for example, sparse rows/ columns) can also be easily encoded since the notion of support $\mathtt{supp}(X)$ captures completely the sparsity structure of a factor.

Despite its importance, Problem (1.1) is challenging due to its non-convexity as well as the discrete nature of $\mathtt{supp}(X_i)$ (which can lead to an exponential number of supports to consider). Existing algorithms to tackle Problem (1.1) directly comprise heuristics such as Proximal Alternating Linearization Minimization (PALM) [3, 27] and its variants [23].

In this work, we consider a restricted class of instances of Problem (1.1), in which just two factors are considered ($N = 2$) and with *prescribed supports*. We call this problem *fixed support (sparse) matrix factorization* (FSMF). In details, given a matrix $A \in \mathbb{R}^{m\times n}$, we look for two sparse factors X, Y that solve the following problem:

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

*Univ Lyon, ENS de Lyon, UCBL, CNRS, Inria, LIP, F-69342, LYON Cedex 07, France (quoc-tung.le@ens-lyon.fr, elisa.riccietti@ens-lyon.fr, remi.gribonval@inria.fr).

where $\|\cdot\|$ is the Frobenius norm, $I \subseteq \llbracket m \rrbracket \times \llbracket r \rrbracket$, $J \subseteq \llbracket n \rrbracket \times \llbracket r \rrbracket$[1] are given support constraints, i.e., $\texttt{supp}(X) \subseteq I$ implies that $\forall (i,j) \notin I, X_{ij} = 0$.

The main aim of this work is to investigate the theoretical properties of (FSMF). 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 the following reasons:

1. The asymptotic behaviour of heuristics such as PALM [3, 27] when applied to Problem (1.1) can be characterized by studying the behaviour of the method on an instance of (FSMF). Indeed, PALM updates the factors alternatively by a projected gradient step onto the set of the constraints. It is experimentally observed that for many instances of the problem, the support becomes constant after a certain number of iterations. Let us illustrate this on an instance of Problem (1.1) with $N = 2, X^i \in \mathbb{R}^{100\times 100}, i = 1,2$ and the constraints $|\texttt{supp}(X^i)| \leq 1000, i = 1,2$. In this setting, running PALM is equivalent to an iterative method in which we consecutively perform one step of gradient descent for *each* factor, while keeping the other fixed, and project that factor onto $\{X \mid X \in \mathbb{R}^{100\times 100}, |\texttt{supp}(X)| \leq 1000\}$ by simple hard-thresholding. Figure 1 illustrates the evolution of the difference between the support of each factor before and after each iteration of PALM through 1000 iterations (the difference between two sets B_1 and B_2 is measured by $|(B_1 \setminus B_2) \cup (B_2 \setminus B_1)|$). We observe that when the iteration counter is large enough, the factor supports do not change (or equivalently they become *fixed*): further iterations of the algorithm simply optimize an instance of (FSMF). Therefore, to develop a more precise understanding of the possible convergence of PALM in such a context, it is necessary to understand properties of (FSMF). For instance, we show that PALM can potentially lead to iterates (X_n^1, X_n^2) diverging to infinity due to the presence of a *spurious local valley* in the landscape of $L(X,Y)$ (cf Remark 4.22). This is not in conflict with the convergence results for PALM in this context [3, 27] since these are established *under the assumption of bounded iterates*.

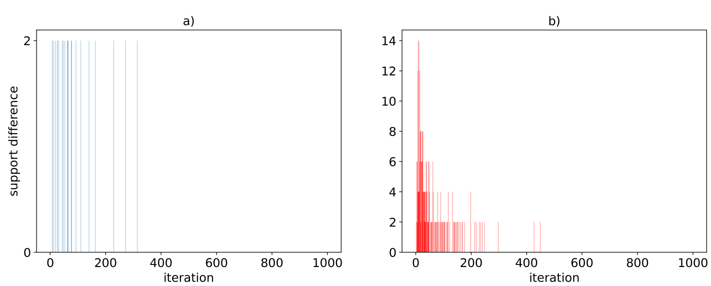

Fig. 1. *Support change of the first (a) and the second (b) factors during PALM.*

2. While (FSMF) is just a class of the general problem (1.1), its coverage includes many other interesting problems:
   - **Low rank matrix approximation (LRMA) [12]:** By taking $I = \llbracket m \rrbracket \times \llbracket r \rrbracket$, $J = \llbracket n \rrbracket \times \llbracket r \rrbracket$, addressing (FSMF) is equivalent to looking for the best rank r matrix approximating A, cf. Figure 2(a). We will refer to this instance in the following as the full support case. This problem is known to be polynomially tractable, cf. Section 3. This work enlarges the family of supports for which (FSMF) remains tractable.

[1] $\forall m \in \mathbb{N}$, $\llbracket m \rrbracket := \{1, \ldots, m\}$

- **LU decomposition [17, Chapter 3.2]:** Considering $m = n = r$ and $I = J = \{(i,j) \mid 1 \le j \le i \le n\}$, it is easy to check that (FSMF) is equivalent to factorizing A into a lower and an upper triangular matrix (X and Y respectively, cf. Figure 2(b)), and in this case, the *infimum* of (FSMF) is always zero. It is worth noticing that there exists a non-empty set of matrices for which this infimum is not attained (or equivalently matrices which do not admit the **LU** decomposition [17]). This behaviour will be further discussed in Section 2 and Section 4. More importantly, our analysis of (FSMF) will cover the non-zero infimum case as well.

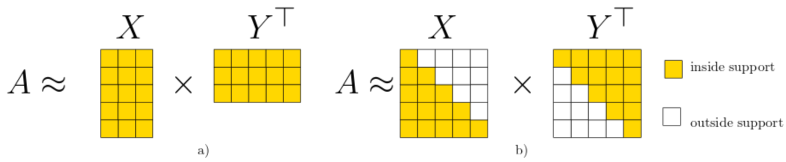

Fig. 2. *Illustrations for(a) LRMA and (b)* **LU** *decomposition as instances of* (FSMF).

- **Butterfly structure and fast transforms [10, 6, 11, 26, 5]:** Many linear operators admit fast algorithms since their associated matrices can be written as a product of sparse factors whose supports are known to possess the *butterfly structure* (and they are *known* in advance). This is the case for instance of the Discrete Fourier Transform (DFT) or the Hadamard transform (HT). For example, a Hadamard transform of size $2^N \times 2^N$ can be written as the product of N factors of size $2^N \times 2^N$ whose factors have two non-zero coefficients per row and per column. Figure 3 illustrates such a factorization for $N = 3$. Although

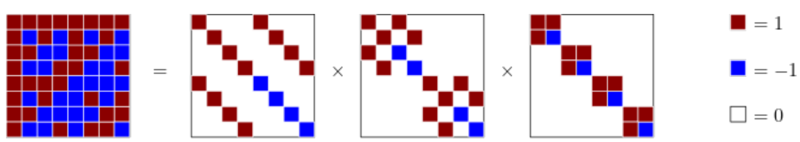

Fig. 3. *The factorization of the Hadamard transform of size* 8×8 ($N = 3$).

  our analysis of (FSMF) only deals with $N = 2$, the butterfly structure allows one to reduce to the case $N = 2$ in a recursive[2] manner [25, 44].
- **Hierarchical $\mathcal{H}$-matrices [19, 20]:** We prove in Appendix E that the class of hierarchically off-diagonal low-rank (HODLR) matrices (defined in [1, Section 3.1], [19, Section 2.3]), a subclass of hierarchical $\mathcal{H}$-matrices, can be expressed as the product of two factors with fixed supports, that are illustrated on Figure 4. Therefore, the task of finding the best $\mathcal{H}$-matrix from this class to approximate a given matrix is reduced to (FSMF).
- **Matrix completion:** We show that matrix completion can be reduced to (FSMF), which is the main result of Section 2.

Our aim is to then study the theoretical properties of (FSMF) and in particular to assess its difficulty. This leads us to consider four complementary aspects.

First, we show the NP-hardness of (FSMF). While this result contrasts with the theory established for coefficient recovery with a fixed support in the classical sparse

[2] While revising this manuscript we heard about the work of Dao et al [9] introducing the "Monarch" class of structured matrices, essentially corresponding to the first stage of the recursion from [25, 44].

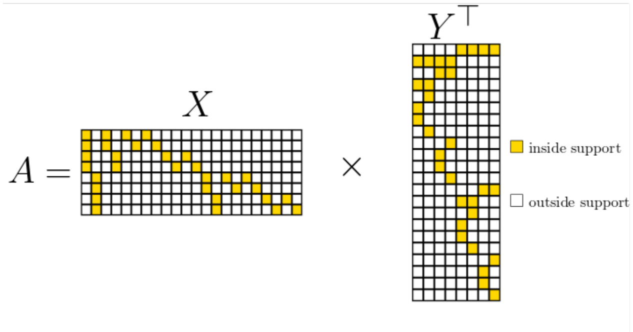

FIG. 4. *Two fixed supports for factors of a HODLR matrix of size* 8×8 *illustration based on analysis of Appendix* E.

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) [41, 36], weighted low rank [15] and matrix completion [15] were all proved to be NP-hard. We prove the NP-hardness by reduction from the 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.

Second, we show that besides its NP-hardness, problem (FSMF) also shares some properties with another hard problem: low-rank tensor approximation [37]. Similarly to the classical example of [37], which shows that the set of rank-two tensors is not closed, we show that there are support constraints I, J such that the set of matrix products $XY^\top$ with "feasible" (X, Y) (i.e., $\{XY^\top \mid \mathtt{supp}(X) \subseteq I, \mathtt{supp}(Y) \subseteq J\}$), is not a closed set. Important examples are the supports (I, J) for which (FSMF) corresponds to **LU** matrix factorization. For such support constraints, there exists a matrix A such that the infimum of $L(X, Y)$ is zero and can only be approached if either X or Y have at least an arbitrarily large coefficient. This is precisely one of the settings leading to a diverging behavior of PALM (cf Remark 4.22).

Third, we show that despite the hardness of (FSMF) in the general case, many pairs of support constraints (I, J) make the problem solvable by an effective direct algorithm based on the block 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 (FSMF) 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 an optimal solution using blockwise SVDs. This technique can be seen in many sparse representations of matrices (for example, hierarchical $\mathcal{H}$-matrices [19, 20]) to allow fast matrix-vector and matrix-matrix multiplication.

The fourth contribution of this paper is the study of the landscape of the objective function L of (FSMF). 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 global landscape of the loss functions for matrix decomposition related problems (matrix sensing [2, 28], phase retrieval [38], matrix completion [14, 13, 7]) and neural network training (either with linear [45, 21, 42] or non-linear activation functions [29, 30]) has been a popular subject of study recently. These works have direct link to ours since matrix factorization *without any support constraint* can be seen either as a matrix decomposition problem or as a specific case of neural network (with two layers, no bias and linear activation function). Notably it has been proved [45] 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 −*strictly*– 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 [29, 42]. In particular, the combination of these results shows the benign landscape for LMRA, a particular instance of (FSMF).

However, to the best of our knowledge, existing analyses of landscape are only proposed for neural network training in general and matrix factorization problem in particular *without support constraints*, cf. [45, 42, 21], while the study of the landscape of (FSMF) remains untouched in the literature and our work can be considered as a generalization of such previous results. Moreover, unlike many existing results of matrix decomposition problems that are proved to hold with high probability under certain random models [2, 28, 38, 14, 13, 7, 8]), our result deterministically ensures the benign landscape for *each* matrix A, under certain conditions on the support constraints (I, J).

To summarize, our main contributions in this paper are:

1) We prove that (FSMF) is NP-hard in Theorem 2.4. In addition, in light of classical results on the **LU** decomposition, we highlight in Section 2 a challenge related to the possible non-existence of an optimal solution of (FSMF) .
2) We introduce families of support constraints (I, J) making (FSMF) tractable (Theorem 3.3 and Theorem 3.8) and provide dedicated polynomial algorithms for those families.
3) We show that the landscape of (FSMF) corresponding to the support pairs (I, J) in these families are free of spurious local valleys, regardless of the factorized matrix A (Theorem 4.12, Theorem 4.13). We also investigate the presence of spurious local minima for such families (Theorem 4.12, Theorem 4.19).
4) These results might suggest a conjecture that holds true for the full support case: an instance of (FSMF) is tractable if and only if its corresponding landscape is benign, i.e. free of spurious objects. We give a counter-example to this conjecture (Remark 4.23) and illustrate numerically that even with support constraints ensuring a benign landscape, state-of-the-art gradient descent methods can be significantly slower than the proposed dedicated algorithm.

1.1. Notations. For $n \in \mathbb{N}$, define $[\![n]\!] := \{1, \dots, n\}$. The notation $\mathbf{0}$ (resp. $\mathbf{1}$) stands for a matrix with all zeros (resp. all ones) coefficients. 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 restricted to the columns indexed in T while $A_T \in \mathbb{R}^{m\times n}$ is the matrix that has the same columns as A for indices in T and is zero elsewhere. 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). If $S \subseteq [\![m]\!], T \subseteq [\![n]\!]$, then $A_{S,T} \in \mathbb{R}^{|S|\times|T|}$ is the submatrix of A restricted to rows and columns indexed in S

and T respectively.

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 $(i,j) \in I$ and 0 otherwise. Both representations will be used interchangeably and the meaning should be clear from the context. For $T \subseteq [\![r]\!]$, we use the notation $I_T := I \cap ([\![m]\!] \times T)$ (this is consistent with the notation A_T introduced earlier).

The notation $\mathtt{supp}(A)$ is used for both vectors and matrices: 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 (FSMF) is NP-hard we use the classical technique to prove NP-hardness: reduction. Our choice of reducible problem is matrix completion with noise [15].

Definition 2.1 (Matrix completion with noise [15]). *Let $W \in \{0,1\}^{m\times n}$ be a binary matrix. Given $A \in \mathbb{R}^{m\times n}, s \in \mathbb{N}$, the matrix completion problem (MCP) is:*

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

This problem is NP-hard even when $s = 1$ [15] by its reducibility from Maximum-Edge Biclique Problem, which is NP-complete [33]. This is given in the following theorem:

Theorem 2.2 (NP-hardness of matrix completion with noise [15]). *Given a binary weighting matrix $W \in \{0,1\}^{m\times n}$ and $A \in [0,1]^{m\times n}$, the optimization problem*

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

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

The following lemma gives a reduction from (MCPO) to (FSMF).

Lemma 2.3. *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 $A \in \mathbb{R}^{m\times n}$, (MCPO) and (FSMF) share the same infimum. I and J can be constructed in polynomial time. Moreover, if one of the problems has a known solution that provides objective function accuracy ϵ, we can find a solution with the same accuracy for the other one in polynomial time.*

Proof sketch. 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:

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

This construction can clearly be made in polynomial time. We show in the supplementary material (Appendix A) that the two problems share the same infimum. ∎

Using Lemma 2.3, we obtain a result of NP-hardness for (FSMF) as follows.

THEOREM 2.4. *When $A \in [0,1]^{m\times n}$, it is NP-hard to solve* (FSMF) *with arbitrary index sets I, J and objective function accuracy less than $\epsilon = 2^{-12}(mn)^{-7}$.*

Proof. Given any instance of (MCPO) (i.e., two matrices $A \in [0,1]^{m\times n}$ and $W \in \{0,1\}^{m\times n}$), we can produce an instance of (FSMF) (the same matrix A and $I \in \{0,1\}^{m\times r}, J \in \{0,1\}^{n\times r}$) such that both have the same infimum (Lemma 2.3). Moreover, for any given objective function accuracy, we can use the procedure of Lemma 2.3 to make sure the solutions of both problems share the same accuracy.

Since all procedures are polynomial, this defines a polynomial reduction from (MCPO) to (FSMF). Because (MCPO) is NP-hard to obtain a solution with objective function accuracy less than ϵ (Theorem 2.2), so is (FSMF). ∎

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.

Lemma 2.3 constructs a hard instance where $(I, J) \in \{0,1\}^{m\times r} \times \{0,1\}^{n\times r}$ and $r = \min(m, n) + 1$. It is also interesting to investigate the hardness of (FSMF) given a fixed r. When $r = 1$, the problem is polynomially tractable since this case is covered by Theorem 3.3 below. On the other hand, when $r \geq 2$, the question becomes complicated due to the fact that the set $\{XY^\top \mid \mathtt{supp}(X) \subseteq I, \mathtt{supp}(Y) \subseteq J\}$ is not always closed. In Remark A.1, we show an instance of (FSMF) where the infimum is zero but cannot be attained. Interestingly enough, this is exactly the example for the non-existence of an exact **LU** decomposition of a matrix in $\mathbb{R}^{2\times 2}$ presented in [17, Chapter 3.2.12]. We emphasize that this is not a mere consequence of the non-coercivity of $L(X, Y)$ – which follows from rescaling invariance, see e.g. Remark 4.2 – as we will also present support constraints for which the problem always admits a global minimizer and can be solved with an efficient algorithm. More generally, one can even show that the set $\mathcal{L}$ of square matrices of size $n \times n$ having an exact **LU** decomposition (*i.e.*, $\mathcal{L} := \{XY^\top \mid \mathtt{supp}(X) \subseteq I, \mathtt{supp}(J) \subseteq J\}$ where $I = J = \{(i,j) \mid 1 \leq j \leq i \leq n\}$) is dense in $\mathbb{R}^{n\times n}$ (since a matrix having all non-zero leading principal minors admits an exact **LU** factorization [17, Theorem 3.2.1]) but $\mathcal{L} \subsetneq \mathbb{R}^{n\times n}$. Thus, $\mathcal{L}$ is not closed.

3. Tractable instances of matrix factorization with fixed support.

Even though (FSMF) is generally NP-hard, when we consider the full support case $I = [\![m]\!] \times [\![r]\!], J = [\![n]\!] \times [\![r]\!]$ the problem is equivalent to LRMA [12], which can be solved using the Singular Value Decomposition (SVD) [16][3]. This section is devoted to enlarge the family of supports for which (FSMF) can be solved by an effective direct algorithm based on blockwise SVDs. We start with an important definition:

DEFINITION 3.1 (Support of rank-one contribution). *Given two support constraints $I \in \{0,1\}^{m\times r}$ and $J \in \{0,1\}^{n\times r}$ of* (FSMF) *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: $\mathcal{S}_k(I,J) = I_{\bullet,k}J_{\bullet,k}^\top$. 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 $\mathtt{supp}(I_{\bullet,k}) \times \mathtt{supp}(J_{\bullet,k})$.*

Given a pair of support constraints I, J, if $\mathtt{supp}(X) \subseteq I, \mathtt{supp}(Y) \subseteq J$, we have: $\mathtt{supp}(X_{\bullet,k}Y_{\bullet,k}^\top) \subseteq \mathcal{S}_k,\ \forall k \in [\![r]\!]$. 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 $XY^\top$ (illustrated in Figure 5).

[3]SVD can be computed to machine precision in $O(mn^2)$ [22], see also [40, Lecture 31, page 236]. It is thus convenient to think of LRMA as polynomially solvable.

In the case of full supports ($\mathcal{S}_k = \mathbf{1}_{m\times n}$ for each $k \in [\![r]\!]$), the optimal solution can be obtained in a greedy manner: indeed, it is well known that Algorithm 3.1 computes factors achieving the best rank-r approximation to A (notice that here the algorithm also works for complex-valued matrices):

Algorithm 3.1 Generic Greedy Algorithm

Require: $A \in \mathbb{R}^{m\times n}$ or $\mathbb{C}^{m\times n}$; $\{\mathcal{S}_k\}_{k\in[\![r]\!]}$ rank-one supports
1: **for** $i \in [\![r]\!]$ **do**
2: $\quad (X_{\bullet,i}, Y_{\bullet,i}) = (u, v)$ where $uv^\top$ is any best rank-one approximation to $A \odot \mathcal{S}_i$
3: $\quad A = A - X_{\bullet,i}, Y_{\bullet,i}^\top$
4: **end for**
5: **return** (X, Y)

Even beyond the full support case, the output of Algorithm 3.1 always satisfies the support constraints due to line 2, however it may not always be the optimal solution of (FSMF). Our analysis of the polynomial tractability conducted below will allow us to show that, under appropriate assumptions on I, J, one can compute in polynomial time an optimal solution of (FSMF) using variants of Algorithm 3.1. The definition of these variants will involve a partition of $[\![r]\!]$ in terms of equivalence classes of rank-one supports:

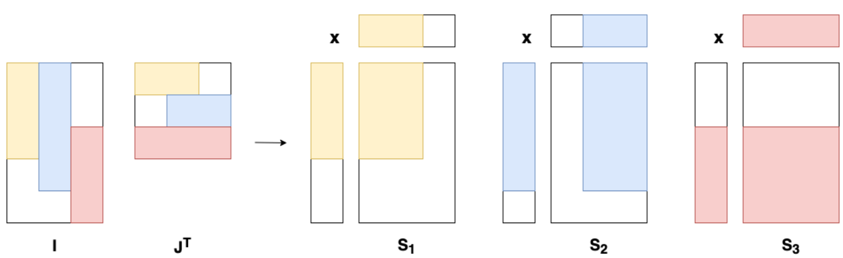

Fig. 5. *Illustration the idea of support of rank-one contribution. Colored rectangles indicate the support constraints (I, J) and the support constraints $\mathcal{S}_k$ on each component matrix $X_{\bullet,k}Y_{\bullet,k}^\top$.*

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 $\mathtt{supp}(I_{\bullet,k}) = R_P$, $\mathtt{supp}(J_{\bullet,k}) = C_P$.

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 $\bar{\mathcal{S}}_{\mathcal{P}'} = ([\![m]\!] \times [\![n]\!]) \backslash \mathcal{S}_{\mathcal{P}'}$.

For instance, in the example in Figure 5 we have three distinct equivalence classes. With the introduction of equivalence classes, one can modify Algorithm 3.1 to make it more efficient, as in Algorithm 3.2: Instead of computing the SVD r times, one can simply compute it only $|\mathcal{P}|$ times. For the full support case, we have $\mathcal{P} = \{[\![r]\!]\}$, thus Algorithm 3.2 is identical to the classical SVD.

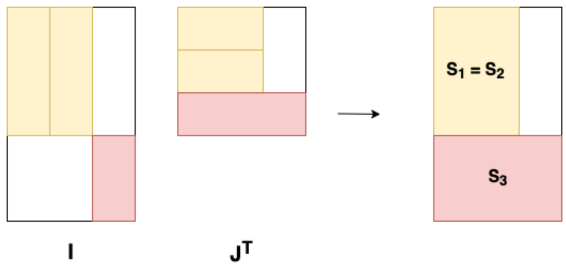

Fig. 6. *An instance of support constraints (I, J) satisfying Theorem 3.3. We use colored rectangles to indicate the support constraints (I, J). The indices belonging to the same equivalence class share the same color.*

Algorithm 3.2 Alternative Generic Greedy Algorithm

Require: $A \in \mathbb{R}^{m\times n}$ or $\mathbb{C}^{m\times n}$; $\{\mathcal{S}_P\}_{P\in\mathcal{P}}$ representative rank-one supports

```
for P ∈ 𝒫 do
    (X_{•,P}, Y_{•,P}) = (U, V) where UV^⊤ is any best rank-|P| approximation to A ⊙ 𝒮_P
    A = A − X_{•,P}, Y_{•,P}^⊤
end for
return (X, Y)
```

A first simple sufficient condition ensuring the tractability of an instance of (FSMF) is stated in the following theorem.

Theorem 3.3. *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 $A \in \mathbb{R}^{m\times n}$.*

Proof. 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 Subsection 1.1). We also use the notations R_P, C_P (Definition 3.2). For each equivalent class P, we have:

$$(X_P Y_P^\top)_{R_P, C_P} = X_{R_P,P} Y_{C_P,P}^\top \tag{3.1}$$

and the product $XY^\top$ can be decomposed as: $XY^\top = \sum_{P\in\mathcal{P}} X_P Y_P^\top$. Due to the hypothesis of this theorem, with $P, P' \in \mathcal{P}, P' \neq P$, we further have:

$$X_{P'} Y_{P'}^\top \odot \mathcal{S}_P = \mathbf{0} \tag{3.2}$$

Algorithm 3.3 Fixed support matrix factorization (under Theorem 3.3 assumptions)

```
procedure SVD_FSMF(A ∈ ℝ^{m×n}, I ∈ {0,1}^{m×r}, J ∈ {0,1}^{n×r})
    Partition ⟦r⟧ into 𝒫 (Definition 3.2) to get {𝒮_P}_{P∈𝒫}
    return (X, Y) using Algorithm 3.2 with input A, {𝒮_P}_{P∈𝒫}
end procedure
```

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

$$\begin{aligned}
\|A - XY^\top\|^2 &= \left(\sum_{P\in\mathcal{P}} \|(A - XY^\top)\odot \mathcal{S}_P\|^2\right) + \|(A - XY^\top)\odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2 \\
&= \left(\sum_{P\in\mathcal{P}} \|(A - \sum_{P'\in\mathcal{P}} X_{P'}Y_{P'}^\top)\odot \mathcal{S}_P\|^2\right) + \|(A - \sum_{P'\in\mathcal{P}} X_{P'}Y_{P'}^\top)\odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2 \\
&\overset{(3.2)}{=} \left(\sum_{P\in\mathcal{P}} \|(A - X_PY_P^\top)\odot \mathcal{S}_P\|^2\right) + \|A\odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2 \\
&= \left(\sum_{P\in\mathcal{P}} \|A_{R_P,C_P} - (X_PY_P^\top)_{R_P,C_P}\|^2\right) + \|A\odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2 \\
&\overset{(3.1)}{=} \left(\sum_{P\in\mathcal{P}} \|A_{R_P,C_P} - X_{R_P,P}Y_{C_P,P}^\top\|^2\right) + \|A\odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2
\end{aligned} \tag{3.3}$$

Therefore, if we ignore the constant term $\|A\odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2$, 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 LRMA instances. Since all the optimized parameters are $\{(X_{R_P,P}, Y_{C_P,P})\}_{P\in\mathcal{P}}$, an 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 a minimizer of $\|A_{R_P,C_P} - X_{R_P,P}Y_{C_P,P}^\top\|^2$ which is computed efficiently using a truncated SVD. Since the blocks associated to distinct P are disjoint, these SVDs can be performed blockwise, in any order, and even in parallel. ∎

For these easy instances, we can therefore recover the factors in polynomial time with the procedure described in Algorithm 3.3. Given a target matrix $A \in \mathbb{R}^{m\times n}$ and support constraints $I \in \{0,1\}^{m\times r}$, $J \in \{0,1\}^{n\times r}$ satisfying the condition in Theorem 3.3, Algorithm 3.3 returns two factors (X,Y) solution of (FSMF).

As simple as this condition is, it is satisfied in some important cases, for instance for a class of Hierarchical matrices (HODLR, cf. Appendix E), or for the so-called *butterfly supports*: in the latter case, the condition is used in [25, 44] to design an efficient hierarchical factorization method, which is shown to outperform first-order optimization approaches commonly used in this context, in terms both of computational time and accuracy.

In the next result, we explore the tractability of (FSMF) while allowing partial intersection between two representative rank-one contribution supports.

DEFINITION 3.4 (Complete equivalence classes of rank-one supports - CEC). *$P \in \mathcal{P}$ is a* complete equivalence class *(or* CEC*) if $|P| \geq \min\{|C_P|, |R_P|\}$ with C_P, R_P as in Definition 3.2. 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 3.5. *Given $I \in \{0,1\}^{m\times r}$, $J \in \{0,1\}^{n\times r}$, consider T, $\mathcal{S}_T$ as in Definition 3.4. 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 $\mathtt{supp}(X) \subseteq I_T$, $\mathtt{supp}(Y) \subseteq J_T$. Such a pair can be computed using Algorithm 3.3 $(X,Y) = SVD\_FSMF(A, I_T, J_T)$.*

The proof of Lemma 3.5 is deferred to the supplementary material (Appendix B.1). The next definition introduces the key properties that the indices $k \in [\![r]\!]$ which are not in any CEC need to satisfy in order to make (FSMF) overall tractable.

Definition 3.6 (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 3.4 and $\bar{T} = [\![r]\!] \setminus T$. 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 $R_k \subseteq [\![m]\!], C_k \subseteq [\![n]\!]$, (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 3.7 (Taxonomy of indices of I and J). *With the notations of Definition 3.6, 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	$I_T = \{(i,k) \mid k \in T, i \in [\![m]\!]\} \cap I$	$J_T = \{(j,k) \mid k \in T, j \in [\![n]\!]\} \cap J$
2	$I_{\bar{T}}^1 = \{(i,k) \mid k \notin T, i \in R_k\} \cap I$	$J_{\bar{T}}^1 = \{(j,k) \mid k \notin T, j \in C_k\} \cap J$
3	$I_{\bar{T}}^2 = \{(i,k) \mid k \notin T, i \notin R_k\} \cap I$	$J_{\bar{T}}^2 = \{(j,k) \mid k \notin T, j \notin C_k\} \cap J$

The following theorem generalizes Theorem 3.3.

Theorem 3.8. *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 $k, l \in \bar{T}$ we have $\mathcal{S}'_k = \mathcal{S}'_l$ or $\mathcal{S}'_k \cap \mathcal{S}'_l = \emptyset$. Then, $(I_{\bar{T}}^1, J_{\bar{T}}^1)$ satisfy the assumptions of Theorem 3.3. Moreover, for any matrix $A \in \mathbb{R}^{m\times n}$, two instances of (FSMF) with data (A, I, J) and $(A \odot \bar{\mathcal{S}}_T, I_{\bar{T}}^1, J_{\bar{T}}^1)$ respectively, share the same infimum. Given an optimal solution of one instance, we can construct the optimal solution of the other in polynomial time. In other word, (FSMF) with (A, I, J) is polynomially tractable.*

Theorem 3.8 is proved in the supplementary material (Appendix B.2). It implies that solving the problem with support constraints (I, J) can be achieved by reducing to another problem, with support constraints satisfying the assumptions of Theorem 3.3. The latter problem can thus be efficiently solved by Algorithm 3.3. In particular, Theorem 3.3 is a special case of Theorem 3.8 when all the equivalent classes (including CECs) have disjoint representative rank-one supports.

Figure 7 shows an instance of (I, J) satisfying the assumptions of Theorem 3.8. The extension in Theorem 3.8 is not directly motivated by concrete examples, but it is rather introduced as a first step to show that the family of polynomially tractable supports (I, J) can be enlarged, as it is not restricted to just the family introduced in Theorem 3.3. An algorithm for instances satisfying the assumptions of Theorem 3.8 is given in Algorithm 3.4 (more details can be found in Corollary B.3 and Remark B.4 in Appendix B in the supplementary material). In Algorithm 3.4, two calls to Algorithm 3.3 are made, they can be done in any order (Line 3 and Line 4 can be switched without changing the result).

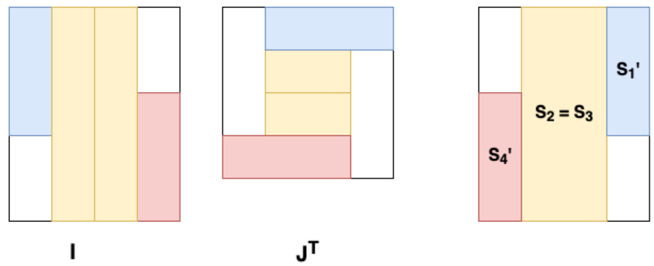

Fig. 7. *An instance of support constraints (I, J) satisfying the assumptions of Theorem 3.8. We have $T = \{2, 3\}$. The supports outside CEC $\mathcal{S}_1'$ and $\mathcal{S}_4'$ are disjoint.*

Algorithm 3.4 Fixed support matrix factorization (under Theorem 3.8's assumptions)

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 $I_T, I_{\bar{T}}^1, I_{\bar{T}}^2$ (and $J_T, J_{\bar{T}}^1, J_{\bar{T}}^2$) (Definition 3.6).
  $(X_T, Y_T) = \text{SVD\_FSMF}(A \odot \mathcal{S}_T, I_T, J_T)$ ($T, \mathcal{S}_T$ as in Definition 3.4).
  $(X_{\bar{T}}^1, Y_{\bar{T}}^1) = \text{SVD\_FSMF}(A \odot \bar{\mathcal{S}_T}, I_{\bar{T}}^1, J_{\bar{T}}^1)$
  return $(X_T + X_{\bar{T}}^1, Y_T + Y_{\bar{T}}^1)$
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, as they may prevent local optimization methods to converge to globally optimal solutions. Previous works [42, 45, 21] 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 much richer family of tractable support constraints for (FSMF) that we introduced in Theorem 3.3. The landscape with the assumptions of Theorem 3.8 is also analyzed. These results might suggest a natural conjecture: an instance of (FSMF) 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 (FSMF) that can be solved efficiently, despite the fact that its corresponding landscape contains spurious objects.

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 [45, 31]). *Consider $L : \mathbb{R}^d \to \mathbb{R}$. A vector $x^* \in \mathbb{R}^d$ is a:*
- ***global minimum*** *(of L) if $L(x^*) \leq L(x), \forall x$.*
- ***local minimum*** *if there is a neighborhood $\mathcal{N}$ of x^* such that $L(x^*) \leq L(x), \forall x \in \mathcal{N}$.*
- ***strict local minimum*** *if there is a neighborhood $\mathcal{N}$ of x^* such that $L(x^*) < L(x), \forall x \in \mathcal{N}, x \neq x^*$.*
- ***(strict) spurious local minimum*** *if x^* is a (strict) 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.

Remark 4.2. With the loss functions $L(X, Y)$ considered in this paper, strict local minima do not exist since for every invertible diagonal matrix D, possibly arbitrarily close to the identity, we have $L(XD, YD^{-1}) = L(X, Y)$.

However, this is not the only undesirable landscape in an optimization problem: spurious local valleys, as defined next, are also challenging.

DEFINITION 4.3 (Sublevel Set [4]). *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 4.4 (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 4.5 (Spurious Local Valley [42, 29]). *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. 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 8 illustrates some other scenarios: as shown on Figure 8a, the segment (approximately) $[10, +\infty)$ creates a

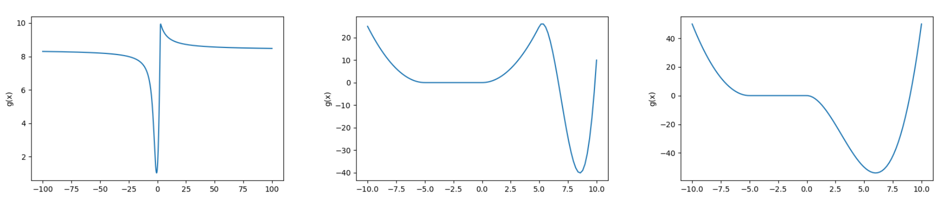

FIG. 8. *Examples of functions with spurious objects.*

spurious local valley, and this function has only one local (and global) minimizer, at zero; in Figure 8b, 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 8c). 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. Previous results on the landscape. Previous works [21, 45] studied the non-existence of spurious local minima of (FSMF) in the classical case of "low rank matrix approximation" (or *full support matrix factorization*)[4]. To prove that a critical point is never a spurious local minimum, previous work used the notion of *strict saddle point* (i.e a point where the Hessian is not positive semi-definite, or equivalently has at least one –*strictly* – negative eigenvalue), see Definition 4.10 below. To prove the non-existence of spurious local valleys, the following lemma was employed in previous works [42, 29]:

LEMMA 4.6 (Sufficient condition for the non-existence of any spurious local valley [42, 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 result is intuitive and a formal proof can be found in [42]. The theorem claims that 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. We remark that although (FSMF) is a constrained optimization problem, Lemma 4.6 is still applicable because one can think of the objective function as defined on a subspace: $L : \mathbb{R}^{|I|+|J|} \to \mathbb{R}$. In this work, to apply Lemma 4.6, the constructed function f has to be a *feasible path*, defined as:

DEFINITION 4.7 (Feasible path). *A feasible path w.r.t the support constraints (I, J) (or simply a feasible path) is a continuous function $f(t) = (X_f(t), Y_f(t)) : [0,1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$ satisfying $\mathtt{supp}(X_f(t)) \subseteq I, \mathtt{supp}(Y_f(t)) \subseteq J, \forall t \in [0,1]$.*

Conversely, we generalize and formalize an idea from [42] into the following lemma, which gives a sufficient condition for the existence of a spurious local valley:

LEMMA 4.8 (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) $\inf\limits_{x \in S_2} L(x) > \inf\limits_{x \in S_3} L(x) > \inf\limits_{x \in S_1} L(x)$.

Then L has a spurious local valley. Moreover, any $x \in S_3$ such that $L(x) < \inf\limits_{x \in S_2} L(x)$ is a point inside 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, and $\Sigma \subseteq S_1$ by the first assumption.

Since $\inf_{x \in S_2} L(x) > \inf_{x \in S_3} L(x)$, there exists $\tau \in S_3, L(\tau) < \inf_{x \in S_2} L(x)$. Consider Φ the path-connected component of the sublevel set $\{x \mid L(x) \leq L(\tau)\}$ that contains τ. Since Φ is a non-empty path-connected component of a level set, it is a local valley. It is thus sufficient to prove that $\Phi \cap \Sigma = \emptyset$ to obtain that it matches the very definition of a spurious local valley.

Indeed, by contradiction, let's assume that there exists $\tau' \in \Phi \cap \Sigma$. Since $\tau, \tau' \in \Phi$ and Φ is path-connected, by definition of path-connectedness there exists a continuous

[4]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., where no support constraints are imposed ($I = [\![m]\!] \times [\![r]\!], J = [\![n]\!] \times [\![r]\!]$)

function $f : [0, 1] \to \Phi$ such that $f(0) = \tau \in S_3, f(1) = \tau' \in \Sigma \subseteq S_1$. 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 (FSMF) and will be used in our subsequent proofs. The questions of the existence of spurious local valleys and spurious local minima were addressed in previous works for full support matrix factorization and deep linear neural networks [42, 29, 45, 21]. We present only results related to our problem of interest.

THEOREM 4.9 (No spurious local valleys in linear networks [42, Theorem 11]). *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 $L(\theta) = \|A - \Phi(B, \theta)\|^2$ satisfies the condition of Lemma 4.6 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).*

DEFINITION 4.10 (Strict saddle property [45, Definition 3]). *Consider a twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$. If each critical point of f is either a global minimum or a* strict *saddle point then f is said to have the strict saddle property. When this property holds, f has no spurious local minimum.*

Even if f has the strict saddle property, it may have no global minimum, consider e.g. the function $f(x) = -\|x\|_2^2$.

THEOREM 4.11 (No spurious local minima in shallow linear networks [45, Theorem 3]). *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 problem:*

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

If B is full row rank, f has the strict saddle property (see Definition 4.10) hence f has no spurious local minimum.

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 4.9, if $K = 2, B = \mathbf{I}_n$ ($n = N$), then the considered function is $L = \|A - W_2 W_1\|^2$. This is (FSMF) without support constraints I and J (and without a transpose on W_1, which does not change the nature of the problem). Theorem 4.9 guarantees that L satisfies the conditions of Lemma 4.6, thus has no spurious local valley.

Similarly, in Theorem 4.11, if $B = \mathbf{I}_{d_0}$ ($d_0 = N$, therefore B is full row rank), we return to the same situation of Theorem 4.9. In general, Theorem 4.11 claims that the landscape of the full support matrix factorization problem has the strict saddle property and thus, does not have spurious local minima.

However, once we turn to (FSMF) with *arbitrary* I and J, such benign landscape is not guaranteed anymore, as we will show in Remark 4.23. Our work in the next subsections studies conditions on the support constraints I and J ensuring the absence / allowing the presence of spurious objects, and can be considered as a generalization of previous results with full supports. [45, 42, 21].

4.3. Landscape of matrix factorization with fixed support constraints. We start with the first result on the landscape in the simple setting of Theorem 3.3.

THEOREM 4.12. *Under the assumption of Theorem 3.3, the function $L(X,Y)$ in (FSMF) does not admit any spurious local valley for any matrix A. In addition, L has the strict saddle property.*

Proof. Recall that under the assumption of Theorem 3.3, all the variables to be optimized are decoupled into "blocks" $\{(X_{R_P,P}, Y_{C_P,P})\}_{P\in\mathcal{P}}$ ($P, \mathcal{P}$ are defined in Definition 3.2). We denote $\mathcal{P} = \{P_1, P_2, \dots, P_\ell\}$, $P_i \subseteq [\![r]\!]$, $1 \le i \le \ell$. From Equation (3.3), we have:

$$\|A - XY^\top\|^2 = \left(\sum_{P\in\mathcal{P}} \|A_{R_P,C_P} - X_{R_P,P}Y_{C_P,P}^\top\|^2\right) + \|A \odot \bar{\mathcal{S}_\mathcal{P}}\|^2 \tag{4.1}$$

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 restricted to the corresponding blocks in A. The global minimizers of L are $\{(X^\star_{R_P,P}, Y^\star_{C_P,P})\}_{P\in\mathcal{P}}$, where for each $P \in \mathcal{P}$ the pair $(X^\star_{R_P,P}, Y^\star_{C_P,P})$ is any global minimizer of $\|A_{R_P,C_P} - X_{R_P,P}Y_{C_P,P}^\top\|^2$.

1) **Non-existence of any spurious local valley**: By Theorem 4.9, from any initial point $(X^0_{R_P,P}, Y^0_{C_P,P})$, there exists a continuous function $f_P(t) = (\tilde{X}_P(t), \tilde{Y}_P(t)) : [0,1] \mapsto \mathbb{R}^{|R_P|\times|P|} \times \mathbb{R}^{|C_P|\times|P|}$ satisfying the conditions in Lemma 4.6, 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 a feasible path (Definition 4.7) $f(t) = (\tilde{X}(t), \tilde{Y}(t)) : [0,1] \mapsto \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n}$ defined in such a way that $\tilde{X}(t)_{R_P,P} = \tilde{X}_P(t)$ for each $P \in \mathcal{P}$ and similarly for $\tilde{Y}(t)$. Since $L \circ f = \sum_{P\in\mathcal{P}} L_P \circ f_P + \|A \odot \bar{\mathcal{S}_\mathcal{P}}\|^2$, f satisfies the assumptions of Lemma 4.6, which shows the non-existence of any spurious local valley.
2) **Non-existence of any spurious local minimum**: Due to the decomposition in Equation (4.1), the gradient and Hessian of $L(X,Y)$ have the following form:

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

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

   Consider a critical point (X,Y) of $L(X,Y)$ that is not a global minimizer. Since (X,Y) is a critical point of $L(X,Y)$, $(X_{R_P,P}, Y_{C_P P})$ is a critical point of the function L_P for all $P \in \mathcal{P}$. Since (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 4.11, $H(L_P)_{|(X_{R_P,P}, Y_{C_P,P})}$ is not positive semi-definite. Hence, $H(L)_{|(X,Y)}$ is not positive semi-definite either (since $H(L)_{|(X,Y)}$ has block diagonal form). This implies that (X,Y) it is a strict saddle point as well (hence, not a spurious local minimum). ∎

For spurious local valleys, we have the same results for the setting in Theorem 3.8. The proof is, however, less straightforward.

THEOREM 4.13. *If I, J satisfy the assumptions of Theorem 3.8, then for each matrix A the landscape of $L(X,Y)$ in (FSMF) has no spurious local valley.*

The following is a concept which will be convenient for the proof of Theorem 4.13.

DEFINITION 4.14 (CEC-full-rank). *A feasible point (X,Y) is said to be* CEC-full-rank *if $\forall P \in \mathcal{P}^\star$, either $X_{R_P,P}$ or $Y_{C_P,P}$ is full row rank.*

We need three following lemmas to prove Theorem 4.13:

LEMMA 4.15. *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 feasible path $f : [0,1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} : f(t) = (X_f(t), Y_f(t))$ such that:*

1) *f connects (X,Y) with a CEC-full-rank point: $f(0) = (X,Y)$, and $f(1)$ is CEC-full-rank.*
2) *$X_f(t)(Y_f(t))^\top = XY^\top, \forall t \in [0,1]$.*

LEMMA 4.16. *Under the assumption of Theorem 3.8, for any CEC-full-rank feasible point (X,Y), there exists feasible path $f : [0,1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} : f(t) = (X_f(t), Y_f(t))$ such that:*

1) *$f(0) = (X,Y)$.*
2) *$L \circ f$ is non-increasing.*
3) *$(A - X_f(1)(Y_f(1))^\top) \odot \mathcal{S}_T = \mathbf{0}$.*

LEMMA 4.17. *Under the assumption of Theorem 3.8, for any CEC-full-rank feasible point (X,Y) satisfying: $(A - XY^\top) \odot \mathcal{S}_T = \mathbf{0}$, there exists a feasible path $f : [0,1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} : f(t) = (X_f(t), Y_f(t))$ such that:*

1) *$f(0) = (X,Y)$.*
2) *$L \circ f$ is non-increasing.*
3) *$f(1)$ is an optimal solution of L.*

The proofs of Lemma 4.15, Lemma 4.16 and Lemma 4.17 can be found in Appendix D.1, Appendix D.2 and Appendix D.3 of the supplementary material.

Proof of Theorem 4.13. Given any initial point (X^0, Y^0), Lemma 4.15 shows the existence of a continuous path along which the product of $XY^\top = X^0(Y^0)^\top$ does not change (thus, $L(X,Y)$ is constant) and ending at a CEC-full-rank point. Therefore it is sufficient to prove the theorem under the additional assumption that (X^0, Y^0) is CEC-full-rank. With this additional assumption, one can employ Lemma 4.16 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:

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

Again, one can assume that (X^1, Y^1) is CEC-full-rank (one can invoke Lemma 4.15 one more time). Therefore, (X^1, Y^1) satisfies the conditions of Lemma 4.17 . 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 4.6, we can conclude that there is no spurious local valley in the landscape of $\|A - XY^\top\|^2$. ◻

The next natural question is whether spurious local minima exist in the setting of Theorem 3.8. While in the setting of Theorem 3.3, all critical points which are not global minima are saddle points, the setting of Theorem 3.8 allows second order critical points (point whose gradient is zero and Hessian is positive semi-definite), which are not global minima.

Example 4.18. Consider the following pair of support contraints I, J and factorized matrix $I = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $J = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $A = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}$. With the notations of Definition 3.4 we have $T = \{1\}$ and one can check that this choice of I and J satisfies the assumptions of Theorem 3.8. 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 feasible point (X_0, Y_0): $X_0 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $Y_0 = \begin{bmatrix} 0 & 10 \\ 0 & 0 \end{bmatrix}$. Since $X_0 Y_0^\top = \begin{bmatrix} 10 & 0 \\ 0 & 0 \end{bmatrix} \neq A$, (X_0, Y_0) is not a global optimal solution. Calculating the gradient of L verifes that (X_0, Y_0) is a critical point:

$$\nabla L(X_0, Y_0) = ((A - X_0 Y_0^\top) Y_0, (A^\top - Y_0 X_0^\top) X_0) = (\mathbf{0}, \mathbf{0})$$

Nevertheless, the Hessian of the function L at (X_0, Y_0) is positive semi-definite. Direct calculation can be found in Appendix D.5 of the supplementary material.

This example shows that if we want to prove the non-existence of spurious local minima in the new setting, one cannot rely on the Hessian. This is challenging since the second order derivatives computation is already tedious. Nevertheless, with Definition 4.14, we can still say something about spurious local minima in the new setting.

THEOREM 4.19. *Under the assumptions of Theorem 3.8, if a feasible point (X, Y) is CEC-full-rank, then (X, Y) is not a spurious local minimum of* (FSMF). *Otherwise there is a feasible path, along which $L(\cdot, \cdot)$ is constant, that joins (X, Y) to some $(\tilde{X}, \tilde{Y})$ which is not a spurious local minimum.*

When (X, Y) is *not* CEC-full-rank, the theorem guarantees that it is not a strict local minimum, since there is path starting from (X, Y) with constant loss. This should however not be a surprise in light of Remark 4.2: indeed, the considered loss function admits no strict local minimum at all. Yet, the path with "flat" loss constructed in the theorem is fundamentally different from the ones naturally due to scale invariances of the problem and captured by Remark 4.2. Further work would be needed to investigate whether this can be used to get a stronger result.

Proof sketch. To prove this theorem, we proceed through two main steps:

1) First, we show that any local minimum satisfies:

$$(A - XY^\top) \odot \mathcal{S}_T = \mathbf{0} \tag{4.2}$$

2) Second, we show that if a point (X, Y) is CEC-full-rank and satisfies Equation (4.2), it cannot be a spurious local minimum.

Combining the above to steps, we obtain as claimed that if a feasible pair (X, Y) is CEC-full-rank, then it is not a spurious local minimum. Finally, if a feasible pair (X, Y) is not CEC-full-rank, Lemma 4.15 yields a feasible path along which L is constant that joins (X, Y) to some feasible $(\tilde{X}, \tilde{Y})$ which is CEC-full-rank, hence (as we have just shown) not a spurious local mimimum.

A complete proof is presented in Appendix D.4 of the supplementary material. □

Although Theorem 4.19 does not exclude completely the existence of spurious local minima, together with Theorem 4.12, we eliminate a large number of such points.

4.4. Absence of correlation between tractability and benign landscape. So far, we have witnessed that the instances of (FSMF) satisfying the assumptions of Theorem 3.8 are not only efficiently solvable using Algorithm 3.4: they also have a landscape with no spurious local valleys and favorable in terms of spurious local minima Theorem 4.19. The question of interest is: Is there a link between such benign landscape and the tractability of the problem? Even if the natural answer could intuitively seem to be positive, as it is the case for the full support case, we prove that this conjecture is not true. We provide a counter example showing that tractability does not imply a benign landscape. First, we establish a sufficient condition for the *existence* of a spurious local valley in (FSMF).

THEOREM 4.20. *Consider function $L(X,Y) = \|A - XY^\top\|^2$ in (FSMF). Given two support constraints $I \in \{0,1\}^{m\times r}$, $J \in \{0,1\}^{n\times r}$, if there exist $i_1 \neq i_2 \in [\![m]\!], j_1 \neq j_2 \in [\![n]\!]$ and $k \in [\![r]\!]$ such that (i_2, j_2) belongs to at least 2 rank-one supports, one of which is $\mathcal{S}_k$, and if $(i_1,j_1),(i_2,j_1),(i_1,j_2)$ belong only to $\mathcal{S}_k$, then:*

1) There exists A such that: $L(X,Y)$ has a spurious local valley.

2) There exists A such that: $L(X,Y)$ has a spurious local minimum.

In both cases, A can be chosen so that the global minimum of $L(X,Y)$ under the considered support constraints is achieved and is zero.

Remark 4.21. Note that the conditions of Theorem 4.20 exclude these of Theorem 3.3 and Theorem 3.8 (which is reasonable since the assumptions of Theorem 3.3 and Theorem 3.8 rule out the possibility of spurious local valleys for any matrix A.).

Proof. Let $l \neq k$ be another rank-one contribution support $\mathcal{S}_l$ that contains (i_1, j_1). Without loss of generality, we can assume $i_1 = j_1 = 1, i_2 = j_2 = 2$ and $k = 1, l = 2$. In particular, let $I' = J' := \{(1,1),(2,1),(2,2)\}$, then $I' \subseteq I, J' \subseteq J$. When $m = n = 2$, these are the support constraints for the **LU** decomposition.

1) We define the matrix A by block matrices as:

$$A = \begin{pmatrix} A' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \text{ where } A' = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \in \mathbb{R}^{2\times 2}. \tag{4.3}$$

The minimum of $L(X,Y) := \|A - XY^\top\|^2$ over feasible pairs is zero and it is attained at $X = \left[\begin{smallmatrix} X' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right], Y = \left[\begin{smallmatrix} Y' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right]$ where $X' = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right], Y' = \left[\begin{smallmatrix} 1 & 0 \\ 1 & -1 \end{smallmatrix}\right]$. (X,Y) is feasible since $\mathtt{supp}(X) = \mathtt{supp}(X') = I' \subseteq I, \mathtt{supp}(Y) = \mathtt{supp}(Y') = J' \subseteq J$. Moreover,

$$XY^\top = \begin{pmatrix} X'Y'^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} A' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = A. \tag{4.4}$$

Using Lemma 4.8 we now prove that this matrix A produces a spurious local valley for $L(X,Y)$ with the considered support constraints (I,J). In fact, since $(1,1),(1,2),(2,1)$ are only in $\mathcal{S}_1$ and in no other support $\mathcal{S}_\ell$, $\ell \neq 1$, one can easily check that for every feasible pair (X,Y) we have:

$$(XY^\top)_{i,j} = X_{i,1}Y_{j,1}, \quad \forall (i,j) \in \{(1,1),(1,2),(2,1)\}. \tag{4.5}$$

Thus, every feasible pair $(X^\star, Y^\star)$ reaching the global optimum $\|A - X^\star(Y^\star)^\top\| = 0$ must satisfy $X^\star_{1,1}Y^\star_{1,1} = X^\star_{2,1}Y^\star_{1,1} = X^\star_{1,1}Y^\star_{2,1} = 1$. This implies $X^\star_{2,1}Y^\star_{2,1} = (X^\star_{2,1}Y^\star_{1,1})(X^\star_{1,1}Y^\star_{2,1})/(X^\star_{1,1}Y^\star_{1,1}) = 1$. Moreover, such an optimum feasible pair also satisfies $0 = A_{2,2} = (X^\star(Y^\star)^\top)_{2,2} = \sum_p X^\star_{2,p}Y^\star_{2,p}$, hence $\sum_{p\neq 1} X^\star_{2,p}Y^\star_{2,p} = -X^\star_{2,1}Y^\star_{2,1} = -1$.

To show the existence of a spurious local valley we use Lemma 4.8 and consider the set $\tilde{S}_\sigma = \{(X,Y) \mid \mathtt{supp}(X) \subseteq I, \mathtt{supp}(Y) \subseteq J, \sum_{p\neq 1} X_{2,p}Y_{2,p} = \sigma\}$. We will show that $S_1 := \tilde{S}_{-1}, S_2 := \tilde{S}_1, S_3 := \tilde{S}_5$ satisfy the assumptions of Lemma 4.8.

To compute $\inf_{(X,Y)\in S_i} L(X,Y)$, we study $g(\sigma) := \inf_{(X,Y)\in\tilde{S}_\sigma} L(X,Y)$. Denoting

$Z = \begin{bmatrix} \mathbf{1}_{2\times 2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \{0,1\}^{m\times n}$ we have:

$$\begin{aligned}
g(\sigma) &= \inf_{(X,Y)\in \tilde{S}_\sigma} \|A - XY^\top\|^2 \\
&\geq \inf_{(X,Y)\in \tilde{S}_\sigma} \|(A - XY^\top)\odot Z\|^2 \\
&\overset{(4.5)}{=} \inf_{(X,Y)\in \tilde{S}_\sigma} \left\| \begin{pmatrix} A_{1,1} - X_{1,1}Y_{1,1} & A_{1,2} - X_{11}Y_{21} \\ A_{2,1} - X_{2,1}Y_{1,1} & A_{2,2} - \sigma - X_{2,1}Y_{2,1} \end{pmatrix} \right\|^2 \\
&= \inf_{X_{1,1},X_{2,1},Y_{1,1},Y_{2,1}} \left\| \begin{pmatrix} 1 - X_{1,1}Y_{1,1} & 1 - X_{11}Y_{21} \\ 1 - X_{2,1}Y_{1,1} & -\sigma - X_{2,1}Y_{2,1} \end{pmatrix} \right\|^2
\end{aligned}$$

Besides Equation (4.5), the third equality exploits the fact that $(XY^\top)_{2,2} = \sum_p X_{2,p}Y_{2,p} = X_{2,1}Y_{2,1} + \sigma$. The last quantity is the loss of the best rank-one approximation of $\tilde{A} = \begin{bmatrix} 1 & 1 \\ 1 & -\sigma \end{bmatrix} \in \mathbb{R}^{2\times 2}$. Since this is a 2×2 symmetric matrix, its eigenvalues can be computed as the solutions of a second degree polynomial, leading to an analytic expression of this last quantity as: $\frac{2(\sigma+1)^2}{(\sigma^2+3)+\sqrt{(\sigma^2+3)^2-4(\sigma+1)^2}}$. Moreover, this infimum can be attained if $\begin{bmatrix} X_{1,1}, X_{2,1} \end{bmatrix} = \begin{bmatrix} Y_{1,1}, Y_{2,1} \end{bmatrix}$ is the first eigenvector of $\tilde{A}$ and the other coefficients of X, Y are set to zero. Therefore,

$$g(\sigma) = \frac{2(\sigma+1)^2}{(\sigma^2+3)+\sqrt{(\sigma^2+3)^2-4(\sigma+1)^2}}. \tag{4.6}$$

We can now verify that S_1, S_2, S_3 satisfy all the conditions of Lemma 4.8.

1) The minimum value of L is zero. As shown above, it is only attained with $\sum_{p\neq 1} X^\star_{2,p}Y^\star_{2,p} = -1$ as shown. Thus, the global minima belong to $S_1 = \tilde{S}_{-1}$.
2) For any feasible path $r : [0,1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} : t \to (X(t), Y(t))$ we have $\sigma_r(t) = \sum_{p\neq 1} X(t)_{2,p}Y(t)_{2,p}$ is also continuous. If $(X(0), Y(0)) \in S_3 = \tilde{S}_5$ and $(X(1), Y(1)) \in S_1 = \tilde{S}_{-1}$ then $\sigma_r(0) = 5$ and $\sigma_r(1) = -1$), hence by the Mean Value Theorem, there must exist $t \in (0,1)$ such that $\sigma_r(t) = 1$, which means $(X(t), Y(t)) \in S_2 = \tilde{S}_1$.
3) Since one can check numerically that $g(1) > g(5) > g(-1)$, we have

$$\inf_{(X,Y)\in S_2} L(X,Y) > \inf_{(X,Y)\in S_3} L(X,Y) > \inf_{(X,Y)\in S_1} L(X,Y).$$

The proof is concluded with the application of Lemma 4.8. In addition, any point (X,Y) satisfying $\sigma = 5$ and $L(X,Y) < g(1) = 2$ is inside a spurious local valley. For example, one of such a point is $X = \begin{bmatrix} X' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, Y = \begin{bmatrix} Y' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ where $X' = \begin{bmatrix} 1 & 0 \\ -5 & 1 \end{bmatrix}, Y' = \begin{bmatrix} -1/5 & 0 \\ 1 & 5 \end{bmatrix}$.

2) We define the matrix A by block matrices as:

$$A = \begin{pmatrix} A' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \text{ where } A' = \begin{pmatrix} b & 0 \\ 0 & a \end{pmatrix} \in \mathbb{R}^{2\times 2}. \tag{4.7}$$

where $a > b > 0$. It is again evident that The infimum of $\|A - XY^\top\|^2$ under the considered support constraints is zero, and is achieved (taking $X = \begin{bmatrix} X' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, Y = \begin{bmatrix} Y' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ where $X' = \begin{bmatrix} b & 0 \\ 0 & a \end{bmatrix}, Y' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and with the same proof as in Equation (4.4), we have $XY^\top = A$.)
Now, we will consider $\tilde{X} = \begin{bmatrix} X' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \tilde{Y} = \begin{bmatrix} Y' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ where $X' = \begin{bmatrix} 0 & 0 \\ a & 0 \end{bmatrix}, Y' = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$. Since $L(\tilde{X}, \tilde{Y}) = b^2 > 0$ it cannot be a global minimum. We will show that $(\tilde{X}, \tilde{Y})$

is indeed a local minimum, which will thus imply that $(\tilde{X}, \tilde{Y})$ is a spurious local minimum. For each feasible pair (X, Y) we have:

$$\begin{aligned}
\|A - XY^\top\|^2 &= \sum_{i,j}(A_{i,j} - (XY^\top)_{i,j})^2 \\
&\geq (A_{1,1} - (XY^\top)_{1,1})^2 + (A_{2,1} - (XY^\top)_{2,1})^2 + (A_{1,2} - (XY^\top)_{1,2})^2 \\
&\overset{(4.5)}{=} (b - X_{1,1}Y_{1,1})^2 + (X_{2,1}Y_{1,1})^2 + (X_{1,1}Y_{2,1})^2 \\
&\geq (X_{1,1}Y_{1,1})^2 - 2bX_{1,1}Y_{1,1} + b^2 + 2(X_{2,1}Y_{2,1})|X_{1,1}Y_{1,1}| \\
&\geq 2(X_{2,1}Y_{2,1} - b)|X_{1,1}Y_{1,1}| + b^2.
\end{aligned}$$

where in the third line we used that for $u = |X_{2,1}|Y_{11}$, $v = X_{11}|Y_{2,1}|$, since $(u-v)^2 \geq 0$ we have $u^2 + v^2 \geq 2uv$. Since $\tilde{X}_{2,1}\tilde{Y}_{2,1} = a > b$, there exists a neighborhood of $(\tilde{X}, \tilde{Y})$ such that $X_{2,1}Y_{2,1} - b > 0$ for all (X, Y) in that neighbourhood. Since $|X_{1,1}Y_{1,1}| \geq 0$ in this neighborhood it follows that $\|A - XY^\top\|^2 \geq b^2 = L(\tilde{X}, \tilde{Y}) > 0$ in that neighborhood. This concludes the proof. ∎

Remark 4.22. Theorem 4.20 is constructed based on the **LU** structure. We elaborate our intuition on the technical proof of Theorem 4.20 as follows: Consider the **LU** decomposition problem of size 2×2 (i.e., $I = J = \{(1,1), (2,1), (2,2)\}$). It is obvious that such (I, J) satisfies the assumptions of Theorem 4.20 (for $i_1 = j_1 = 1, i_2 = j_2 = 2$). We consider three matrices of size 2×2:

$$A_1 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix},\ A_2 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix},\ A_3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

A_1 (resp. A_2) is simply the matrix A' in (4.3) (resp. in (4.7), with $a = 2, b = 1$) in the proof of Theorem 4.20. A_3 is a matrix which does not admit an **LU** decomposition. We plot the graphs of $g_i(\sigma) = \inf_{X_{2,2}Y_{2,2}=\sigma} \|A_i - XY^\top\|$ (this is exactly $g(\sigma)$ introduced in the proof of Theorem 4.20) in Figure 9.

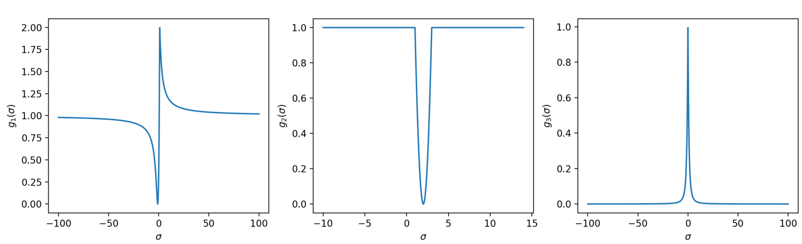

FIG. 9. *Illustration of the functions $g_i(\sigma), i = 1, 2, 3$ from left to right.*

In particular, the spurious local valley constructed in the proof of Theorem 4.20 with A_1 is a spurious local valley extending to infinity. With A_2, one can see that $g_2(\sigma)$ has a plateau with value $1 = b^2$. The local minimum that we consider in the proof of Theorem 4.20 is simply a point in this plateau (where $\sigma = 0$). Lastly, since the matrix A_3 does not admit an **LU** decomposition, there is no optimal solution. Nevertheless, the infimum zero can be approximated with arbitrary precision when σ tends to infinity (two valleys extending to $\pm\infty$).

For the cases with the matrices A_1 and A_3, once initialized inside the valleys of their landscapes, any sequence (X_k, Y_k) with sufficiently small steps associated to a decreasing loss $L(X_k, Y_k)$ will have the corresponding parameter σ converging to infinity. As a consequence, at least one parameter of either X_k or Y_k has to diverge. This is thus a setting in which PALM (and other optimization algorithms which seek to locally decrease their objective function in a monotone way) can diverge.

We can now exhibit the announced counter-example to the mentioned conjecture:

Remark 4.23. Consider the **LU** decomposition as an instance of (FSMF) with $m = n = r$, $I = J = \{(i, j) \mid 1 \leq j \leq i \leq n\}$, taking $i_1 = j_1 = 1, i_2 = j_2 = 2$ shows that the **LU** decomposition satisfies the condition of Theorem 4.20. Consequently, there exists a matrix A such that the global optimum of $L(X, Y)$ is achieved (and is zero), yet the landscape of $L(X, Y)$ will have spurious objects. Nevertheless, a polynomial algorithm to compute the **LU** decomposition exists [32]. This example is in the same spirit of a recent result presented in [43], where a polynomially solvable instance of Matrix Completion is constructed, whose landscape can have an exponential number of spurious local minima.

The existence of spurious local valleys shown in Theorem 4.20 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 initially. The answer is positive, as shown in the following theorem.

THEOREM 4.24. *Given any I, J, A such that the infimum of* (FSMF) *is attained, every 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. In particular, $(\mathbf{0}, \mathbf{0})$ is never in any spurious local valley.*

Proof. Let (X^*, Y^*) be a minimizer of (FSMF), which exists due to our assumptions. We only prove the result for the initialization $(X, \mathbf{0}), \mathtt{supp}(X) \subseteq I$. The case of the initialization $(\mathbf{0}, Y)$, $\mathtt{supp}(Y) \subseteq J$ can be dealt with similarly.

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

1) $f(0) = (X, \mathbf{0})$.
2) $f(1) = (X^*, Y^*)$.
3) $L \circ f$ is non-increasing w.r.t t.

Indeed, if such 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 $L \circ f$ is non-increasing).

We will construct such a function feasible path f as a concatenation of two functions feasible paths $f_1 : [0, 1/2] \to \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$, $f_2 : [1/2, 1] \to \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$, 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 $f_1(1/2) = f_2(1/2) = (X^*, \mathbf{0})$. Also, $L \circ f$ is non-increasing on $[0, 1]$ since:

1) $L(f_1(t)) = \|A - ((1 - 2t)X + 2tX^*)\mathbf{0}^\top\|^2 = \|A\|^2$ is constant for $t \in [0, 1/2]$.
2) $L(f_2(t)) = \|A - (2t - 1)X^*Y^*\|^2$ 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 (FSMF)). As a result, $t \mapsto L(f_2(t))$ is non-increasing on $[1/2, 1]$. ∎

Yet, such an initialization does not guarantee that first-order methods converge to a global minimum. Indeed, while in the proof of this result we do show that there exists a feasible path joining this "smart" initialization to an optimal solution without increasing the loss function, the value of the objective function is "flat" in the first part of this feasible path. Thus, even if such initialization is completely outside any spurious local valley, it is not clear whether local information at the initialization allows to "guide" optimization algorithms towards the global optimum to blindly find such a path. In fact, first-order methods are not bound to follow our constructive continuous path.

5. Numerical illustration: landscape and behaviour of gradient descent. As a numerical illustration of the practical impact of our results, we compare the performance of Algorithm 3.4 to other popular first-order methods on problem (FSMF).

We consider two types of instances of (FSMF): $I_1 = \mathbf{1}_{2^a \times 2^a} \otimes \mathbf{I}_{2^b \times 2^b}, J_1 = \mathbf{I}_{2^a \times 2^a} \otimes \mathbf{1}_{2^b \times 2^b}$ where $\otimes$ denotes the Kronecker product, $a = \lceil N/2 \rceil, b = \lfloor N/2 \rfloor$ (hence $a + b = N$) and $I_2 = \mathbf{1}_{2\times 2} \otimes \mathbf{I}_{2^{N-1}}, J_2 = \mathbf{I}_2 \otimes \mathbf{1}_{2^{N-1}\times 2^{N-1}}$. These supports are interesting because they are those taken at the first two steps of the hierarchical algorithm in [25, 44] for approximating a matrix by a product of N butterfly factors [25]. The first pair of support constraints (I_1, J_1) is also equivalent to the recently proposed Monarch parameterization [9]. Both pairs (I_1, J_1) and (I_2, J_2) are proved to satisfy Theorem 3.3 [44, Lemma 3.15].

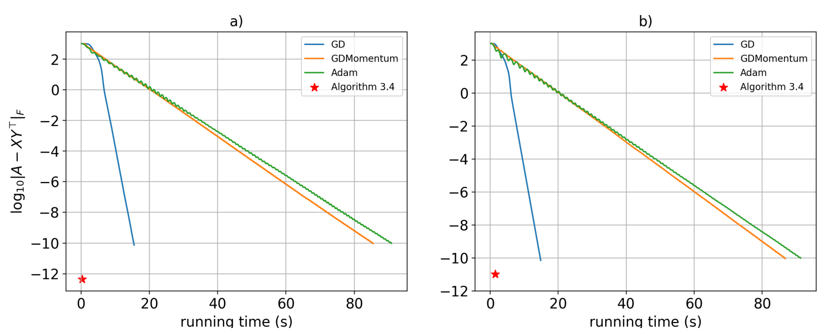

FIG. 10. *Evolution of* $\log_{10} \|A - XY^\top\|_F$ *for three variants of gradient descent and* Algorithm 3.4 *with support constraints* (I_1, J_1) *(left) and* (I_2, J_2) *(right) for* $N = 10$.

We consider A as the Hadamard matrix of size $2^N \times 2^N$, which is known to admit an exact factorization with each of the considered support constraints, and we employ Algorithm 3.4 to factorize A in these two settings. We compare Algorithm 3.4 to three variants of gradient descent: vanilla gradient descent (GD), gradient descent with momentum (GDMomentum) and ADAM [18, Chapter 8]. We use the efficient implementation of these iterative algorithms available in Pytorch 1.11. For each matrix size 2^N, learning rates for iterative methods are tuned by grid search: we run all the factorizations with all learning rates in $\{5 \times 10^{-k}, 10^{-k} \mid k = 1, \dots, 4\}$. Matrix X (resp. Y) is initialized with i.i.d. random coefficients inside its support I (resp. J) drawn according to the law $\mathcal{N}(0, 1/R_I)$ (resp. $\mathcal{N}(0, 1/R_J)$) where R_I, R_J are respectively the number of elements in each column of I and of J. All these experiments are run on an Intel Core i7 CPU 2,3 GHz. In the interest of reproducible research, our implementation is available in open source [24]. Since A admits an exact factorization with both the supports (I_1, J_1) and (I_2, J_2), we set a threshold $\epsilon = 10^{-10}$ for these iterative algorithms (i.e if $\log_{10}(\|A - XY^\top\|_F) \leq -10$, the algorithm is terminated and considered to have found an optimal solution). This determines the running time for a given iterative algorithm for a given dimension 2^N and a given learning rate. For each dimension 2^N we report the best running time over all learning rates. The reported running times do not include the time required for hyperparameters tuning. The

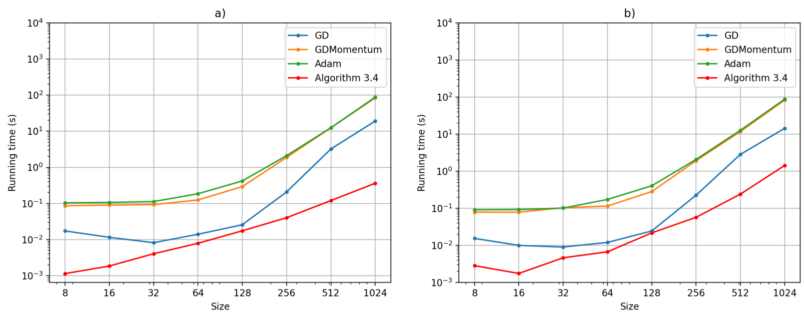

FIG. 11. *Running time (in logarithmic scale, contrary to Figure 10) of three variants of gradient descent and Algorithm 3.4 to reach a precision* $\log_{10}(\|A - XY^\top\|_F) \leq -10$*;* $N \in \{3, \ldots, 10\}$ *with support constraints* (I_1, J_1) *(left) and* (I_2, J_2) *(right).*

experiments illustrated in Figure 10 for $N = 10$ confirm our results on the landscape presented in Subsection 4.3: the assumptions of theorem Theorem 3.3 are satisfied so the landscape is benign and all variants of gradient descent are able to find a good factorization for A from a random initialization.

Figure 10 also shows that Algorithm 3.4 is consistently better than the considered iterative methods in terms of running time, regardless of the size of A, cf. Figure 11. A crucial advantage of Algorithm 3.4 over gradient methods is also that it is free of hyperparameter tuning, which is critical for iterative methods to perform well, and may be quite time consuming (we recall that the time required for hyperparameters tuning of these iterative methods is *not* considered in Figure 11). In addition, Algorithm 3.4 can be further accelerated since its main steps (cf Algorithm 3.2) rely on block SVDs that can be computed in parallel (in these experiments, our implementation of Algorithm 3.4 is not parallelized yet). Interested readers can find more applications of Algorithm 3.4 on the problem of fixed-support multilayer sparse factorization in [25].

6. Conclusion.

In this paper, we studied the problem of two-layer matrix factorization with fixed support. We showed that this problem is 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 (FSMF) with these support constraints. Although it would have seemed natural to assume an equivalence between tractability and benign landscape of (FSMF), 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 the proposed algorithm over state-of-the-art first order optimization methods usually employed in this context. We refer the reader to [25] where we propose an extension of Algorithm 3.3 to fixed-support multilayer sparse factorization and show the superiority of the resulting method in terms of both accuracy and speed compared to the state of the art [10].

REFERENCES

[1] J. Ballani and D. Kressner, *Matrices with Hierarchical Low-Rank Structures*, vol. 2173, 01 2016, https://doi.org/10.1007/978-3-319-49887-4_3.

[2] S. Bhojanapalli, B. Neyshabur, and N. Srebro, *Global optimality of local search for low rank matrix recovery*, in Advances in Neural Information Processing Systems, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds., vol. 29, Curran Associates, Inc., 2016, https://proceedings.neurips.cc/paper/2016/file/b139e104214a08ae3f2ebcce149cdf6e-Paper.pdf.

[3] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming, 146 (2014), pp. 459–494, https://doi.org/10.1007/s10107-013-0701-9, https://hal.inria.fr/hal-00916090.

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

[5] E. Candès, L. Demanet, and L. Ying, *A fast butterfly algorithm for the computation of fourier integral operators*, Multiscale Modeling & Simulation, 7 (2009), pp. 1727–1750, https://doi.org/10.1137/080734339, https://doi.org/10.1137/080734339, https://arxiv.org/abs/https://doi.org/10.1137/080734339.

[6] B. Chen, T. Dao, K. Liang, J. Yang, Z. Song, A. Rudra, and C. Ré, *Pixelated butterfly: Simple and efficient sparse training for neural network models*, in International Conference on Learning Representations, 2022, https://openreview.net/forum?id=Nfl-iXa-y7R.

[7] J. Chen and X. Li, *Model-free nonconvex matrix completion: Local minima analysis and applications in memory-efficient kernel PCA*, Journal of Machine Learning Research, 20 (2019), pp. 1–39, http://jmlr.org/papers/v20/17-776.html.

[8] Y. Chi, Y. M. Lu, and Y. Chen, *Nonconvex optimization meets low-rank matrix factorization: An overview*, Trans. Sig. Proc., 67 (2019), p. 5239–5269, https://doi.org/10.1109/TSP.2019.2937282, https://doi.org/10.1109/TSP.2019.2937282.

[9] T. Dao, B. Chen, N. Sohoni, A. Desai, M. Poli, J. Grogan, A. Liu, A. Rao, A. Rudra, and C. Ré, *Monarch: Expressive structured matrices for efficient and accurate training*, 2022, https://doi.org/10.48550/ARXIV.2204.00595, https://arxiv.org/abs/2204.00595.

[10] T. Dao, A. Gu, M. Eichhorn, A. Rudra, and C. Ré, *Learning fast algorithms for linear transforms using butterfly factorizations*, in Proceedings of the 36th International Conference on Machine Learning, vol. 97, 2019, pp. 1517–1527, http://proceedings.mlr.press/v97/dao19a.html.

[11] T. Dao, N. Sohoni, A. Gu, M. Eichhorn, A. Blonder, M. Leszczynski, A. Rudra, and C. Ré, *Kaleidoscope: An efficient, learnable representation for all structured linear maps*, in International Conference on Learning Representations, 2020, https://openreview.net/forum?id=BkgrBgSYDS.

[12] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218, https://doi.org/10.1007/BF02288367.

[13] R. Ge, C. Jin, and Y. Zheng, *No spurious local minima in nonconvex low rank problems: A unified geometric analysis*, in Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, JMLR.org, 2017, p. 1233–1242.

[14] R. Ge, J. D. Lee, and T. Ma, *Matrix completion has no spurious local minimum*, in Advances in Neural Information Processing Systems, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds., vol. 29, Curran Associates, Inc., 2016, https://proceedings.neurips.cc/paper/2016/file/7fb8ceb3bd59c7956b1df66729296a4c-Paper.pdf.

[15] 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), https://doi.org/10.1137/110820361.

[16] 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 (1965), pp. 205–224, http://www.jstor.org/stable/2949777.

[17] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, third ed., 1996.

[18] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

[19] W. Hackbusch, *A sparse matrix arithmetic based on h-matrices. part i: Introduction to h-matrices*, Computing, 62 (1999), pp. 89–108.

[20] W. Hackbusch and B. N. Khoromskij, *A sparse h-matrix arithmetic. part ii: Application to multi-dimensional problems*, Computing, 64 (2000), pp. 21–47.

[21] K. Kawaguchi, *Deep learning without poor local minima*, in Advances in Neural In-

formation Processing Systems 29, 2016, pp. 586–594, http://papers.nips.cc/paper/6112-deep-learning-without-poor-local-minima.pdf.

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

[23] Q. Le and R. Gribonval, *Structured Support Exploration For Multilayer Sparse Matrix Factorization*, in ICASSP 2021 - IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, Ontario, Canada, June 2021, IEEE, pp. 1–5, https://doi.org/10.1109/ICASSP39728.2021.9414238, https://hal.inria.fr/hal-03132013. This paper is associated to code for reproducible research available at https://hal.inria.fr/hal-03572265.

[24] Q. Le, R. Gribonval, and E. Riccietti, *Code for reproducible research - "Spurious Valleys, NP-hardness, and Tractability of Sparse Matrix Factorization With Fixed Support"*, May 2022, https://hal.inria.fr/hal-03667186.

[25] Q. Le, L. Zheng, E. Riccietti, and R. Gribonval, *Fast learning of fast transforms, with guarantees*, in ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, Singapore, May 2022, https://hal.inria.fr/hal-03438881. This paper is associated to code for reproducible research available at https://hal.inria.fr/hal-03552956.

[26] 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), 2015, pp. 3287–3291, https://doi.org/10.1109/ICASSP.2015.7178579.

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

[28] Q. Li, Z. Zhu, and G. Tang, *The non-convex geometry of low-rank matrix optimization*, Information and Inference: A Journal of the IMA, 8 (2018), pp. 51–96, https://doi.org/10.1093/imaiai/iay003, https://doi.org/10.1093/imaiai/iay003, https://arxiv.org/abs/https://academic.oup.com/imaiai/article-pdf/8/1/51/28053147/iay003.pdf.

[29] Q. Nguyen, *On connected sublevel sets in deep learning*, in Proceedings of the 36th International Conference on Machine Learning, vol. 97, 2019, pp. 4790–4799, http://proceedings.mlr.press/v97/nguyen19a.html.

[30] Q. Nguyen and M. Hein, *The loss surface of deep and wide neural networks*, 2017, https://arxiv.org/abs/1704.08045.

[31] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, second ed., 2006.

[32] P. Okunev and C. Johnson, *Necessary and sufficient conditions for existence of the LU factorization of an arbitrary matrix*, arXiv preprint, (2005).

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

[34] R. Rubinstein, A. Bruckstein, and M. Elad, *Dictionaries for sparse representation modeling*, Proceedings of the IEEE, 98 (2010), pp. 1045 – 1057, https://doi.org/10.1109/JPROC.2010.2040551.

[35] R. Rubinstein, M. Zibulevsky, and M. Elad, *Double sparsity: Learning sparse dictionaries for sparse signal approximation*, IEEE Transactions on Signal Processing, 58 (2010), pp. 1553 – 1564, https://doi.org/10.1109/TSP.2009.2036477.

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

[37] V. Silva and L.-H. Lim, *Tensor rank and the ill-posedness of the best low-rank approximation problem*, SIAM Journal on Matrix Analysis and Applications, 30 (2006), p. 1084–1127, https://doi.org/10.1137/06066518X.

[38] J. Sun, Q. Qu, and J. Wright, *A geometric analysis of phase retrieval*, in 2016 IEEE International Symposium on Information Theory (ISIT), 2016, pp. 2379–2383, https://doi.org/10.1109/ISIT.2016.7541725.

[39] I. Tošić and P. Frossard, *Dictionary learning*, IEEE Signal Processing Magazine, 28 (2011), pp. 27–38, https://doi.org/10.1109/MSP.2010.939537.

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

[41] S. A. Vavasis, *On the complexity of nonnegative matrix factorization*, SIAM Journal on Optimization, 20 (2010), pp. 1364–1377, https://doi.org/10.1137/070709967, https://doi.org/10.1137/070709967, https://arxiv.org/abs/https://doi.org/10.1137/070709967.

[42] L. Venturi, A. S. Bandeira, and J. Bruna, *Spurious valleys in one-hidden-layer neural network optimization landscapes*, Journal of Machine Learning Research, 20 (2019), pp. 1–34, http://jmlr.org/papers/v20/18-674.html.

[43] B. Yalcin, H. Zhang, J. Lavaei, and S. Sojoudi, *Factorization approach for low-complexity matrix completion problems: Exponential number of spurious solutions and failure of gradient methods*, 2021, https://doi.org/10.48550/ARXIV.2110.10279, https://arxiv.org/abs/2110.10279.

[44] L. Zheng, R. Gribonval, and E. Riccietti, *Identifiability in exact multilayer sparse matrix factorization*, CoRR, abs/2110.01230 (2021), https://arxiv.org/abs/2110.01230, https://arxiv.org/abs/2110.01230.
[45] 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 (2019), pp. 279–292.

Appendix A. Proof of Lemma 2.3. 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 2.3.

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

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

Figure 12 illustrates an example of support constraints built from W.

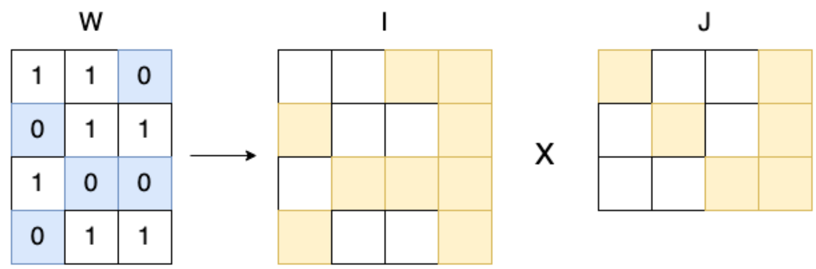

Fig. 12. *Factor supports I and J constructed from the weighted matrix $W \in \{0,1\}^{4\times 3}$. Colored squares in I and J are positions in the supports.*

We consider the (FSMF) with the same matrix A and I, J defined as in Equation (A.1). 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:

$$(XY^\top) \odot W = (X_{\bullet,n+1}Y^\top_{n+1,\bullet}) \odot W \tag{A.2}$$

We will prove that (FSMF) and (MCPO) share the same infimum[5]. 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:

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

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:

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

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

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

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

Therefore, $\mu_2 \le \mu_1$.

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

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

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

This shows that $\mu_1 = \mu_2$. Moreover, the proofs of $\mu_1 \le \mu_2$ and $\mu_2 \le \mu_1$ 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 2.3, we focus on the infimum instead of minimum since there are cases where the infimum is not attained. Indeed, consider the following instance of (FSMF) with: $A = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$, $I = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$, $J = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$. The infimum of this problem is zero, which can be shown by choosing: $X_k = \left[\begin{smallmatrix} -k & k \\ 0 & \frac{1}{k} \end{smallmatrix}\right]$, $Y_k = \left[\begin{smallmatrix} k & k \\ 0 & \frac{1}{k} \end{smallmatrix}\right]$. In the limit, when k goes to infinity, we have:

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

Yet, there does not exist any couple (X, Y) such that $\|A - XY^\top\|^2 = 0$. Indeed, any such couple would need to satisfy: $X_{1,2}Y_{2,2} = 1, X_{2,2}Y_{1,2} = 1, X_{2,2}Y_{2,2} = 0$. 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 weight binary matrix $W = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ (the construction is similar to one in the proof of Lemma 2.3). Problem (MCPO) with (A, W) has unattainable infimum as well. Note that this choice of (I, J) also makes this instance of (FSMF) equivalent to the problem of **LU** decomposition of matrix A.

Appendix B. Proofs for section 3.

B.1. Proof of Lemma 3.5. Denote $\mathcal{P}$ the partition of $[\![r]\!]$ into equivalence classes defined by the rank-one supports associated to (I, J), and $\mathcal{P}^\star \subseteq \mathcal{P}$ the corresponding CECs. Since $T \subseteq [\![r]\!]$ is precisely the set of indices of CECs, and since I_T (resp. J_T) is the restriction of I (resp. of J) to columns indexed by T, the partition of $[\![r]\!]$ into equivalence classes *w.r.t* (I_T, J_T) is precisely $\mathcal{P}^\star$, and for $P \in \mathcal{P} \backslash \mathcal{P}^*$, we have

$\mathcal{S}_P = \emptyset$. WLOG, we assume $\mathcal{P}^\star = \{P_i \mid 1 \leq i \leq \ell\}$. Denote $\mathcal{P}_k = \{P_1, \ldots, P_k\}$, $\mathcal{S}_{\mathcal{P}_k} = \cup_{1 \leq i \leq k} \mathcal{S}_{P_i}$ for $1 \leq k \leq \ell$ and $\mathcal{S}_{P_0} = \emptyset$. We prove below that $(X, Y) = \text{SVD\_FSMF}(A, I_T, J_T)$ satisfies:

$$X_{P_k} Y_{P_k}^\top = A \odot (\mathcal{S}_{\mathcal{P}_k} \setminus \mathcal{S}_{\mathcal{P}_{k-1}}), \forall\, 1 \leq k \leq \ell, \tag{B.1}$$

which implies: $XY^\top = \sum_{P \in \mathcal{P}^\star} X_P Y_P^\top = \sum_{k=1}^{\ell} A \odot (\mathcal{S}_{\mathcal{P}_k} \setminus \mathcal{S}_{\mathcal{P}_{k-1}}) = A \odot \mathcal{S}_\ell = A \odot \mathcal{S}_T = A$ (since we assume $\mathtt{supp}(A) = \mathcal{S}_T$). This yields the conclusion since $\mathtt{supp}(X) \subseteq I_T$ and $\mathtt{supp}(Y) \subseteq J_T$ by definition of SVD_FSMF$(\cdot)$.

We prove Equation (B.1) by induction on ℓ. To ease the reading, in this proof, we denote C_{P_k}, R_{P_k} (Definition 3.4) by C_k, R_k respectively.

For $\ell = 1$ it is sufficient to consider $k = 1$: we have $\mathcal{S}_{\mathcal{P}_1} \setminus \mathcal{S}_{\mathcal{P}_0} = C_1 \times R_1$. Since $\min(|R_1|, |C_1|) \leq |P_1|$ (Definition 3.4), taking the best rank-$|P_1|$ approximation of $A \odot (R_1 \times C_1)$ (whose rank is at most $\min(|R_1|, |C_1|)$) yields $X_{P_1} Y_{P_1}^\top = A \odot (R_1 \times C_1) = A \odot (\mathcal{S}_{\mathcal{P}_1} \setminus \mathcal{S}_{\mathcal{P}_0})$.

Assume that Equation (B.1) holds for $\ell - 1$. We prove its correctness for ℓ. Consider: $A' := A - \sum_{k<\ell} X_{P_k} Y_{P_k}^\top = A - A \odot \mathcal{S}_{\mathcal{P}_{\ell-1}} = A \odot \bar{\mathcal{S}}_{\mathcal{P}_{\ell-1}}$. Therefore, $A' \odot \mathcal{S}_{P_\ell} = A \odot (\mathcal{S}_{\mathcal{P}_\ell} \setminus \mathcal{S}_{\mathcal{P}_{\ell-1}})$. Again, since $\min(|R_\ell|, |C_\ell|) \leq |P_\ell|$ (Definition 3.4), taking the best rank-$|P_\ell|$ approximation of $A' \odot \mathcal{S}_{P_\ell} = A' \odot (R_\ell \times C_\ell)$ (whose rank is at most $\min(|R_\ell|, |C_\ell|)$) yields $X_{P_\ell} Y_{P_\ell}^\top = A' \odot (R_\ell \times C_\ell) = A \odot (\mathcal{S}_{P_\ell} \setminus \mathcal{S}_{P_{\ell-1}})$. That implies Equation (B.1) is correct for all ℓ.

B.2. Proof of Theorem 3.8. First, we decompose the factors X and Y using the taxonomy of indices from Definition 3.7.

DEFINITION B.1. *Given I_T, J_T and $I_{\bar{T}}^i, J_{\bar{T}}^i, i = 1, 2$ as in Definition 3.7, consider (X, Y) a feasible point of (FSMF), we denote:*
1) $X_T = X \odot I_T, X_{\bar{T}}^i = X \odot I_{\bar{T}}^i$, for $i = 1, 2$.
2) $Y_T = Y \odot I_T, Y_{\bar{T}}^i = Y \odot I_{\bar{T}}^i$, for $i = 1, 2$.
with $\odot$ the Hadamard product between a matrix and a support constraint (introduced in subsection 1.1).

The following is a technical result.

LEMMA B.2. *Given I, J support constraints of (FSMF), consider $T, \mathcal{S}_T, \mathcal{S}_\mathcal{P}$ as in Definition 3.2, $X_T, X_{\bar{T}}^i, Y_T, Y_{\bar{T}}^i$ as in Definition 3.6 and assume that for all $k \in \bar{T}$, $\mathcal{S}'_k$ is rectangular. It holds:*

C1 $\mathtt{supp}(X_T Y_T^\top) \subseteq \mathcal{S}_T$.
C2 $\mathtt{supp}(X_{\bar{T}}^1 (Y_{\bar{T}}^1)^\top) \subseteq \mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T$.
C3 $\mathtt{supp}(X_{\bar{T}}^i (Y_{\bar{T}}^j)^\top) \subseteq \mathcal{S}_T, \forall 1 \leq i, j \leq 2, (i, j) \neq (1, 1)$.

Proof. We justify (**C1**)-(**C3**) as follow:

- **C1**: Since $X_T Y_T^\top = \sum_{i \in T} X_{\bullet,i} Y_{\bullet,i}^\top$, $\mathtt{supp}(X_T Y_T^\top) \subseteq \cup_{i \in T} \mathcal{S}_k = \mathcal{S}_T$.
- **C2**: Consider the coefficient (i, j) of $(X_{\bar{T}}^1)(Y_{\bar{T}}^1)^\top$

$$((X_{\bar{T}}^1)(Y_{\bar{T}}^1)^\top)_{i,j} = \sum_k (X_{\bar{T}}^1)_{i,k} (Y_{\bar{T}}^1)_{j,k} = \sum_{(i,k) \in I_{\bar{T}}^1, (j,k) \in J_{\bar{T}}^1} X_{i,k} Y_{j,k}$$

By the definition of $I_{\bar{T}}^1, J_{\bar{T}}^1$, $(X_{\bar{T}}^1)(Y_{\bar{T}}^1)_{i,j}^\top \neq 0$ iff $(i, j) \in \cup_{\ell \in \bar{T}} R_\ell \times C_\ell = \mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T$.
- **C3**: We prove for the case of $(X_{\bar{T}}^1)(Y_{\bar{T}}^2)^\top$. Others can be proved similarly.

$$((X_{\bar{T}}^1)(Y_{\bar{T}}^2)^\top)_{i,j} = \sum_k (X_{\bar{T}}^1)_{i,k} (Y_{\bar{T}}^2)_{j,k} = \sum_{(i,k) \in I_{\bar{T}}^1, (j,k) \in J_{\bar{T}}^2} X_{i,k} Y_{j,k} \tag{B.2}$$

Since $\forall \ell \in \bar{T}, \mathcal{S}'_\ell$ is rectangular, $\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T = \cup_{\ell \in \bar{T}} \mathcal{S}'_\ell = \cup_{\ell \in \bar{T}} R_\ell \times C_\ell$. If $(i,j) \in \mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T$, Equation (B.2) shows that $((X^1_{\bar{T}})(Y^2_{\bar{T}})^\top)_{i,j} = 0$ since there is no k such that $(i,k) \in I^2_{\bar{T}}, (j,k) \in J^2_{\bar{T}}$ due to the definition of $I^1_{\bar{T}}, J^2_{\bar{T}})$. Moreover, $\mathtt{supp}((X^1_{\bar{T}})(Y^2_{\bar{T}})^\top) \subseteq \mathcal{S}_\mathcal{P}$ (since $\mathtt{supp}(X^1_{\bar{T}}) \subseteq I, \mathtt{supp}(Y^2_{\bar{T}}) \subseteq J$). Thus, it shows that $\mathtt{supp}((X^1_{\bar{T}})(Y^2_{\bar{T}})^\top) \subseteq \mathcal{S}_\mathcal{P} \setminus (\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T) = \mathcal{S}_T$. □

Here, we present the proof of Theorem 3.8.

Proof of Theorem 3.8. Given X, Y feasible point of the input (A, I, J), 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 (FSMF) with (A, I, J) and with $(A', I^1_{\bar{T}}, J^1_{\bar{T}})$ $(A' = A \odot \bar{\mathcal{S}}_T)$ respectively.

First, we remark that $I^1_{\bar{T}}$ and $J^1_{\bar{T}}$ satisfy the assumptions of Theorem 3.3. 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 3.3.

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 $C \odot D = \mathbf{0}$, $C \neq D, C, D \in \{\mathcal{S}_T, \mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T, \bar{\mathcal{S}}_\mathcal{P}\}$. From the definition of A' it holds $A' \odot \bar{\mathcal{S}}_\mathcal{P} = A \odot \bar{\mathcal{S}}_\mathcal{P}$ and $A' \odot \mathcal{S}_T = \mathbf{0}$. Moreover, it holds $(X^1_{\bar{T}})(Y^1_{\bar{T}})^\top \odot \mathcal{S}_T \cup \bar{\mathcal{S}}_\mathcal{P} = \mathbf{0}$ due to **C2**.

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

$$XY^\top = X_T Y_T^\top + \sum_{1 \leq i,j \leq 2} (X^i_{\bar{T}})(Y^j_{\bar{T}})^\top. \tag{B.3}$$

Consider the loss function of (FSMF) with input $(A', I^1_{\bar{T}}, J^1_{\bar{T}})$ and solution $(X^1_{\bar{T}}, Y^1_{\bar{T}})$:

$$\begin{aligned}
&\|A' - X^1_{\bar{T}}(Y^1_{\bar{T}})^\top\|^2 \\
&= \|(A' - X^1_{\bar{T}}(Y^1_{\bar{T}})^\top) \odot \mathcal{S}_T\|^2 + \|(A' - X^1_{\bar{T}}(Y^1_{\bar{T}})^\top) \odot (\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T)\|^2 \\
&\qquad + \|(A' - X^1_{\bar{T}}(Y^1_{\bar{T}})^\top) \odot \bar{\mathcal{S}}_\mathcal{P}\|^2 \\
&\overset{\mathbf{C2}}{=} \|(A' - (X^1_{\bar{T}})(Y^1_{\bar{T}})^\top) \odot \mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T\|^2 + \|A' \odot \bar{\mathcal{S}}_\mathcal{P}\|^2 \\
&\overset{\mathbf{C1}+\mathbf{C3}}{=} \|(A - X_T Y_T^\top - \sum_{1 \leq i,j \leq 2} (X^i_{\bar{T}})(Y^j_{\bar{T}})^\top) \odot (\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T)\|^2 + \|A \odot \bar{\mathcal{S}}_\mathcal{P}\|^2 \\
&\overset{(B.3)}{=} \|(A - XY^\top) \odot (\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T)\|^2 + \|A \odot \bar{\mathcal{S}}_\mathcal{P}\|^2
\end{aligned} \tag{B.4}$$

Perform the same calculation with (A, I, J) and solution (X, Y):

$$\begin{aligned}
&\|(A - XY^\top)\|^2 \\
&= \|(A - XY^\top) \odot \mathcal{S}_T\|^2 + \|(A - XY^\top) \odot (\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T)\|^2 + \|(A - XY^\top) \odot \bar{\mathcal{S}}_\mathcal{P}\|^2 \\
&= \|(A - XY^\top) \odot \mathcal{S}_T\|^2 + \|(A - XY^\top) \odot (\mathcal{S}_\mathcal{P} \setminus \mathcal{S}_T)\|^2 + \|A \odot \bar{\mathcal{S}}_\mathcal{P}\|^2
\end{aligned} \tag{B.5}$$

where the last equality holds since $\mathtt{supp}(XY^\top) \subseteq \mathcal{S}_\mathcal{P}$. Therefore, for any feasible point (X, Y) of instance (A, I, J), we can choose $\tilde{X} = X^1_{\bar{T}}, \tilde{Y} = Y^1_{\bar{T}}$ feasible point of $(A', I^1_{\bar{T}}, J^1_{\bar{T}})$ such that $\|A - XY^\top\| \geq \|A' - \tilde{X}\tilde{Y}^\top\|$ (Equation (B.4) and Equation (B.5)). This shows $\mu_1 \geq \mu_2$.

On the other hand, given any feasible point $(\tilde{X}, \tilde{Y})$ of instance $(A', I^1_{\bar{T}}, J^1_{\bar{T}})$, we can construct a feasible point (X, Y) for instance (A, I, J) such that $\|A - XY^\top\|^2 = \|A' - X'Y'^\top\|^2$. We construct $(X, Y) = (X_T + X^1_{\bar{T}} + X^2_{\bar{T}}, Y_T + Y^1_{\bar{T}} + Y^2_{\bar{T}})$ where:

1) $X^1_{\bar{T}} = \tilde{X}, Y^1_{\bar{T}} = \tilde{Y}$,
2) $X^2_{\bar{T}}, Y^2_{\bar{T}}$ can be chosen arbitrarily such that $\mathtt{supp}(X^2_{\bar{T}}) \subseteq I^2_{\bar{T}}, \mathtt{supp}(Y^2_{\bar{T}}) \subseteq J^2_{\bar{T}}$
3) X_T and Y_T such that $\mathtt{supp}(X_T) \subseteq I_T, \mathtt{supp}(Y_T) \subseteq J_T$ and:

$$X_T Y_T^\top = (A - (X^1_{\bar{T}} + X^2_{\bar{T}})(Y^1_{\bar{T}} + Y^2_{\bar{T}})^\top) \odot \mathcal{S}_T$$

(X_T, Y_T) exists due to Lemma 3.5. By Lemma B.2, with this choice we have:

$$\begin{aligned} (A - XY^\top) \odot \mathcal{S}_T &\overset{(B.3)}{=} (A - (X^1_{\bar{T}} + X^2_{\bar{T}})(Y^1_{\bar{T}} + Y^2_{\bar{T}})^\top - X_T Y_T^\top) \odot \mathcal{S}_T \\ &\overset{\mathbf{C1}}{=} (A - (X^1_{\bar{T}} + X^2_{\bar{T}})(Y^1_{\bar{T}} + Y^2_{\bar{T}})^\top) \odot \mathcal{S}_T) - X_T Y_T^\top = \mathbf{0} \end{aligned} \tag{B.6}$$

Therefore $\|A - XY^\top\|^2 = \|A' - \tilde{X}\tilde{Y}^\top\|^2$ (Equation (B.4) and Equation (B.5)). Thus, $\mu_2 \geq \mu_1$. We obtain $\mu_1 = \mu_2$. In addition, given (X, Y) an optimal solution of (FSMF) with instance (A, I, J), we have shown how to construct an optimal solution $(\tilde{X}, \tilde{Y})$ with instance $(A \odot \bar{\mathcal{S}}_T, I^1_{\bar{T}}, J^1_{\bar{T}})$ and vice versa. That completes our proof. ∎

The following Corollary is a direct consequence of the proof of Theorem 3.8.

COROLLARY B.3. *With the same assumptions and notations as in Theorem 3.8, a feasible point (X, Y) (i.e., such that $\mathtt{supp}(X) \subseteq I, \mathtt{supp}(Y) \subseteq J$) is an optimal solution of (FSMF) if and only if:*
1) $(X \odot I^1_{\bar{T}}, Y \odot J^1_{\bar{T}})$ is an optimal solution of (FSMF) with $(A \odot \bar{\mathcal{S}}_T, I^1_{\bar{T}}, J^1_{\bar{T}})$.
2) The following equation holds: $(A - XY^\top) \odot \mathcal{S}_T = \mathbf{0}$

Remark B.4. In the proof of Theorem 3.8, for an optimal solution, one can choose $X^2_{\bar{T}}, Y^2_{\bar{T}}$ arbitrarily. If we choose $X^2_{\bar{T}} = \mathbf{0}, Y^2_{\bar{T}} = \mathbf{0}$, thanks to (B.6), X_T and Y_T has to satisfy:

$$X_T Y_T^\top = (A - (X^1_{\bar{T}} + X^2_{\bar{T}})(Y^1_{\bar{T}} + Y^2_{\bar{T}})^\top) \odot \mathcal{S}_T = (A - X^1_{\bar{T}}(Y^1_{\bar{T}})^\top) \odot \mathcal{S}_T \overset{\mathbf{C2}}{=} A \odot \mathcal{S}_T$$

Appendix C. Proofs for a key lemma. In this section, we will introduce an important technical lemma. It is used extensively for the proof of the tractability and the landscape of (FSMF) under the assumptions of Theorem 3.8, cf. Appendix D.4.

LEMMA C.1. *Consider I, J support constraints of (FSMF) such that $\mathcal{P}^\star = \mathcal{P}$. For any CEC-full-rank feasible point (X, Y) and continuous function $g : [0,1] \to \mathbb{R}^{m \times n}$ satisfying $\mathtt{supp}(g(t)) \subseteq \mathcal{S}_T$ (Definition 3.4) and $g(0) = XY^\top$, there exists a feasible 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:*
A1 *$f(0) = (X_T, Y_T)$.*
A2 *$g(t) = X_f(t) Y_f(t)^\top, \forall t \in [0,1]$.*
A3 *$\|f(z) - f(t)\|^2 \leq \mathcal{C}\|g(z) - g(t)\|^2, \forall t, z \in [0,1]$.*
where $\mathcal{C} = \max\limits_{P \in \mathcal{P}^\star} \left(\max \left(\left|\left|\left| X^\dagger_{R_P,P} \right|\right|\right|^2, \left|\left|\left| Y^\dagger_{C_P,P} \right|\right|\right|^2 \right) \right)$ ($D^\dagger$ and $|||D|||$ denote the pseudo-inverse and operator norm of a matrix D respectively).

Lemma C.1 consider the case where $\mathcal{P}$ only contains CECs. Later in other proofs, we will control the factors (X, Y) by decomposing $X = X_T + X_{\bar{T}}$ (and $Y = Y_T + X_{\bar{T}}$) ($T, \bar{T}$ defined in Definition 3.4) and manipulate (X_T, Y_T) and $(X_{\bar{T}}, Y_{\bar{T}})$ separately. Since the supports of (X_T, Y_T) satisfy Lemma C.1, it provides us a tool to work with (X_T, Y_T).

The proof of Lemma C.1 is carried out by induction. We firstly introduce and prove two other lemmas: Lemma C.2 and Lemma C.3. While Lemma C.2 is Lemma C.1 without support constraints, Lemma C.3 is Lemma C.1 where $|\mathcal{P}^\star| = 1$.

LEMMA C.2. *Let $X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{n\times r}, \min(m,n) \le r$ 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 $f : [0,1] \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} : f(t) = (X_f(t), Y_f(t))$ such that:*
1) $f(0) = (X, Y)$.
2) $g(t) = X_f(t)Y_f(t)^\top, \forall t \in [0,1]$.
3) $\|f(z) - f(t)\|^2 \le \mathcal{C}\|g(z) - g(t)\|^2, \forall t, z \in [0,1]$.
where $\mathcal{C} = \max\left(\left|\left|\left|X^\dagger\right|\right|\right|^2, \left|\left|\left|Y^\dagger\right|\right|\right|^2\right)$.

Proof. WLOG, we can assume that X has full row rank. We define f as:

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

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. Since $\|f(z) - f(t)\|^2 = \|Y_f(z) - Y_f(t)\|^2 = \|X^\dagger(g(z) - g(t))\|^2$, the third one is also satisfied as:

$$\|f(z) - f(t)\|^2 = \|X^\dagger(g(z) - g(t))\|^2 \le \left|\left|\left|X^\dagger\right|\right|\right|^2 \|g(z) - g(t)\|^2 \le \mathcal{C}\|g(z) - g(t)\|^2$$ ◻

LEMMA C.3. *Consider I, J support of* (FSMF) *where $\mathcal{P}^\star = \mathcal{P} = \{P\}$, for any feasible CEC-full-rank point (X, Y) and continuous function $g : [0,1] \to \mathbb{R}^{m\times n}$ satisfying $\mathtt{supp}(g(t)) \subseteq \mathcal{S}_P$ (Definition 3.2) and $g(0) = XY^\top$, there exists a feasible 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:*
B1 *$f(0) = (X, Y)$.*
B2 *$g(t) = X_f(t)Y_f(t)^\top, \forall t \in [0,1]$.*
B3 *$\|f(z) - f(t)\|^2 \le \mathcal{C}\|g(z) - g(t)\|^2$.*
where $\mathcal{C} = \max\left(\left|\left|\left|X^\dagger_{R_P,P}\right|\right|\right|^2, \left|\left|\left|Y^\dagger_{C_P,P}\right|\right|\right|^2\right)$.

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

Since $\mathcal{P}^\star = \mathcal{P} = \{P\}$, a continuous feasible function $f(t)$ must have the form: $X_f(t) = \left[\begin{smallmatrix}\tilde{X}_f(t) & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{smallmatrix}\right]$ and $Y_f(t) = \left[\begin{smallmatrix}\tilde{Y}_f & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{smallmatrix}\right]$ where $\tilde{X}_f : [0,1] \to \mathbb{R}^{|R_P|\times|P|}, \tilde{Y}_f : [0,1] \to \mathbb{R}^{|C_P|\times|P|}$ are continuous functions. f is fully determined by $(\tilde{X}_f(t), \tilde{Y}_f(t))$.

Moreover, if $g : [0,1] \to \mathbb{R}^{m\times n}$ satisfying $\mathtt{supp}(g(t)) \subseteq \mathcal{S}_T$, then g has to have the form: $g(t) = \left[\begin{smallmatrix}\tilde{g} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{smallmatrix}\right]$ where $\tilde{g} : [0,1] \to \mathbb{R}^{|R_P|\times|C_P|}$ is a continuous function.

Since $g(0) = XY^\top$, $\tilde{g}(0) = (X_{R_P,P})(Y_{C_P,P})^\top$. Thus, to satisfy each constraint **B1-B3**, it is sufficient to find $\tilde{X}_f$ and $\tilde{Y}_f$ such that:

B1: $\tilde{X}_f(0) = X_{R_P,P}, \tilde{Y}_f(0) = Y_{C_P,P}$.

B2: $\tilde{g}(t) = \tilde{X}_f(t)\tilde{Y}_f(t)^\top, \forall t \in [0,1]$ because:

$$X_f(t)Y_f(t)^\top = \begin{pmatrix}\tilde{X}_f(t)\tilde{Y}_f(t)^\top & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} = \begin{pmatrix}\tilde{g}(t) & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} = g(t)$$

B3: $\|X'(z) - X'(t)\|^2 + \|Y'(z) - Y'(t)\|^2 \le \mathcal{C}\|A'(z) - A'(t)\|^2$ since $\|X'_f(z) - X_f(t)\|^2 + \|Y'_f(z) - Y_f(t)\|^2 = \|f(z) - f(t)\|^2$ and $\|A'(z) - A'(t)\|^{=}\|g(z) - g(t)\|^2$.

Such function exists thanks Lemma C.2 (since we assume $X_{R_P,P}$ has full rank). ◻

Proof of Lemma C.1. We prove by induction on the size $\mathcal{P}$. By Lemma C.3 the result is true if $|\mathcal{P}| = 1$. Assume the result is true if $|\mathcal{P}| \leq p$. We consider the case where $|\mathcal{P}| = p + 1$. Let $P \in \mathcal{P}$ and partition $\mathcal{P}$ into $\mathcal{P}' = \mathcal{P} \setminus \{P\}$ and $\{P\}$. Let $T' = \cup_{P' \in \mathcal{P}'} P' = T \setminus P$. Since $|\mathcal{P}'| = p$, we can use induction hypothesis. Define:

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

We verify that the function $h_1(t)$ satisfying the hypotheses to use induction step: h_1 continuous, $\mathtt{supp}(h_1(t)) \subseteq \mathcal{S}_{\mathcal{P}'}$ and finally $h_1(0) = (g(0) - X_P Y_P^\top) \odot \mathcal{S}_{\mathcal{P}'} = X_{T'} Y_{T'}^\top \odot \mathcal{S}_{\mathcal{P}'} = X_{T'} Y_{T'}^\top$. Using the induction hypothesis with $\mathcal{P}'$, there exists a function $f_1 : [0,1] \to \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r} : f_1(t) = (X_f^1(t), Y_f^1(t))$ such that:

1) $\mathtt{supp}(X_f^1(t)) \subseteq I_{T'}, \mathtt{supp}(Y_f^1(t)) \subseteq J_{T'}$.
2) $f_1(0) = (X_{T'}, Y_{T'})$.
3) $h_1(t) = X_f^1(t) Y_f^1(t)^\top, \forall t \in [0,1]$.
4) $\|f_1(z) - f_1(t)\|^2 \leq \mathcal{C}' \|h_1(z) - h_1(t)\|^2$.

where $\mathcal{C}' = \max\limits_{P' \in \mathcal{P}'} \left(\max \left(\left|\left|\left| X_{R_{P'},P'}^\dagger \right|\right|\right|^2, \left|\left|\left| Y_{C_P,P}^\dagger \right|\right|\right|^2 \right) \right)$.

On the other hand, $h_2(t)$ satisfies the assumptions of Lemma C.3: $h_2(t)$ is continuous and $\mathtt{supp}(h_2(t)) = \mathtt{supp}(X_P Y_P^\top \odot \mathcal{S}_{\mathcal{P}'} + g(t) \odot \mathcal{S}_P \setminus \mathcal{S}_{\mathcal{P}'}) \subseteq \mathtt{supp}(X_P Y_P^\top) \cup (\mathcal{S}_P \setminus \mathcal{S}_{\mathcal{P}'}) = \mathcal{S}_P$.

In addition, since $g(0) \odot \mathcal{S}_P \setminus \mathcal{S}_{\mathcal{P}'} = (XY^\top) \odot \mathcal{S}_P \setminus \mathcal{S}_{\mathcal{P}'} = (X_{T'} Y_{T'}^\top + X_P Y_P^\top) \odot \mathcal{S}_P \setminus \mathcal{S}_{\mathcal{P}'} = X_P Y_P^\top \odot \mathcal{S}_P \setminus \mathcal{S}_{\mathcal{P}'}$, we have $h_2(0) = X_P Y_P^\top \odot \mathcal{S}_{\mathcal{P}'} + g(0) \odot \mathcal{S}_P \setminus \mathcal{S}_{\mathcal{P}'} = X_P Y_P^\top \odot (\mathcal{S}_{\mathcal{P}'} + \mathcal{S}_P \setminus \mathcal{S}_{\mathcal{P}'}) = X_P Y_P^\top$. Invoking Lemma C.3 with the singleton $\{P\}$, there exists a function $(X_f^2(t), Y_f^2(t))$ such that:

1) $\mathtt{supp}(X_f^2(t)) \subseteq I_P, \mathtt{supp}(Y_f^2(t)) \subseteq J_P$.
2) $f_2(0) = (X_P, Y_P)$.
3) $h_2(t) = X_f^2(t) Y_f^2(t)^\top, \forall t \in [0,1]$.
4) $\|f_2(z) - f_2(t)\|^2 \leq \max \left(\left|\left|\left| X_{R_P,P}^\dagger \right|\right|\right|^2, \left|\left|\left| Y_{C_P,P}^\dagger \right|\right|\right|^2 \right) \|h_2(z) - h_2(t)\|^2$.

We construct the functions $f(t) = (X_f(t), Y_f(t))$ as:

$$X_f(t) = X_f^1(t) + X_f^2(t), \qquad Y_f(t) = Y_f^1(t) + Y_f^2(t)$$

We verify the validity of this construction. f is clearly feasible due to the supports of $X_f^i(t), Y_f^i(t), i = 1, 2$. The remaining conditions are:

A1:

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

A2:

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

A3:

$$\begin{aligned}
&\|f(z) - f(t)\|^2\\
&= \|f_1(z) - f_1(t)\|^2 + \|f_2(z) - f_2(t)\|^2\\
&\leq \mathcal{C}'\|h_1(z) - h_1(t)\|^2 + \max\left(\left|\left|\left|X^{\dagger}_{R_P,P}\right|\right|\right|^2, \left|\left|\left|Y^{\dagger}_{C_P,P}\right|\right|\right|^2\right)\|h_2(z) - h_2(t)\|^2\\
&\leq \mathcal{C}(\|h_1(z) - h_1(t)\|^2 + \|h_2(z) - h_2(t)\|^2)\\
&= \mathcal{C}(\|(g(z) - g(t)) \odot \mathcal{S}_{\mathcal{P}'}\|^2 + \|(g(z) - g(t)) \odot \mathcal{S}_P \setminus \mathcal{S}_{\mathcal{P}'}\|^2)\\
&= \mathcal{C}\|g(z) - g(t)\|^2
\end{aligned}$$

□

Appendix D. Proofs for section 4.

D.1. Proof of Lemma 4.15. The proof relies on two intermediate results that we state first: Lemma D.1 and Corollary D.2. The idea of Lemma D.1 can be found in [42]. Since it is not formally proved as a lemma or theorem, we reprove it here for self-containedness. In fact, Lemma D.1 and Corollary D.2 are special cases of Lemma 4.15 with no support contraints and $\mathcal{P}^\star = \mathcal{P} = \{P\}$ respectively.

LEMMA D.1. *Let $X \in \mathbb{R}^{R\times p}, Y \in \mathbb{R}^{C\times p}, \min(R, C) \leq p$. There exists a continuous function $f(t) = (X_f(t), Y_f(t))$ on $[0, 1]$ such that:*
- *$f(0 = (X, Y)$.*
- *$XY^\top = X_f(t)(Y_f(t))^\top, \forall t \in [0, 1]$.*
- *$X_f(1)$ or $Y_f(1)$ has full row rank.*

Proof. 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 a matrix $\tilde{Y}$ by their columns as follow:

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

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

$$f_1(t) = (X, (1 - t)Y + t\tilde{Y})$$

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

$$X((1 - t)Y^\top + t\tilde{Y}^\top) = (1 - t)XY^\top + tX\tilde{Y}^\top = XY^\top.$$

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

$$f_2(t) = ((1 - t)X + t\tilde{X}, \tilde{Y})$$

also has their product unchanged (since first q columns of $(1-t)X + t\tilde{X}$ are constant and last $r-q$ rows of $\tilde{Y}$ are zero). Moreover, $f_2(0) = (\tilde{X}, \tilde{Y})$ where $\tilde{X}$ 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. ∎

Corollary D.2. *Consider I, J support constraints of (FSMF) with $\mathcal{P}^\star = \mathcal{P} = \{P\}$. There is a feasible continuous function $f : [0,1] \mapsto \mathbb{R}^{m\times r} \times \mathbb{R}^{n \times r} : f(t) = (X_f(t), Y_f(t))$ such that:*

1. *$f(0) = (X, Y)$;*
2. *$X_f(t)(Y_f(t))^\top = XY^\top, \forall t \in [0,1]$;*
3. *$(X_f(1))_{R_P,P}$ or $(Y_f(1))_{C_P,P}$ has full row rank.*

Proof of Corollary D.2. WLOG, up to permuting columns, we can assume $P = [\![|P|]\!], R_P = [\![|R_P|]\!]$ and $C_P = [\![|C_P|]\!]$ (R_P and C_P are defined in Definition Definition 3.2). A feasible function $f = (X_f(t), Y_f(t))$ has the form:

$$X_f(t) = \begin{pmatrix} \tilde{X}_f(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, Y_f(t) = \begin{pmatrix} \tilde{Y}_f(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

where $\tilde{X}_f : [0,1] \mapsto \mathbb{R}^{R_P \times P}$, $\tilde{Y}_f : [0,1] \mapsto \mathbb{R}^{C_P \times P}$.

Since P is a CEC, we have $p \geq \min(R_P, C_P)$. Hence we can use Lemma D.1 to build $(\tilde{X}_f(t), \tilde{Y}_f(t))$ satisfying all conditions of Lemma D.1. Such $(\tilde{X}_f(t), \tilde{Y}_f(t))$ fully determines f and make f our desirable function. ∎

Proof of Lemma 4.15. First, we decompose X and Y as:

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

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 $P \in \mathcal{P}^\star$, (I_P, J_P) contains one CEC. By applying Corollary D.2, we can build continuous functions $(X_f^P(t), Y_f^P(t))$, $\mathtt{supp}(X_f^P(t)) \subseteq I_P, \mathtt{supp}(Y_f^P(t)) \subseteq J_P, \forall t \in [0,1]$ such that:

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

Our desirable $f(t) = (X_f(t), Y_f(t))$ is defined 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 $f = (X_f(t), Y_f(t))$ is feasible, $f(0) = (X, Y)$, $f(1)$ is CEC-full-rank and $X_f(t)Y_f(t)^\top = XY^\top, \forall t \in [0,1]$. ∎

D.2. Proof of Lemma 4.16. Denote $Z = XY^\top$, we construct f such that $X_f(t)Y_f(t)^\top = B(t)$, where $B(t) = Z \odot \bar{\mathcal{S}}_T + (At + Z(1-t)) \odot \mathcal{S}_T$. Such function f makes $L(X_f(t), Y_f(t))$ non-increasing since:

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

Thus, the rest of the proof is devoted to show that such a function f exists by using Lemma C.1. 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:

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

which shows $\mathtt{supp}(g(t)) \subseteq \mathcal{S}_T$. Since (X_T, Y_T) is CEC-full-rank (by our assumption, (X, Y) is CEC-full-rank), invoking Lemma C.1 with (I_T, J_T), there exists $f^T(t) = (X_f^T(t), Y_f^T(t))$ such that:

D1 $\mathtt{supp}(X_f^T(t)) \subseteq I_T, \mathtt{supp}(Y_f^C(t)) \subseteq J_T$.

D2 $f^T(0) = (X_T, Y_T)$.

D3 $g(t) = X_f^T(t)(Y_f^T(t))^\top, \forall t \in [0, 1]$.

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

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

f is clearly feasible due to (**D1**). The remaining condition to be checked is:

- First condition:

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

- Second condition: holds thanks to Equation (D.1) and:

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

- Third condition:

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

D.3. Proof of Lemma 4.17. Consider $X_T, X_{\bar{T}}^i, Y_T, Y_{\bar{T}}^i, i = 1, 2$ as in Definition B.1. We redefine $A' = A \odot \bar{\mathcal{S}}_T, I' = I_{\bar{T}}^1, J' = J_{\bar{T}}^1$ as in Theorem 3.8.

In light of Corollary B.3, an optimal solution $(\tilde{X}, \tilde{Y})$ has the following form:

1) $\tilde{X}_{\bar{T}}^1 = \tilde{X} \odot I_{\bar{T}}^1, \tilde{Y}_{\bar{T}}^1 = \tilde{Y} \odot J_{\bar{T}}^1$ is an optimal solution of (FSMF) with (A', I', J').
2) $\tilde{X}_{\bar{T}}^2 = \tilde{X} \odot I_{\bar{T}}^2, \tilde{Y}_{\bar{T}}^2 = \tilde{Y} \odot J_{\bar{T}}^2$ can be arbitrary.
3) $\tilde{X}_T = \tilde{X} \odot I_T, \tilde{Y}_T = \tilde{Y} \odot J_T$ satisfy:

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

Since (I', J') has its support constraints satisfying Theorem 3.3 assumptions as shown in Theorem 3.8, by Theorem 4.12, there exists a function $(X_f^{\bar{T}}(t), Y_f^{\bar{T}}(t))$ such that:

1) $\mathtt{supp}(X_f^{\bar{T}}(t)) \subseteq I_{\bar{T}}^1, \mathtt{supp}(Y_f^{\bar{T}}(t)) \subseteq J_{\bar{T}}^1$.
2) $X_f^{\bar{T}}(0) = X_{\bar{T}}^1, Y_f^{\bar{T}}(0) = Y_{\bar{T}}^1$.
3) $L'(X_f^{\bar{T}}(t), Y_f^{\bar{T}}(t)) = \|A' - X_f^{\bar{T}}(t) Y_f^{\bar{T}}(t)^\top\|^2$ is non-increasing.
4) $(X_f^{\bar{T}}(1), Y_f^{\bar{T}}(1))$ is an optimal solution of the instance of (FSMF) with (A', I', J').

Consider the function $g(t) = \left(A - (X_f^{\bar{T}}(t) + X_{\bar{T}}^2)(Y_f^{\bar{T}}(t) + Y_{\bar{T}}^2)^\top\right) \odot \mathcal{S}_T$. This construction makes $g(0) = X_T Y_T^\top$. Indeed,

$$\begin{aligned} g(0) &= \left(A - (X_f^{\bar{T}}(0) + X_{\bar{T}}^2)(Y_f^{\bar{T}}(0) + Y_{\bar{T}}^2)^\top\right) \odot \mathcal{S}_T \\ &= \left(A - (X_{\bar{T}}^1 + X_{\bar{T}}^2)(Y_{\bar{T}}^1 + Y_{\bar{T}}^2)^\top\right) \odot \mathcal{S}_T \\ &\overset{(1)}{=} \left(XY^\top - (X_{\bar{T}}^1 + X_{\bar{T}}^2)(Y_{\bar{T}}^1 + Y_{\bar{T}}^2)^\top\right) \odot \mathcal{S}_T \\ &\overset{(2)}{=} X_T Y_T^\top \end{aligned}$$

where (1) holds by the hypothesis $(A - XY^\top) \odot \mathcal{S}_T = \mathbf{0}$, and (2) holds by Equation (B.3) and $\mathtt{supp}(X_T Y_T^\top) \subseteq \mathcal{S}_T$. Due to our hypothesis (X, Y) is CEC-full-rank, (X_T, Y_T) is CEC-full-rank. In addition, $g(t)$ continuous, $\mathtt{supp}(g(t)) \subseteq \mathcal{S}_T$ and $g(0) = X_T Y_T^\top$. Invoking Lemma C.1 with (I_T, J_T), there exist functions $(X_f^C(t), Y_f^C(t))$ satisfying:

1) $\mathtt{supp}(X_f^T(t)) \subseteq I_T, \mathtt{supp}(Y_f^T(t)) \subseteq J_T$.
2) $f^T(0) = (X_T, Y_T)$.
3) $g(t) = X_f^T(t) Y_f^T(t)^\top, \forall t \in [0, 1]$.

Finally, one can define the function $X_f(t), Y_f(t)$ satisfying Lemma 4.17 as:

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

f is feasible 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 remaining conditions are satisfied as:

- First condition:

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

- Second condition:

$$\begin{aligned} \|A - X_f(t)Y_f(t)^\top\|^2 &= \|A - X_f^T(t)(Y_f^T(t))^\top - (X_f^{\bar{T}}(t) + X_{\bar{T}}^2)(Y_f^{\bar{T}}(t) + Y_{\bar{T}}^2)^\top\|^2 \\ &= \|g(t) - X_f^T(t)Y_f^T(t)^\top\|^2 + \|(A - X_f^{\bar{T}}(t)(Y_f^{\bar{T}}(t))^\top) \odot \mathcal{S}_{\mathcal{P}} \setminus \mathcal{S}_T\|^2 + \|A \odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2 \\ &= \|(A' - X_f^{\bar{T}}(t)(Y_f^{\bar{T}}(t))^\top) \odot \mathcal{S}_{\mathcal{P}} \setminus \mathcal{S}_T\|^2 + \|A \odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2 \\ &\overset{(B.4)}{=} \|A' - X_f^{\bar{T}}(t)(Y_f^{\bar{T}}(t))^\top\|^2 \end{aligned}$$

  Since $\|A' - X_f^{\bar{T}}(t)(Y_f^{\bar{T}}(t))^\top\|^2$ is non-increasing, so is $\|A - X_f(t)Y_f(t)^\top\|^2$.
- Third condition: By Theorem 3.8, $(X_f(1), Y_f(1))$ is a global minimizer since $\|A - X_f(1)Y_f(1)^\top\|^2 = \|A' - X_f^{\bar{T}}(1)(Y_f^{\bar{T}}(1))^\top\|^2$ where $(X_f^{\bar{T}}(1), Y_f^{\bar{T}}(1))$ is an optimal solution of the instance of (FSMF) with (A', I', J').

D.4. Proof of Theorem 4.19. The following corollary is necessary for the proof of Theorem 4.19.

COROLLARY D.3. *Consider I, J support constraints of* (FSMF)*, such that $\mathcal{P}^\star = \mathcal{P}$. Given any feasible CEC-full-rank point (X, Y) and any B satisfying $\mathtt{supp}(B) \subseteq \mathcal{S}_{\mathcal{P}}$, there exists $(\tilde{X}, \tilde{Y})$ such that:*

E1 $\mathtt{supp}(\tilde{X}) \subseteq I, \mathtt{supp}(\tilde{Y}) \subseteq J$

E2 $\tilde{X}\tilde{Y}^\top = B$.

E3 $\|X - \tilde{X}\|^2 + \|Y - \tilde{Y}\|^2 \leq \mathcal{C}\|XY^\top - B\|^2$.

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. Corollary D.3 is an application of Lemma C.1. Consider the function $g(t) = (1-t)XY^\top + tB$. By construction, $g(t)$ is continuous, $g(0) = XY^\top$ and $\texttt{supp}(g(t)) \subseteq \texttt{supp}(XY^\top) \cup \texttt{supp}(B) = \mathcal{S}_\mathcal{P}$. Since (X,Y) is CEC-full-rank, there exists a feasible function $f(t) = (X_f(t), Y_f(t))$ satisfying **A1** - **A3** by using Lemma C.1.

We choose $(\tilde{X}, \tilde{Y}) = (X_f(1), Y_f(1))$. The verification of constraints is as follow:

E1: f is feasible.

E2: $\tilde{X}\tilde{Y}^\top = X_f(1)Y_f(1)^\top \overset{\textbf{A2}}{=} g(1) = B$.

E3: $\|X-\tilde{X}\|^2 + \|Y-\tilde{Y}\|^2 \overset{\textbf{A1}}{=} \|f(1)-f(0)\|^2 \overset{\textbf{A3}}{\leq} \mathcal{C}\|g(0)-g(1)\|^2 \leq \mathcal{C}\|XY^\top - B\|^2$. ◻

Proof of Theorem 4.19. As mentioned in the sketch of the proof, given any (X,Y) not CEC-full-rank, Lemma 4.15 shows the existence of a path f along which L is constant and f connects (X,Y) to some CEC-full-rank $(\tilde{X},\tilde{Y})$. Therefore, this proof will be entirely devoted to show that a feasible CEC-full-rank solution (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:

$$L(X,Y) = \|A - XY^\top\|^2 = \|A - \sum_{P'\in\mathcal{P}^\star} X_{P'}Y_{P'}^\top - X_{\bar{T}}Y_{\bar{T}}^\top\|^2$$

If (X,Y) is truly a local minimum, then $\forall P \in \mathcal{P}^\star$, (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 equal to L but we optimize only w.r.t (X_P, Y_P) while fixing the other coefficients. In other words, (X_P,Y_P) is a local minimum of the problem:

$$\begin{aligned} &\underset{X'\in\mathbb{R}^{m\times r}, Y'\in\mathbb{R}^{n\times r}}{\text{Minimize}} && L'(X',Y') = \|B - X'Y'^\top\|^2 \\ &\text{Subject to:} && \texttt{supp}(X') \subseteq I_P \text{ and } \texttt{supp}(Y') \subseteq J_P \end{aligned}$$

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 4.12). Global minima are attained when $X_PY_P^\top = B \odot \mathcal{S}_P$ due to the expressivity of a CEC (Lemma 3.5). Thus, for any $P \in \mathcal{P}^\star$, $\forall (i,j) \in \mathcal{S}_P$, we have:

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

which implies Equation (4.2).
SECOND STEP: In this step, we assume that Equation (4.2) holds. Consider $X_T, X_{\bar{T}}^i, Y_T, Y_{\bar{T}}^i, i = 1,2$ as in Definition 3.7. Let $A' = A \odot \bar{\mathcal{S}}_T, I' = I_{\bar{T}}^1, J' = J_{\bar{T}}^1$.
We consider two possibilities. First, if $(X_{\bar{T}}^1, Y_{\bar{T}}^1)$ is an optimal solution of the instance of (FSMF) with (A', I', J'), by Corollary B.3, (X,Y) is an optimal solution of (FSMF) with (A,I,J) (since Equation (4.2) holds). 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 (FSMF) with (A', I', J'). We show that in this case, in

any neighborhood of (X, Y), there exists a point (X', Y') such that $\mathtt{supp}(X') \subseteq I$, $\mathtt{supp}(Y') \subseteq J'$ and $L(X, Y) > L(X', Y')$. Thus (X, Y) cannot be a local minimum. Since $(I^1_{\bar{T}}, J^1_{\bar{T}})$ satisfies Theorem 3.3 assumptions, (FSMF) has no spurious local minima (Theorem 4.12). As $(X^1_{\bar{T}}, Y^1_{\bar{T}})$ is not an optimal solution, it cannot be a local minimum either, i.e., in any neighborhood of $(X^1_{\bar{T}}, Y^1_{\bar{T}})$, there exists $(\dot{X}, \dot{Y})$ with $\mathtt{supp}(\tilde{X}^1_{\bar{T}}) \subseteq I', \mathtt{supp}(\tilde{Y}^1_{\bar{T}}) \subseteq J'$ and

$$\|A' - X^1_{\bar{T}}(Y^1_{\bar{T}})^\top\|^2 > \|A' - \tilde{X}^1_{\bar{T}}(\tilde{Y}^1_{\bar{T}})^\top\|^2 \tag{D.2}$$

By Equation (B.4), we have:

$$\begin{aligned}\|A' - (X^1_{\bar{T}})(Y^1_{\bar{T}})^\top\|^2 &= \|(A - (X^1_{\bar{T}})(Y^1_{\bar{T}})^\top) \odot \mathcal{S}_{\mathcal{P}} \setminus \mathcal{S}_T\|^2 + \|A \odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2 \\ \|A' - (\tilde{X}^1_{\bar{T}})(\tilde{Y}^1_{\bar{T}})^\top\|^2 &= \|(A - (\tilde{X}^1_{\bar{T}})(\tilde{Y}^1_{\bar{T}})^\top) \odot \mathcal{S}_{\mathcal{P}} \setminus \mathcal{S}_T\|^2 + \|A \odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2\end{aligned} \tag{D.3}$$

By Equation (D.2) and Equation (D.3) we have:

$$\|(A - (X^1_{\bar{T}})(Y^1_{\bar{T}})^\top) \odot \mathcal{S}_{\mathcal{P}} \setminus \mathcal{S}_T\|^2 > \|(A - \tilde{X}^1_{\bar{T}}(\tilde{Y}^1_{\bar{T}})^\top) \odot \mathcal{S}_{\mathcal{P}} \setminus \mathcal{S}_T\|^2 \tag{D.4}$$

Consider the matrix: $B := \left(A - (\tilde{X}^1_{\bar{T}} + X^2_{\bar{T}})(\tilde{Y}^1_{\bar{T}} + Y^2_{\bar{T}})^\top\right) \odot \mathcal{S}_T$. Since $\mathtt{supp}(B) \subseteq \mathcal{S}_T$ and (X_T, Y_T) is CEC-full-rank (we assume (X, Y) is CEC-full-rank), by Corollary D.3, there exists $(\tilde{X}_T, \tilde{Y}_T)$ such that:

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

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

$$\tilde{X} = \tilde{X}_T + \tilde{X}^1_{\bar{T}} + X^2_{\bar{T}}, \qquad \tilde{Y} = \tilde{Y}_T + \tilde{Y}^1_{\bar{T}} + Y^2_{\bar{T}}$$

The point $(\tilde{X}, \tilde{Y})$ still satisfies Equation (4.2). Indeed,

$$\begin{aligned}(A - \tilde{X}\tilde{Y}^\top) \odot \mathcal{S}_T &= \left(A - \tilde{X}_T\tilde{Y}_T^\top - (\tilde{X}^1_{\bar{T}} + X^2_{\bar{T}})(\tilde{Y}^1_{\bar{T}} + Y^2_{\bar{T}})^\top\right) \odot \mathcal{S}_T \\ &= (B - \tilde{X}_T\tilde{Y}_T^\top) \odot \mathcal{S}_T = \mathbf{0}.\end{aligned} \tag{D.5}$$

It is clear that $(\tilde{X}, \tilde{Y})$ satisfies $\mathtt{supp}(\tilde{X}) \subseteq I$, $\mathtt{supp}(\tilde{Y}) \subseteq J$ due to the support of its components $(\tilde{X}_T, \tilde{Y}_T), (\tilde{X}^1_{\bar{T}}, \tilde{Y}^1_{\bar{T}}), (X^2_{\bar{T}}, Y^2_{\bar{T}})$. Moreover, we have:

$$\begin{aligned}\|A - \tilde{X}\tilde{Y}^\top\|^2 &= \|(A - \tilde{X}\tilde{Y}^\top) \odot \mathcal{S}_T\|^2 + \|(A - \tilde{X}\tilde{Y}^\top) \odot \mathcal{S}_{\mathcal{P}} \setminus \mathcal{S}_T\|^2 + \|A \odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2 \\ &\overset{(D.5)}{=} \|(A - \tilde{X}^1_{\bar{T}}(\tilde{Y}^1_{\bar{T}})^\top) \odot \mathcal{S}_{\mathcal{P}} \setminus \mathcal{S}_T\|^2 + \|A \odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2 \\ &\overset{(D.4)}{<} \|(A - X^1_{\bar{T}}(Y^1_{\bar{T}})^\top) \odot \mathcal{S}_{\mathcal{P}} \setminus \mathcal{S}_T\|^2 + \|A \odot \bar{\mathcal{S}}_{\mathcal{P}}\|^2 \\ &= \|A - XY^\top\|^2.\end{aligned}$$

Lastly, we show that $(\tilde{X}, \tilde{Y})$ can be chosen arbitrarily close to (X, Y) by choosing $(\tilde{X}^1_{\bar{T}}, \tilde{Y}^1_{\bar{T}})$ 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:

$$\begin{aligned}\|X - \tilde{X}\|^2 + \|Y - \tilde{Y}\|^2 &= \|X_T - \tilde{X}_T\|^2 + \|Y_T - \tilde{Y}_T\|^2 + \|X^1_{\bar{T}} - \tilde{X}^1_{\bar{T}}\|^2 + \|Y^1_{\bar{T}} - \tilde{Y}^1_{\bar{T}}\|^2 \\ &\leq \mathcal{C}\|X_TY_T^\top - B\|^2 + \epsilon\end{aligned}$$

We will bound the value $\|X_T Y_T^\top - B\|^2$. By using Equation (4.2), we have:

$$\begin{aligned}(A - \sum_{1\le i,j\le 2}(X_{\bar T}^i)(Y_{\bar T}^j)^\top)\odot \mathcal{S}_T - X_T Y_T^\top &= (A - X_T Y_T^\top - \sum_{1\le i,j\le 2}(X_{\bar T}^i)(Y_{\bar T}^j)^\top)\odot \mathcal{S}_T \\ &= (A - XY^\top)\odot \mathcal{S}_T \overset{(4.2)}{=} \mathbf{0}\end{aligned}$$

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

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

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

D.5. Proof for Remark 4.23. Direct calculation of the Hessian of L at point (X_0, Y_0) is given by:

$$H(L)_{|(X_0,Y_0)} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 100 & 0 & 0 & 0 & 10 & 0 \\ 0 & 0 & 100 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

which is indeed positive semi-definite.

Appendix E. Expressing any hierarchically off-diagonal low-rank matrix (HODLR) as a product of 2 factors with fixed supports. In the following, we report the definition of HODLR matrices. For convenience, we report the definition only for a *square* matrix whose size is a power of two, i.e $n = 2^J, J \in \mathbb{N}$.

Definition E.1 (HODLR matrices). *A matrix $A \in \mathbb{R}^{2^N \times 2^N}$ is called an HODLR matrix if either of the following two holds:*
- *$N = 0$, i.e., $A \in \mathbb{R}^{1\times 1}$.*
- *A has the form $A = \left[\begin{smallmatrix} A_{11} & A_{12} \\ A_{21} & A_{22}\end{smallmatrix}\right]$ for $A_{i,j} \in \mathbb{R}^{2^{N-1}\times 2^{N-1}}, 1 \le i,j \le 2$ such that A_{21}, A_{12} are of rank at most one and $A_{11}, A_{22} \in \mathbb{R}$ are HODLR matrices.*

We prove that any HODLR matrix is a product of two factors with fixed support. The result is proved when A_{12}, A_{21} are of rank at most one, but more generally, if we allow A_{12} and A_{21} to have rank $k \ge 1$, the general scheme of the proof of Lemma E.2 below still works (with the slight modification $|I| = |J| = O(kn\log n)$, $I, J \in \{0,1\}^{2^N \times k(3\times 2^N - 2)}$). We prove that any HODLR matrix is a product of two factors with fixed support.

Lemma E.2. *For each $N \ge 1$ there exists $I, J \in \{0,1\}^{2^N \times (3\times 2^N - 2)}$ support constraints such that for any HODLR matrix $A \in \mathbb{R}^{2^N\times 2^N}$, we have:*
1) A admits a factorization $XY^\top$ and $\mathtt{supp}(X) \subseteq I, \mathtt{supp}(Y) \subseteq J$.
2) $|I| = |J| = O(n\log n)$ $(n = 2^N)$.

3) (I,J) satisfies the assumption of Theorem 3.3.

Proof. The proof is carried out by induction.

1) For $N=1$, one can consider $(I,J) \in \{0,1\}^{2\times 4} \times \{0,1\}^{2\times 4}$ defined (in the binary matrix form) as follows:

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

Any (X,Y) constrained to (I,J) will have the following form:

$$X = \begin{pmatrix} x_1 & 0 & x_3 & 0 \\ 0 & x_2 & 0 & x_4 \end{pmatrix}, \; Y = \begin{pmatrix} 0 & y_2 & y_3 & 0 \\ y_1 & 0 & 0 & y_4 \end{pmatrix}, \; XY^\top = \begin{pmatrix} x_3y_3 & x_1y_1 \\ x_2y_2 & x_4y_4 \end{pmatrix}.$$

Given any matrix $A \in \mathbb{R}^{2\times 2}$ (and in particular, given any HODLR matrix in this dimension) it is easy to see that A can be represented as $XY^\top$ such that $\mathtt{supp}(X) \subseteq I, \mathtt{supp}(Y) \subseteq J$ (take e.g. $x_3 = a_{11}$, $x_1 = a_{12}$, $x_2 = a_{21}$, $x_4 = a_{22}$ and all $y_i = 1$). It is also easy to verify that this choice of (I,J) makes all the supports of the rank-one contributions pairwise disjoint, so that the assumptions of Theorem 3.3 are fulfilled. Finally, we observe that $|I_N| = |J_N| = 4$.

2) Suppose that our hypothesis is correct for $N-1$, we need to prove its correctness for N. Let (I_{N-1}, J_{N-1}) be the pair of supports for $N-1$, we construct (I_N, J_N) (still in binary matrix form) as follows:

$$I_N = \begin{pmatrix} \mathbf{1}_{n/2\times 1} & \mathbf{0}_{n/2\times 1} & I_{N-1} & \mathbf{0}_{n/2\times(3n/2-2)} \\ \mathbf{0}_{n/2\times 1} & \mathbf{1}_{n/2\times 1} & \mathbf{0}_{n/2\times(3n/2-2)} & I_{N-1} \end{pmatrix}$$
$$J_N = \begin{pmatrix} \mathbf{0}_{n/2\times 1} & \mathbf{1}_{n/2\times 1} & J_{N-1} & \mathbf{0}_{n/2\times(3n/2-2)} \\ \mathbf{1}_{n/2\times 1} & \mathbf{0}_{n/2\times 1} & \mathbf{0}_{n/2\times(3n/2-2)} & J_{N-1} \end{pmatrix}$$

where $n = 2^N$ and $\mathbf{1}_{p\times q}$ (resp. $\mathbf{0}_{p\times q}$) is the matrix of size $p \times q$ full of ones (resp. of zeros). Since I_{N-1} and J_{N-1} are both of dimension $2^{N-1} \times (3 \times 2^{N-1} - 2) = (n/2)(3n/2-2)$, the dimensions of I_N and J_N are both equal to $(n, 2\times(3n/2-2)+2) = (n, 3n-2)$. Moreover, the cardinalities of I_N and J_N satisfy the following recursive formula:

$$|I_N| = n + 2|I_{N-1}|, \qquad |J_N| = n + 2|J_{N-1}|,$$

which justifies the fact that $|I_N| = |J_N| = O(n \log n)$. Finally, any factors (X,Y) respecting the support constraints (I_N, J_N) need to have the following form:

$$X = \begin{pmatrix} X_1 & \mathbf{0}_{n/2\times 1} & X_3 & \mathbf{0}_{n/2\times(3n/2-2)} \\ \mathbf{0}_{n/2\times 1} & X_2 & \mathbf{0}_{n/2\times(3n/2-2)} & X_4 \end{pmatrix}$$
$$Y = \begin{pmatrix} \mathbf{0}_{n/2\times 1} & Y_2 & Y_3 & \mathbf{0}_{n/2\times(3n/2-2)} \\ Y_1 & \mathbf{0}_{n/2\times 1} & \mathbf{0}_{n/2\times(3n/2-2)} & Y_4 \end{pmatrix}$$

where $X_i, Y_i \in \mathbb{R}^{n/2}, 1 \le i \le 2$, and for $3 \le j \le 4$ we have $X_j, Y_j \in \mathbb{R}^{n/2\times(3n/2-2)}$, $\mathtt{supp}(X_j) \subseteq I_{N-1}, \mathtt{supp}(Y_j) \subseteq J_{N-1}$. Their product yields:

$$XY^\top = \begin{pmatrix} X_3Y_3^\top & X_1Y_1^\top \\ X_2Y_2^\top & X_4Y_4^\top \end{pmatrix}.$$

Given an HODLR matrix $A \in \mathbb{R}^{n\times n}$, since $A_{12}, A_{21} \in \mathbb{R}^{n/2\times n/2}$ are of rank at most one, one can find $X_i, Y_i \in \mathbb{R}^{n/2}, 1 \le i \le 2$ such that $A_{12} = X_1Y_1^\top, A_{21} = X_2Y_2^\top$.

Since $A_{11}, A_{22} \in \mathbb{R}^{n/2 \times n/2}$ are HODLR, by the induction hypothesis, one can also find $X_i, Y_i \in \mathbb{R}^{n/2 \times (3n/2-2)}$, $3 \leq i \leq 4$ such that $\mathtt{supp}(X_i) \subseteq I_{N-1}$, $\mathtt{supp}(Y_i) \subseteq I_{N-1}$ and $A_{11} = X_3 Y_3^\top, A_{22} = X_4 Y_4^\top$. Finally, this construction also makes all the supports of the rank-one contributions pairwise disjoint: the first two rank-one supports are $\mathcal{S}_1 = \{n/2+1, \ldots, n\} \times [\![n/2]\!], \mathcal{S}_2 = [\![n/2]\!] \times \{n/2+1, \ldots, n\}$, and the remaining ones are inside $[\![n/2]\!] \times [\![n/2]\!]$ and $\{n/2+1, \ldots, n\} \times \{n/2+1, \ldots, n\}$ which are disjoint by the induction hypothesis. $\square$