FAST SPARSE NONNEGATIVE MATRIX FACTORIZATION WITH MANIFOLD ACCELERATION

Juhao Bai[1], *Shixiang Chen*[1], *Shiqian Ma*[2]

[1]School of Mathematical Sciences, University of Science and Technology of China, Hefei, China
[2]Department of Computational Applied Mathematics and Operations Research,
Rice University, Houston, USA

ABSTRACT

In this paper, we propose a fast sparse Nonnegative Matrix Factorization algorithm incorporating manifold identification techniques. Within an alternating update framework, it adaptively leverages the algorithm's inherent manifold identification information to accelerate subproblem solutions, thereby enhancing computational efficiency. Numerical experiments demonstrate that our algorithm shows superior performance compared to existing methods, achieving better solutions with faster convergence rates, particularly under high sparsity requirements. We provide a global convergence guarantee for the algorithm. Regarding the locally linear convergence observed experimentally, under a set of assumptions, we develop a proof strategy for general cases. Furthermore, we furnish a complete proof for the vector case.

***Index Terms*—** Sparse NMF, Manifold identification, Kurdyka-Losiajewicz property, Morse-Bott condition

1. INTRODUCTION

Nonnegative Matrix Factorization (NMF) is a popular decomposition tool and has a wide range of applications, e.g., clustering [1, 2], face recognition [3, 4, 5], signal processing [6], etc. To enhance the interpretability and sparsity of the factorized components, sparse NMF variants have also been extensively studied [7, 8]. In a seminal work, Lee and Seung [9] demonstrated how nonnegative matrix factorization automatically learns parts-based decompositions of faces, allowing each learned component (e.g., eyes, nose, mouth) to be interpreted as a physical "part" rather than an abstract algebraic factor.

Specifically, standard NMF aims to solve the following problem:

$$\min_{X\in\mathbb{R}_+^{n\times r},\ Y\in\mathbb{R}_+^{r\times m}} F(X,Y) = \frac{1}{2}\|A - XY\|_F^2, \tag{1}$$

where $\mathbb{R}_+^{n\times r}$ denotes the nonnegative orthant of $\mathbb{R}^{n\times r}$, i.e., the set of all real $n \times r$ matrices with nonnegative entries. Although the multiplication update algorithm in [10] has been one of most commonly used for NMF, some issues related to its perfoormance [11, 12, 13] and problems with convergence were reported. In recent years, several algorithms adopt an alternating nonnegative least squares (ANLS) framework [11, 12, 13] to fully leverage the problem's structure, which were introduced with good performance. The alternating nonnegative least squares is to optimize X and Y by alternatingly solving the following nonnegative least squares problems:

$$X^{k+1} = \operatorname{argmin}_{X\in\mathbb{R}_+^{n\times m}} \frac{1}{2}\|A - XY^k\|_F^2, \tag{2}$$

$$Y^{k+1} = \operatorname{argmin}_{Y\in\mathbb{R}_+^{r\times m}} \frac{1}{2}\|A - X^{k+1}Y\|_F^2. \tag{3}$$

These algorithms exhibit favorable convergence properties because every limit point produced by the ANLS framework is a stationary point [11]. Moreover, given that sparsity is prevalent in numerous practical problems—such as signal processing and text clustering—designing algorithms capable of generating sparse solutions has indeed demonstrated superior performance in real-world applications. A prominent contribution in this area is the work by Bolte et al [14], who proposed the Proximal Alternating Linearized Minimization method (PALM). This algorithm is capable of solving a broad class of nonsmooth nonconvex problems, including NMF with sparsity constraints. Due to its favorable performance, PALM has been extensively employed in practical applications. More recently, another notable work by Teboulle and Vaisbourd [15] introduced novel algorithms such as Co-HALS and Co-MU. By incorporating Bregman distances, their approach circumvents the requirement for gradient Lipschitz continuity and exhibits excellent experimental performance.

On the other hand, sparsity, much like non-negativity, is an intrinsic property of many real-world datasets. In the context of NMF, researchers often incorporate regularization terms to obtain sparse solutions, which in practice demonstrate superior empirical performance. Specifically, the sparse NMF problem can be formulated as follows:

$$\min_{X\in\mathbb{R}_+^{n\times r},\ Y\in\mathbb{R}_+^{r\times m}} F(X,Y) \triangleq \frac{1}{2}\|A - XY\|_F^2 + R_1(X) + R_2(Y). \tag{4}$$

where $R_1(X)$ and $R_2(Y)$ are regularization terms designed to induce sparsity in the factors X and Y, respectively. In [7], sparsity is achieved by incorporating the ℓ_1 regularization term, which is known to produce a sparse representation [16].

Regarding the use of sparsity as a crucial property, a significant piece of work is [17], wherein the sparsity induced by regularization, as described above, is extended to the so-called manifold identification property. By leveraging manifold identification, the original nonsmooth optimization problem is restricted to a smooth manifold, which not only enables the use of more efficient Riemannian optimization techniques but also significantly reduces the effective dimension of the problem, thereby accelerating convergence. It intertwine Riemannian Newton-like methods with proximal gradient steps to drastically boost the convergence and prove the algorithm's superlinear convergence when solving nondegenerate nonsmooth nonconvex optimization problems in [17]. By incorporating suitable regularization subproblems, the aforementioned method can

Corresponding author: Shixiang Chen (shxchen@ustc.edu.cn)

be employed to solve subproblem (2) and (3), which constitutes our primary approach.

This paper proposes a novel manifold-accelerated algorithm for sparse nonnegative matrix factorization (SNMF) that adaptively exploits underlying manifold structures within an alternating proximal minimization framework. We establish global convergence to a stationary point under standard assumptions and. To the best of our knowledge, we provide the first theoretical analysis of local linear convergence around the global optimal solution under the restricted case where the inner dimension r equals the nonnegative rank of A [18]. By verifying the Morse–Bott property [19], we show that the Kurdyka–Lojasiewicz exponent is $\frac{1}{2}$ for the case $r = 1$. Empirical results on real-world datasets demonstrate the superiority of our method over existing approaches in convergence speed and solution accuracy, particularly under high sparsity. The proposed framework is generalizable to a broad class of structured nonsmooth nonconvex optimization problems.

Notation. Unless otherwise specified, the matrix inner product is defined as the trace inner product $\langle U, V\rangle = \mathrm{tr}(U^\top V)$. We use $\|\cdot\|$ to denote the ℓ_2-norm (for vectors), $\|\cdot\|_F$ for the Frobenius norm (for matrices), $\|\cdot\|_0$ for the entrywise ℓ_0-"norm" (counting the number of nonzero entries), and $\|\cdot\|_1$ for the entrywise ℓ_1-norm (sum of absolute values). For a set C, we use $\delta_C(x)$ to denote the indicator function, where $\delta_C(x) = 0$ if $x \in C$ and $\delta_C(x) = \infty$ otherwise. We use $\mathbb{R}_+^{p\times q}$ to denote nonnegative matrix in $\mathbb{R}^{p\times q}$, and $[n] = \{1, \cdots, n\}$. For a function g, we use ∂g to denote the subdifferential. For a matrix M, we use X_{ij} to denote its element in the i-th row and j-th column. For a real symmetric matrix $M \in \mathbb{R}^{n\times n}$, we use $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ to denote its minimum eigenvalue and maximum eigenvalue.

2. PROPOSED ALGORITHM

As previously stated, this paper aims to solve the sparse NMF problem in the following form:

$$\min_{X,Y} F(X,Y) \triangleq \frac{1}{2}\|A - XY\|_F^2 + R_1(X) + R_2(Y). \tag{5}$$

where $R_1(X)$ and $R_2(Y)$ are regularization terms designed to induce sparsity in the factors $X \in \mathbb{R}^{n\times r}$ and $Y \in \mathbb{R}^{r\times m}$, respectively. For simplicity, we incorporate the non-negative constraints into the regularization terms R_1 and R_2 via indicator functions $\delta_{\mathbb{R}_+^{n\times r}}(X)$ and $\delta_{\mathbb{R}_+^{r\times m}}(Y)$. To promote structured sparsity and nonnegativity, we consider regularizers R_1 and R_2 that act column-wise on X and row-wise on Y, respectively. A widely used convex choice is the ℓ_1-based regularizer: $R_1(X) = \lambda_1\|X\|_1 + \delta_{\mathbb{R}_+^{n\times r}}(X)$, while structured ℓ_0-based constraints are also common [15, 14]: $R_1(X) = \sum_{i=1}^r \left[\delta_{\mathbb{R}_+^n}(x_i) + \delta_{\{\|x_i\|_0\le\alpha_i\}}\right]$. We unify these formulations by defining

$$R_1(X) := \sum_{i=1}^{r}\left[\delta_{\mathbb{R}_+^n}(x_i) + \rho_i(x_i)\right],$$

where $\rho_i : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ is selected as

$$\rho_i(x) = \begin{cases} \delta_{\{\|x\|_0\le\alpha_i\}}, & \text{(hard } \ell_0 \text{ constraint)} \\ \lambda_i\|x\|_0, & (\ell_0 \text{ regularization}) \\ \delta_{\{\|x\|_1\le\tau_i\}}, & \text{(hard } \ell_1 \text{ constraint)} \\ \lambda_i\|x\|_1, & (\ell_1 \text{ regularization}) \end{cases} \tag{6}$$

depending on the model.

We propose a Proximal Alternating Linearized Minimization Algorithm with Newton acceleration (PALM-NA) to solve problem (5). The detailed procedure is summarized in Algorithm 1. We employ an alternating update scheme to update variables X and Y. For instance, in Step 5 of PALM-NA, we get W^{k+1} by performing a proximal gradient step:

$$\mathrm{prox}_{c_k}^{R_1}(U^k) = \operatorname*{argmin}_X \left\{\frac{c_k}{2}\|X - U^k\|^2 + R_1(X)\right\}$$

Step 5 provides both the current point W^{k+1} and the manifold $\mathcal{M}_{W^{k+1}}$ where it lies. Specifically, the manifold $\mathcal{M}_{W^{k+1}}$ is defined as follows:

$$\mathcal{M}_{W^{k+1}} \triangleq \{X \in \mathbb{R}_+^{n\times r} : S_X = S_{W^{k+1}}\} \tag{7}$$

where S_X and $S_{W^{k+1}}$ denotes the support set of X and W^{k+1}. Next, in Step 6, $\mathrm{ManAcc}_{\mathcal{M}_{W^{k+1}}}$ denotes a second-order optimization step on $\mathcal{M}_{W^{k+1}}$. On such a smooth manifold, the problem dimension is significantly reduced by restricting attention to the nonzero components. This enables the use of more advanced methods and leads to substantially faster local convergence. In this work, we employ the Riemannian trust-region method [20] as $\mathrm{ManAcc}_{\mathcal{M}_{W^{k+1}}}$. The updates to Y follow a similar pattern. Refer to Appendix 7 for $\mathrm{ManAcc}_{\mathcal{M}_{W^{k+1}}}$ implementation details.

Algorithm 1 PALM with Newton Acceleration (PALM-NA)

1: Input: Data matrix A, initialized random starting point (X^0, Y^0), parameters $\gamma_1, \gamma_2 > 1$
2: **while** True **do**
3: Let $c_k = \gamma_1 \left\|Y^k(Y^k)^T\right\|_F$
4: Compute $U^k = X^k - \frac{1}{c_k}\left(X^kY^k - A\right)(Y^k)^T$
5: Update $W^{k+1} \in \mathrm{prox}_{c_k}^{R_1}(U^k)$
6: Update $X^{k+1} = \mathrm{ManAcc}_{\mathcal{M}_{W^{k+1}}}(W^{k+1})$
7: Let $d_k = \gamma_2 \left\|X^{k+1}(X^{k+1})^T\right\|_F$
8: Compute $V^k = Y^k - \frac{1}{d_k}(X^{k+1})^T\left(X^{k+1}Y^k - A\right)$
9: Update $H^{k+1} \in \mathrm{prox}_{c_k}^{R_2}(V^k)$
10: Update $Y^{k+1} = \mathrm{ManAcc}_{\mathcal{M}_{H^{k+1}}}(H^{k+1})$
11: **if** convergence criterion is satisfied **then**
12: break
13: **end if**
14: **end while**

3. THEORETICAL RESULTS

This section establishes the convergence of PALM-NA. All the proofs can be found in Appendix 7. We first show global convergence to a stationary point. To explain the empirically observed local linear convergence in certain special cases—previously lacking theoretical justification—we introduce a unified analytical framework, and provide a complete proof of local linear convergence in the vector case ($r = 1$). In advance of the theoretical analysis, we first present the following definitions and assumptions.

Definition 1 (KL inequality)**.** *The function G is said to have the Kurdyka–Lojasiewicz (KL) property at $\bar{x} \in$ dom ∂G with an exponent of α, if there exist $c > 0$, a neighborhood U of $\bar{x}$ such that for all $x \in U$, inequality holds*

$$\mathit{dist}(0, \partial G(x)) \ge c(G(x) - G(\bar{x}))^\alpha. \tag{8}$$

The Kurdyka–Lojasiewicz inequality generalizes the original Lojasiewicz inequality [22] to a broader class of functions[1]. A crucial special case is the Polyak-Lojasiewicz (PL) inequality [24], which is recovered when the exponent $\alpha = \frac{1}{2}$ and serves as a standard tool for analyzing convergence rates. Regarding the KL exponent $\alpha \in [0, 1)$, different values of α imply that the algorithm exhibits varying rates of convergence [25].

Assumption 1 (Full rank)**.** *We assume that in the generated sequence $\{(X^k, Y^k)\}_k$, X^k has full column rank and Y^k has full row rank. Furthermore, $\forall k, \exists \mu_i^+ > 0, \mu_i^- > 0 (i = 1, 2)$, s.t.*

$$\begin{aligned}
&\inf\{\lambda_{\min}((X^k)^T X^k) : k \in \mathbb{N}\} \geq \mu_2^-, \\
&\inf\{\lambda_{\min}(Y^k (Y^k)^T) : k \in \mathbb{N}\} \geq \mu_1^-, \\
&\sup\{\lambda_{\max}((X^k)^T X^k) : k \in \mathbb{N}\} \leq \mu_2^+, \\
&\sup\{\lambda_{\max}(Y^k (Y^k)^T) : k \in \mathbb{N}\} \leq \mu_1^+.
\end{aligned}$$

This assumption is mild and necessary, and Assumption 1 is likely to happen when r is chosen to be much smaller than $\min(m, n)$. Moreover, it is very common in other NMF algorithms [26, 27] and empirical studies in [27] show this assumption is satisfied in practice.

3.1. Global Convergence

Theorem 1. *Let $\{(X^k, Y^k)\}_k$ be the sequence generated by the PALM-NA algorithm to minimize the objective function F. Assume that this sequence satisfies Assumption 1, and that F possesses the KL property. Then the following conclusions hold:*

(i) The sequence $\{Z^k\}_{k\in\mathbb{N}} = \{(X^k, Y^k)\}_{k\in\mathbb{N}}$ has finite length, i.e., $\sum_{k=1}^{\infty} \|Z^{k+1} - Z^k\|_F < \infty$.

(ii) The sequence $\{Z^k\}_{k\in\mathbb{N}}$ converges to a stationary point $Z^* = (X^*, Y^*)$ of F.

While Theorem 1 establishes global convergence under Assumption 1 and KL property, we note that these assumptions are mild and commonly adopted in the analysis of nonconvex optimization methods. The full rank assumption on iterates, though technically necessary, is often satisfied in practice with proper initialization and regularization. And the KL property of NMF problem is also discussed in [14]. Although Theorem 1 guarantees convergence to a stationary point, it does not characterize the rate of convergence. In the following subsection, we address this gap by analyzing the local linear convergence behavior for $r = 1$ under additional structural assumptions, supported by theoretical analysis.

3.2. Local convergence

Before proceeding to the theoretical analysis, we first present the following definitions.

Definition 2 (Morse-Bott condition)**.** *Let $\Phi(x) : \mathcal{M} \to \mathbb{R}$ is least C^1, where $\mathcal{M}$ is a Riemannian manifold, and $\bar{x}$ be a local minimum of Φ. We say Φ satisfies the Morse–Bott property at $\bar{x}$ if $\mathcal{S}$ is a C^1 submanifold around $\bar{x}$ and $\ker \nabla^2 \Phi(\bar{x}) = T_{\bar{x}}\mathcal{S}$, where $\mathcal{S}$ is a set of local optimal solution*

$$\mathcal{S} \triangleq \{x \in \mathcal{M} : x \text{ is a local minimum of } \Phi \text{ and } \Phi(\bar{x}) = \Phi_{\mathcal{S}}\}. \quad (9)$$

Unlike PL inequality, quadratic growth condition, etc., which are widely used in optimization to study gradient flows, Morse-Bott (MB) condition has received relatively little attention. Early work by Shapiro [28] analyzes perturbations of optimization problems assuming a property similar to MB. There is also a mention of gradient flow under MB in [29]. From the existing results, we know that when the function f is C^2, the PL condition is equivalent to the above MB property [19].

[1]The original Lojasiewicz inequality was introduced in [23]

Definition 3 (Nonnegative rank)**.** *For a given nongative matrix $M \in \mathbb{R}^{p\times q}$, The nonnegative rank of M is equal to the smallest number s such there exists a nonnegative $p \times s$ matrix P and a nonnegative $s \times q$ matrix C such that $M = PQ$.*

The concept of nonnegative rank is referred to [18]. In this work, the nonnegative rank condition is leveraged to simplify the structure of the local solution set, thereby facilitating the analysis of the solution set form in certain special cases.

We begin by outlining our proof strategy as follows: Our primary goal is to establish the local linear convergence of the sequence generated by the PALM-NA when solving the ℓ_0/ℓ_1-regularized NMF problem (5). As indicated in reference [22], proving linear convergence is equivalent to showing that the objective function F has a KL exponent of $\frac{1}{2}$ at the optimal point. Since F is nonconvex and nonsmooth, a direct proof is exceedingly difficult. On the other hand, reference [19] points out that for at least $\mathcal{C}^2$ functions Φ, the Polyak-Lojasiewicz (PL) condition—which applies in the smooth case—implies a KL exponent of $\frac{1}{2}$, and this is equivalent to the Morse-Bott (MB) condition. Furthermore, applying the manifold identification results from reference [30], we can verify that the sequence generated by the PALM-NA possesses the finite identification property. This allows us to define a manifold $\mathcal{M}_*$ induced by the limit point (X^*, Y^*). After a finite number of iterations, the algorithm effectively reduces to optimizing over the manifold $\mathcal{M}_*$. This inspires us to define the restriction of F to the manifold $\mathcal{M}_*$ denoted by $F|_{\mathcal{M}_*}$ which is $\mathcal{C}^\infty$-smooth on $\mathcal{M}_*$, where $\mathcal{M}_* \triangleq \{(X \in \mathbb{R}_+^{n\times r}, Y \in \mathbb{R}_+^{r\times m}) : S_X = S_{X^*}, S_Y = S_{Y^*}\}$. The problem thus reduces to two aspects: first, verifying that $F|_{\mathcal{M}_*}$ satisfies the MB condition at (X^*, Y^*), and second, establishing the connection between the local behavior of $F|_{\mathcal{M}_*}$ and that of F near the optimal point. The latter is guaranteed under Assumption 2, a mild condition that has been employed and empirically studied in [27], and is often satisfied in practice.

Assumption 2 (Strict complementary slackness)**.** *For limit point (X^*, Y^*) of the iterative sequence $\{(X^k, Y^k)\}_k$ generated by PALM-NA, we assume that if $X_{ij}^* = 0$, then $(\frac{\partial f(X^*,Y^*)}{\partial X})_{ij} > 0$, where $f(X, Y) = \frac{1}{2}\|A - XY\|_F^2$, and analogous assumption applies to the variable Y.*

And the former is further analyzed below, we need to verify three conditions for the MB property: (i) The function F is constant on the local solution set $\mathcal{S}$. (ii) $\mathcal{S}$ is a smooth embedded submanifold of $\mathcal{M}_*$. (iii) The Riemannian Hessian of F is zero along the tangent space $T_{(X,Y)}\mathcal{S}$, and positive definite on the normal space $N_{(X,Y)}\mathcal{S}$. Due to the intricate manifold structure of the solution set caused by varying distributions of zero entries, we postulate condition (ii) as the following assumption.

Assumption 3. *Let $\mathcal{S}$ denote the local solution set of $F|_{\mathcal{M}_*}$. We assume that $\mathcal{S}$ is an embedded submanifold of $\mathcal{M}$ around (X^*, Y^*).*

Assumption 3 is relatively strong in general. However, for the special case $r = 1$, we can rigorously show that it indeed forms a one-dimensional smooth submanifold of $\mathcal{M}_*$; see the full version [21] for details. For the general case $r > 1$, the local solution set may exhibit diverse structures due to uncertainty in the zero-entry patterns of the optimal solution. Nevertheless, it is sufficient for the submanifold property to hold locally, which makes the assumption more practically plausible.

To verify condition (iii) of the Morse–Bott property, it is necessary to understand the structure of the local solutions, which constitutes one of the primary challenges in the analysis. We begin by considering a restricted case: the scenario of exact factorization, i.e., $XY = A$. This naturally leads us to introduce the concept of the nonnegative rank of a nonnegative matrix M. Moreover, to account for the sparse constraint present in our problem, we incorporate it explicitly as the following assumption.

Assumption 4. *For* $R_1(X) = \sum_{i=1}^{r}\left[\delta_{\mathbb{R}_+^n}(x_i) + \rho_i(x_i)\right]$, *and* $R_2(Y) = \sum_{i=1}^{r}\left[\delta_{\mathbb{R}_+^m}(y_i) + \rho_i(y_i)\right]$, *we assume that the inner dimension* r *equals the nonnegative rank of matrix* A. *Furthermore, even under the* ℓ_0 *constraint, the global optimal solution* (X^*, Y^*) *still satisfies the exact factorization condition* $X^*Y^* = A$.

This assumption is practically reasonable. Setting the inner dimension to the nonnegative rank prevents loss of information, while requiring the global optimum to remain an exact factorization ensures that sparsity does not compromise the essential features of the data.

Theorem 2. *For the iterative sequence* $\{(X^k, Y^k)\}_k$ *generated by PALM-NA with* ℓ_0*-norm constraint, let* (X^*, Y^*) *be its limit point. Assume that* $0 \in \operatorname{rint} \partial F(X^*, Y^*)$, *then the manifold* $\mathcal{M}_{(X^*,Y^*)}$ *will be identified within finitely many iterations, i.e., there exists* $K \in \mathbb{N}$ *such that for all* $k \geq K$, *we have:*

$$\mathcal{M}_{(X^k,Y^k)} = \mathcal{M}_{(X^*,Y^*)} = \mathcal{M}_*. \tag{10}$$

In Theorem 2, the relative interior point assumption, i.e., $0 \in \operatorname{rint} \partial F(X^*, Y^*)$, is a common assumption [17], and Theorem 2 serves as a crucial guarantee for smoothing the original nonsmooth problem.

Theorem 3. *Let* $\{(X^k, Y^k)\}_k$ *be the sequence generated by PALM-NA for minimizing* F *with an* ℓ_0*-norm constraint, and suppose it converges to a global optimum* (X^*, Y^*). *If* $0 \in \operatorname{rint} \partial F(X^*, Y^*)$ *and Assumptions 2, 3, and 4 hold, then* $F|_{\mathcal{M}_*}$ *satisfies the Morse–Bott property at* (X^*, Y^*), *and the sequence* $\{(X^k, Y^k)\}_k$ *converges locally linearly.*

Theorem 3 states our main result on local linear convergence. While Assumption 3 is relatively strong, we have verified its validity in the simple case $r = 1$.

4. NUMERICAL EXPERIMENTS

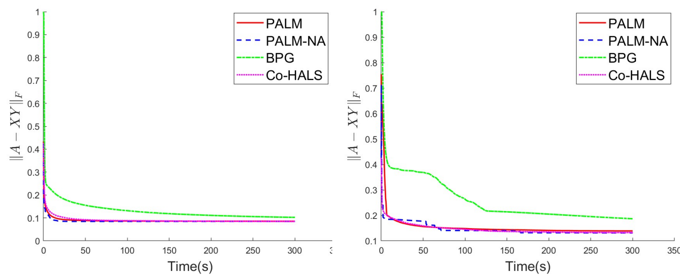
We have used a real dataset in order to compare the methods. Here, we report the results for Center for Biological and Computational Learning (CBCL) dataset [9]. The CBCL dataset contains 2429 images. The size of each image is 19×19 pixels. Thus, the resulting data matrix A is of size 361×2429. We have used $r = 49$ in our experimental setting. Each entry in the initial matrices was generated uniformly in the interval $[0, 1]$. We have also adopted some simple yet effective initialisation strategies that will help us improve the quality of the solutions [15]. The data matrix was normalized such that $\|A\|_F = 1$. And by applying a rescaling procedure we ensured that the initial matrices X^0 and Y^0 satisfy the following two conditions: (i) $\|x_i^0\| = \|y_i^0\|$ for all $i \in [r]$, (ii) $\operatorname{argmin}_t \|A - tX^0Y^0\| = 1$, where x_i^0 denotes the i-th column of X^0, and y_i^0 denotes the i-th row of Y^0.

In this subsection, we consider setting the regularization terms to the ℓ_0-norm. And we set $R_1(X) = \sum_{i=1}^{r}[\delta_{\mathbb{R}_+^m}(x_i) + \delta_{\mathbb{B}_0^{\alpha_i}}(x_i)]$ and $R_2(Y) = 0$. We set sparsity parameter $\alpha_i = \{0.2, 0.05\}$, and the corresponding Hessian matrix sizes in ManAcc are $\{3528 \times 3528, 882 \times 882\}$. In both PALM and PALM-NA, the step-size scaling factors were set to $\gamma_1 = \gamma_2 = 1.001$, and for Co-HALS, the smoothing parameter was chosen as $\epsilon = 10^{-6}$. We execute 10 runs for any method we examine, each time initializing it with a different point, and report average results. For a fair comparison we use the same initial points for all methods. We compare our algorithm with existing algorithms, namely BPG [15], PALM [14], and Co-HALS [15], which all use alternating updates to solve the sparse NMF problem (5).

In Tables 1 and 2, we present the corresponding values sampled at 15, 30, 60, and 300 seconds for sparsity levels of 0.8 and 0.95, respectively. From table 1, PALM-NA demonstrates superior performance, achieving the best value(0.0851) in just 30 seconds and maintaining it consistently. This indicates fast convergence and high stability. Both PALM and Co-HALS converge more slowly but nearly match PALM-NA's result by 300 seconds. BPG performs significantly worse than all other methods at every time interval.

	15s	30s	60s	300s
BPG	0.2065	0.1777	0.1500	0.1019
PALM	0.0974	0.0900	0.0868	0.0851
Co-HALS	0.1127	0.0995	0.0913	0.0852
PALM-NA	0.0888	**0.0851**	**0.0851**	**0.0851**

Table 1. sparsity parameter $\alpha_i = 0.2, \dim(\mathcal{M}^*) = 0.2(nr, mr)$.



(1a) $\alpha_i = 0.2$ (1b) $\alpha_i = 0.05$

In Figure 1 we present the average Frobenius norm of the residual matrix $A - XY$ for the sparse model with sparsity 80% and 95%. It is worth noting that in Figure 1(b), a sharp drop in the curve of Algorithm 1 occurs around 50 seconds. This can be attributed to the manifold acceleration used for the subproblem with one factor fixed: such acceleration becomes notably effective only when the current iterate reaches a favorable position. It is conjectured that local minima are more likely to exhibit such characteristics. This phenomenon is also reflected in the results of Table 2, where Algorithm 1 demonstrates a clear advantage at the 60-second mark, further supporting the observation that the algorithm achieves significant performance improvement once it enters a favorable region.

	15s	30s	60s	300s
BPG	0.3807	0.3743	0.3366	0.1848
PALM	0.1836	0.1654	0.1519	0.1350
Co-HALS	0.1895	0.1758	0.1615	0.1320
PALM-NA	0.1785	0.1704	**0.1584**	**0.1314**

Table 2. sparsity parameter $\alpha_i = 0.05, \dim(\mathcal{M}^*) = 0.05(nr, mr)$.

5. CONCLUSION

This paper presents a fast sparse Nonnegative Matrix Factorization algorithm that integrates manifold identification within an alternating minimization framework. The approach adaptively accelerates subproblem solving using second-order information on the manifold. Global convergence to a stationary point is guaranteed, and local linear convergence is established for the case where the inner dimension equals the nonnegative rank. Evaluations on the CBCL dataset show superior convergence speed and accuracy under high sparsity compared to existing methods.

6. REFERENCES

[1] X. Xu, W. Liu and Y. Gong, "Document clustering based on non-negative matrix factorization," 2003, SIGIR '03, p. 267–273.

[2] Shuai Wang, Tsung-Hui Chang, Ying Cui, and Jong-Shi Pang, "Clustering by orthogonal non-negative matrix factorization: A sequential non-convex penalty approach," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5576–5580.

[3] S.Z. Li, Xin Wen Hou, Hong Jiang Zhang, and Qian Sheng Cheng, "Learning spatially localized, parts-based representation," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 2001, vol. 1, pp. I–I.

[4] David Guillamet and Jordi Vitrià, "Non-negative matrix factorization for face recognition," in *Topics in Artificial Intelligence*, M. Teresa Escrig, Francisco Toledo, and Elisabet Golobardes, Eds., Berlin, Heidelberg, 2002, pp. 336–344, Springer Berlin Heidelberg.

[5] Jiwen Lu and Yap-Peng Tan, "Doubly weighted nonnegative matrix factorization for imbalanced face recognition," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 877–880.

[6] Jonathan Le Roux, John R. Hershey, and Felix Weninger, "Deep nmf for speech separation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 66–70.

[7] Patrik O. Hoyer, "Non-negative sparse coding," *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, pp. 557–565, 2002.

[8] Thomas Guthier, Adrian Šošić, Volker Willert, and Julian Eggert, "snn-lds: Spatio-temporal non-negative sparse coding for human action recognition," in *Artificial Neural Networks and Machine Learning – ICANN 2014*, Stefan Wermter, Cornelius Weber, Włodzisław Duch, Timo Honkela, Petia Koprinkova-Hristova, Sven Magg, Günther Palm, and Alessandro E. P. Villa, Eds., Cham, 2014, pp. 185–192, Springer International Publishing.

[9] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[10] Daniel Lee and H Sebastian Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, 2000.

[11] Chih-Jen Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Computation*, vol. 19, pp. 2756–2779, 2007.

[12] Dongmin Kim, Suvrit Sra, and Inderjit S. Dhillon, *Fast Newton-type Methods for the Least Squares Nonnegative Matrix Approximation Problem*, pp. 343–354.

[13] Hyunsoo Kim and Haesun Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 2, pp. 713–730, 2008.

[14] Jérôme Bolte, Shoham Sabach, and Marc Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1, pp. 459–494, 2014.

[15] Marc Teboulle and Yakov Vaisbourd, "Novel proximal gradient methods for nonnegative matrix factorization with sparsity constraints," *SIAM Journal on Imaging Sciences*, vol. 13, no. 1, pp. 381–421, 2020.

[16] David L. Donoho and Michael Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 2197 – 2202, 2003.

[17] Gilles Bareilles, Franck Iutzeler, and Jérôme Malick, "Newton acceleration on manifolds identified by proximal gradient methods," *Mathematical Programming*, vol. 200, no. 1, pp. 37–70, 2023.

[18] Joel E. Cohen and Uriel G. Rothblum, "Nonnegative ranks, decompositions, and factorizations of nonnegative matrices," *Linear Algebra and its Applications*, vol. 190, pp. 149–168, 1993.

[19] Quentin Rebjock and Nicolas Boumal, "Fast convergence to non-isolated minima: four equivalent conditions for C^2 functions," *Mathematical Programming*, vol. 213, no. 1, pp. 151–199, 2025.

[20] P.-A. Absil, C.G. Baker, and K.A. Gallivan, "Trust-region methods on riemannian manifolds," *Foundations of Computational Mathematics*, vol. 7, no. 3, pp. 303–330, 2007.

[21] JuHao Bai ShiXiang Chen and Shiqian Ma, "Fast sparse nonnegative matrix factorization with manifold acceleration," `https://github.com/sjubg/SNMF`.

[22] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-lojasiewicz inequality," *Math. Oper. Res.*, vol. 35, pp. 438–457, 2008.

[23] Stanisław Lojasiewicz, "Ensembles semi-analytiques," 1965.

[24] Hamed Karimi, Julie Nutini, and Mark W. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-Lojasiewicz condition," in *ECML/PKDD*, 2016.

[25] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis, "The Lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 1205–1223, 2007.

[26] Hyunsoo Kim and Haesun Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM J. Matrix Anal. Appl.*, vol. 30, pp. 713–730, 2008.

[27] Pinghua Gong and Changshui Zhang, "Efficient nonnegative matrix factorization via projected newton method," *Pattern Recognition*, vol. 45, no. 9, pp. 3557–3565, 2012, Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011).

[28] Alexander Shapiro, "Perturbation theory of nonlinear programs when the set of optimal solutions is not a singleton," *Applied Mathematics and Optimization*, vol. 18, pp. 215–229, 1988.

[29] Uwe Helmke and John B.Moore, "Optimization and dynamical systems," 2012, Prop. 12.3.

[30] Warren Hare and Adrian Lewis, "Identifying active constraints via partial smoothness and prox-regularity," *J. Convex Anal.*, vol. 11, pp. 251–266, 01 2004.

[31] Trond Steihaug, "The conjugate gradient method and trust regions in large scale optimization," *SIAM Journal on Numerical Analysis*, vol. 20, no. 3, pp. 626–637, 1983.

[32] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis, "A nonsmooth morse–sard theorem for subanalytic functions," *Journal of Mathematical Analysis and Applications*, vol. 321, no. 2, pp. 729–740, 2006.

7. APPENDIX

7.1. Global Convergence

Here, we provide a brief explanation of some notations that will be used later. First, for the convergence analysis of our algorithm, we consider the sequence of all iterates generated during the iterations, denoted by ${Z^k}_{k\in\mathbb{N}}$. Specifically, in our algorithm, for any $k \in \mathbb{N}$, the update sequence is as follows:

$$(X^k, Y^k) \xrightarrow{PGD} (W^k, Y^k) \xrightarrow{RTR} (X^{k+1}, Y^k) \xrightarrow{PGD} (X^{k+1}, H^k) \xrightarrow{RTR} (X^{k+1}, Y^{k+1})$$

For instance, for any $k \in \mathbb{N}$, $Z^{4k+1} = (W^k, Y^k)$ and $Z^{4k+3} = (X^{k+1}, Y^k)$. Additionally, we denote $Z_X \in \mathbb{R}^{n\times r}$ and $Z_Y \in \mathbb{R}^{r\times m}$ as the projections of Z onto the X and Y components, respectively, i.e., $Z = (Z_X, Z_Y)$.
Here, we prove that our algorithm is guaranteed to converge to a first-order stationary point. Before proceeding with the proof, we state the following proposition concerning the trust-region update step, which will be frequently used in the subsequent analysis.
First,the objective function $F(X, Y)$ and its smooth component $f(X, Y)$ satisfy the following properties:

Proposition 1. *(a) $F(X, Y)$ is bounded below:*

$$\inf_{\mathbb{R}^{n\times r}\times\mathbb{R}^{r\times m}} F(X, Y) > -\infty. \tag{11}$$

(b) The smooth term $f(X, Y) = \frac{1}{2}\|A - XY\|_F^2$ has block-wise Lipschitz continuous gradients. Specifically, for any fixed Y, the partial gradient $\nabla_X f(X, Y)$ is Lipschitz continuous with respect to X:

$$\|\nabla_X f(X_1, Y) - \nabla_X f(X_2, Y)\| \le L_1(Y)\|X_1 - X_2\|, \quad \forall X_1, X_2 \in \mathbb{R}^{n\times r}.$$

Similarly, for any fixed X, $\nabla_Y f(X, Y)$ is Lipschitz continuous with respect to Y.

(c) The Lipschitz constants $L_1(Y)$ and $L_2(X)$ are uniformly bounded. That is, there exist positive constants $\theta_1^\pm$ and $\theta_2^\pm$ such that for all $k \in \mathbb{N}$:

$$\theta_1^- \le L_1(Z_Y^k) \le \theta_1^+,$$
$$\theta_2^- \le L_2(Z_X^k) \le \theta_2^+.$$

(d) The full gradient $\nabla f(X, Y) = (\nabla_X f(X, Y), \nabla_Y f(X, Y))$ is jointly Lipschitz continuous on any bounded set $B \subset \mathbb{R}^{n\times r} \times \mathbb{R}^{r\times m}$. That is, there exists a constant $M > 0$ such that for all $(X_1, Y_1), (X_2, Y_2) \in B$:

$$\|\nabla f(X_1, Y_1) - \nabla f(X_2, Y_2)\| \le M \|(X_1 - X_2, Y_1 - Y_2)\|.$$

Proposition 2. *For the trust-region step update, when the step is accepted, the update direction is found at the trust-region boundary only finitely many times.*

Proof. In the trust-region update, we consider the case where the direction is accepted. Note that we only take one trust-region step, and the reduction in the quadratic model satisfies the following inequality[31]:

$$m_0(0) - m_0(\eta) \ge \frac{1}{2}\|\operatorname{grad} f(x_0)\| \min\left(\Delta_0, \frac{\|\operatorname{grad} f(x_0)\|}{\|H_0\|}\right) \tag{12}$$

If the update direction is found at the trust-region boundary, then the $\min$ term in the above inequality achieves Δ_0. Combining inequality (12) with the acceptance condition of the trust-region algorithm, we have:

$$F(Z^k) - F(Z^{k+1}) \ge \frac{1}{2}\rho'\Delta_0^2\|H_0\| \tag{13}$$

Here, $\rho' > 0$ is the threshold for accepting the update, Δ_0 is the trust-region radius, and H_0 denotes the Riemannian Hessian at the current iterate. The right-hand side is a positive constant. If there were infinitely many steps where the update is found at the boundary, it would contradict Proposition 1. Therefore, after a sufficiently large number of iterations, whenever an update direction is accepted, the trust-region update will remain strictly inside the trust region. ☐

Proposition 3. *Under Assumption 1, in the trust-region update, if the solution obtained by the tCG method lies strictly inside the trust region, then we have the following estimate:*

$$\frac{1-\kappa}{\max(\theta_1^+, \theta_2^+)}\|g\| \le \frac{1-\kappa}{\lambda_{\max}(H_0)}\|g\| \le \|p\| \le \frac{1+\kappa}{\lambda_{\min}(H_0)}\|g\| \le \frac{1+\kappa}{\min(\theta_1^-, \theta_2^-)}\|g\| \tag{14}$$

Here, $0 < \kappa < 1$ is a given parameter in the tCG algorithm, and p, g, and H_0 denote the update direction computed by tCG, the Riemannian gradient, and the Riemannian Hessian at the current trust-region subproblem, respectively.

Proof. Under our algorithmic settings, if the obtained update direction lies within the trust region, the tCG algorithm ensures that the following inequality holds:

$$\|H_0 p + g\| \leq \kappa \|g\| \tag{15}$$

This implies that:

$$\|H_0 p\| - \|g\| \leq \|H_0 p + g\| \leq \kappa \|g\| \quad \text{and} \quad \|g\| - \|H_0 p\| \leq \|H_0 p + g\| \leq \kappa \|g\| \tag{16}$$

The conclusion then follows by applying Assumption 1. ☐

Below, we employ the same strategy as in [14] to prove the global convergence of the PALM-NA algorithm. Before presenting the main theorem, we introduce the following fundamental yet important lemmas:

Lemma 1. *Under Assumptions 1 and Proposition 1, the following conclusions hold:*

(i) The sequence $\left\{F\left(X^k, Y^k\right)\right\}_{k\in\mathbb{N}}$ is non-increasing. Furthermore, we have the estimate:

$$\frac{c_1}{2}\left(\left\|X^{k+1} - W^k\right\|^2 + \left\|Y^{k+1} - H^k\right\|^2 + \left\|W^k - X^k\right\|^2 + \left\|H^k - Y^k\right\|^2\right) \leq F(X^k, Y^k) - F(X^{k+1}, Y^{k+1}), \quad \forall k \geq 0,$$

where c_1 is given by:

$$c_1 = \min\left\{\frac{\rho'(\mu_1^-)^2}{(1+\kappa)^2\mu_1^+}, \frac{\rho'(\mu_2^-)^2}{(1+\kappa)^2\mu_2^+}, (\gamma_1 - 1)\theta_1^-, (\gamma_2 - 1)\theta_2^-\right\}.$$

(ii) Moreover, we have:

$$\sum_{k=0}^{\infty}\left(\left\|X^{k+1} - W^k\right\|^2 + \left\|Y^{k+1} - H^k\right\|^2 + \left\|W^k - X^k\right\|^2 + \left\|H^k - Y^k\right\|^2\right) < \infty,$$

which implies

$$\lim_{k\to\infty}\left\|X^{k+1} - W^k\right\| = \lim_{k\to\infty}\left\|Y^{k+1} - H^k\right\| = \lim_{k\to\infty}\left\|W^k - X^k\right\| = \lim_{k\to\infty}\left\|H^k - Y^k\right\| = 0.$$

Proof. (i) It is straightforward to see that we only need to prove one side, as the updates on the other side are similar. Without loss of generality, we focus on the side with fixed Y. First, for the update from (X^k, Y^k) to (W^k, Y^k), which is a single step of PGD, the sufficient decrease property of the PGD algorithm directly gives:

$$F(W^k, Y^k) \leq F(X^k, Y^k) - \frac{1}{2}(\gamma_1 - 1)L_1(Y^k)\|W^k - X^k\|^2 \tag{17}$$

Next, for the update from (W^k, Y^k) to (X^{k+1}, Y^k), which is a trust-region step. By Proposition 2, we may consider only the case where the update direction lies strictly inside the trust region. Then, using Proposition 3 and inequality (12), we obtain:

$$\begin{aligned} F(W^k, Y^k) - F(X^{k+1}, Y^k) &\geq \frac{\rho'|g|^2}{2|H_0|} && (18)\\ &\geq \frac{\rho'\lambda_{\min}^2(H_0)}{2(1+\kappa)^2\lambda_{\max}(H_0)}\|W^k - X^{k+1}\|^2 && (19)\\ &\geq \frac{\rho'(\mu_1^-)^2}{2(1+\kappa)^2\mu_1^+}\|W^k - X^{k+1}\|^2 && (20) \end{aligned}$$

(ii) The result follows directly by summing the inequalities in (i) and applying Proposition 1. ☐

Lemma 2. *Under Proposition 1, for any positive integer k, there exist A_X^k and A_Y^k such that $(A_X^k, A_Y^k) \in \partial F\left(X^{k+1}, Y^{k+1}\right)$, and we have the following bound:*

$$\begin{aligned} \left\|(A_X^k, A_Y^k)\right\| &\leq \left\|A_X^k\right\| + \left\|A_Y^k\right\| && (21)\\ &\leq (2M + c_2)\left(\left\|X^{k+1} - W^k\right\| + \left\|Y^{k+1} - H^k\right\| + \left\|W^k - X^k\right\| + \left\|H^k - Y^k\right\|\right) && (22) \end{aligned}$$

where c_2 is given by:

$$c_2 = \max\left\{\gamma_1\theta_1^+, \gamma_2\theta_2^+\right\}$$

Proof. We similarly only need to prove the update on the X-side. First, for the update from (X^k, Y^k) to (W^k, Y^k), which is a proximal gradient update:

$$W^k \in \operatorname{argmin}_{X\in\mathbb{R}^{n\times r}} \left\{ \langle X - X^k, \nabla_X f(X^k, Y^k)\rangle + \frac{c_k}{2}\|X - X^k\|^2 + \sigma_1(X) \right\}. \tag{23}$$

By the optimality condition, we have:

$$\nabla_X f(X^k, Y^k) + c_k(W^k - X^k) + u^k = 0 \tag{24}$$

where $u^k \in \partial\sigma_1(W^k)$. Thus, we obtain:

$$\partial_X F(W^k, Y^k) \ni \nabla_X f(W^k, Y^k) + u^k = \nabla_X f(W^k, Y^k) - \nabla_X f(X^k, Y^k) + \nabla_X f(X^k, Y^k) + u^k \tag{25}$$

$$= \nabla_X f(W^k, Y^k) - \nabla_X f(X^k, Y^k) - c_k(W^k - X^k) \tag{26}$$

Using Proposition 1, we then have:

$$\|\nabla_X f(W^k, Y^k) + u^k\| = \|\nabla_X f(W^k, Y^k) - \nabla_X f(X^k, Y^k) - c_k(W^k - X^k)\| \tag{27}$$

$$\leq \|\nabla_X f(W^k, Y^k) - \nabla_X f(X^k, Y^k)\| + c_k\|W^k - X^k\| \tag{28}$$

$$\leq (M + \gamma_1\theta_1^+)\|W^k - X^k\| \tag{29}$$

Now consider the point (X^k, Y^k) to (X^{k-1}, Y^{k-1}). Again, we only need to consider the X-side case. We have:

$$\nabla_X f(X^{k+1}, Y^{k+1}) + v^k = \nabla_X f(X^{k+1}, Y^{k+1}) - \nabla_X f(W^k, H^k) + \nabla_X f(W^k, H^k) - \nabla_X f(W^k, Y^k) \tag{30}$$

$$+ v^k - u^k + \nabla_X f(W^k, Y^k) + u^k \tag{31}$$

where $v^k \in \partial\sigma_1(X^{k+1})$. Taking $A_X^k = \nabla_X f(X^{k+1}, Y^{k+1}) + v^k$, and using Proposition 1 and inequality (29), we obtain the estimate:

$$\|A_X^k\| \leq \|\nabla_X f(X^{k+1}, Y^{k+1}) - \nabla_X f(W^k, H^k)\| + \|\nabla_X f(W^k, H^k) - \nabla_X f(W^k, Y^k)\| \tag{32}$$

$$+ \|v^k - u^k\| + \|\nabla_X f(W^k, Y^k) + u^k\| \tag{33}$$

$$\leq M(\|X^{k+1} - W^k\| + \|Y^{k+1} - H^k\| + \|H^k - Y^k\|) + (M + \gamma_1\theta_1^+)\|W^k - X^k\| \tag{34}$$

Note that due to the regularizer being $\|\cdot\|_1$ or $\delta_{\mathbb{B}_0^\alpha}(\cdot)$, and the manifold selection in our trust-region step, we can take $\|v^k - u^k\| = 0$. Similarly, combining with the estimate on the Y-side, we obtain:

$$\left\|(A_X^k, A_Y^k)\right\| \leq \|A_X^k\| + \|A_Y^k\| \tag{35}$$

$$\leq (2M + c_2)\left(\|X^{k+1} - W^k\| + \|Y^{k+1} - H^k\| + \|W^k - X^k\| + \|H^k - Y^k\|\right) \tag{36}$$

□

Lemma 3. *Under Proposition 1, let $\omega(Z^0)$ denote the set of limit points of the sequence generated from the initial point (X^0, Y^0). Then the following conclusions hold:*

(i) $\emptyset \neq \omega\left(Z^0\right) \subset \operatorname{crit} F$

(ii) We have

$$\lim_{k\to\infty} \operatorname{dist}\left(Z^k, \omega\left(Z^0\right)\right) = 0.$$

(iii) The set $\omega(Z^0)$ is non-empty, compact, and connected.

(iv) The objective function F is constant on the set $\omega(Z^0)$.

Proof. We only provide the proof for the case where the regularization term takes the ℓ_1-norm, and the reasoning for the ℓ_0-norm is analogous.(i) First, by Lemma 1, it is clear that the sequence $\{Z^k\}$ is bounded, and thus the set of limit points $\omega(Z^0)$ is non-empty. Furthermore, suppose there exists a subsequence $\{Z^{k_n}\}$ converging to a limit point $Z^* = (X^*, Y^*)$. Similar to before, we first consider the X-side. By the lower semicontinuity of the regularizer σ_1, we have:

$$\liminf_{n\to\infty} \sigma_1\left(Z_X^{k_n}\right) \geq \sigma_1\left(X^*\right) \tag{37}$$

Next, consider Z^{k_n}. If $Z_X^{k_n} = W^{l_n}$, then from (23), we have:

$$\left\langle W^k - X^k, \nabla_X f(X^k, Y^k)\right\rangle + \frac{c_k}{2}\left\|W^k - X^k\right\|^2 + \sigma_1\left(W^k\right) \tag{38}$$

$$\leq \left\langle X^* - X^k, \nabla_X f(X^k, Y^k)\right\rangle + \frac{c_k}{2}\left\|X^* - X^k\right\|^2 + \sigma_1\left(X^*\right). \tag{39}$$

Setting $k = l_n$ and taking the upper limit on both sides, and combining with the result of Lemma 1, we obtain:

$$\limsup_{n\to\infty} \sigma_1\left(W^{l_n}\right) \leq \limsup_{n\to\infty} \left(\left\langle X^* - X^{l_n}, \nabla_X f(X^{l_n}, Y^{l_n})\right\rangle + \frac{c_{l_n}}{2}\|X^* - X^{l_n}\|^2\right) + \sigma_1(X^*) \tag{40}$$

By Proposition 1, Lemma 1, and noting that $X^{l_n} - X^* = X^{l_n} - W^{l_n} + W^{l_n} - X^*$, we readily obtain:

$$\limsup_{n\to\infty} \sigma_1\left(W^{l_n}\right) \leq \sigma_1(X^*) \tag{41}$$

For the case $Z^{k_n} = X^{p_n}$, similarly using Lemma 1 and the result from Lemma 2, we have the following estimate:

$$\begin{aligned}
\limsup_{n\to\infty} \sigma_1(X^{p_n}) &\leq \limsup_{n\to\infty} (\sigma_1(X^{p_n}) - \sigma_1(W^{p_n-1})) + \limsup_{n\to\infty} \sigma_1(W^{p_n-1}) &(42)\\
&\leq \lambda_1\sqrt{nr} \limsup_{n\to\infty} \|X^{p_n} - W^{p_n-1}\| + \sigma_1(X^*) &(43)\\
&= \sigma_1(X^*) &(44)
\end{aligned}$$

Here, for the first term in the last inequality, from the manifold selection in the trust-region Newton step update, we know:

$$\mathcal{I}_{X^{p_n}} \subset \mathcal{I}_{W^{p_n-1}} \tag{45}$$

Thus, we have:

$$\begin{aligned}
|\sigma_1(X^{p_n}) - \sigma_1(W^{p_n-1})| &= \sigma_1(X^{p_n} - W^{p_n-1}) &(46)\\
&\leq \lambda_1\sqrt{nr}\|X^{p_n} - W^{p_n-1}\| &(47)
\end{aligned}$$

The second term in (43) can be obtained similarly using the optimality of the proximal gradient step, which we omit for brevity.

Combining (41) and (44), we obtain:

$$\limsup_{n\to\infty} \sigma_1(Z_X^{k_n}) \leq \sigma_1(X^*) \tag{48}$$

Then, from (48) and (37), it follows that:

$$\lim_{n\to\infty} \sigma_1(Z_X^{k_n}) = \sigma_1(X^*) \tag{49}$$

A similar argument holds for the Y-side. Consequently, we readily obtain:

$$\begin{aligned}
\lim_{n\to\infty} F\left(Z^{k_n}\right) &= \lim_{n\to\infty} \left\{f\left(Z^{k_n}\right) + \sigma_1\left(Z_X^{k_n}\right) + \sigma_2\left(Z_Y^{k_n}\right)\right\} &(50)\\
&= f\left(X^*, Y^*\right) + \sigma_1\left(X^*\right) + \sigma_2\left(Y^*\right) &(51)\\
&= F\left(X^*, Y^*\right) &(52)
\end{aligned}$$

Finally, by Lemma 2 and the closedness of ∂F, we have $0 \in \partial F(Z^*)$, which completes the proof.

(ii) This follows directly from the definition of the limit.

(iii), (iv) These are standard results with proofs available in (palm paper), which are omitted here. ☐

With the previous lemmas as a foundation, we now present the following theorem:

Theorem 4. *Suppose that F is a KL function and satisfies Proposition 1 as well as Assumption 1. Then the following conclusions hold:*

(i) The sequence $\left\{Z^k\right\}_{k\in\mathbb{N}}$ has finite length; in other words:

$$\sum_{k=1}^{\infty} \left\|Z^{k+1} - Z^k\right\| < \infty.$$

(ii) The sequence $\{Z^k\}_{k\in\mathbb{N}}$ converges to a stationary point $Z^* = (X^*, Y^*)$ of F.

Proof. First, from Lemma 3, we know that:

$$\lim_{k\to\infty} F\left(Z^k\right) = F\left(X^*, Y^*\right). \tag{53}$$

Suppose that $F(Z^{k_0}) = F(X^*, Y^*)$. Since our algorithm guarantees a decrease at each step, we have $F(Z^k) = F(X^*, Y^*)$ for all $k \geq k_0$, and the conclusion obviously holds in this case. Therefore, we may assume without loss of generality that $F(Z^k) > F(X^*, Y^*)$ holds for all k. Then, by the [[14] lemma 6] and Lemma 3, there exists k_0 such that for all $k \geq k_0$, if we take Ω in [[14] lemma 6] as $\omega(Z^0)$, the following holds:

$$\varphi'\left(F\left(X^k, Y^k\right) - F\left(Z^*\right)\right) \mathrm{dist}\left(0, \partial F\left(X^k, Y^k\right)\right) \geq 1. \tag{54}$$

Furthermore, by Lemma 2, we obtain:

$$\varphi'\left(F\left(X^k,Y^k\right)-F\left(Z^*\right)\right) \geq \frac{1}{(2M+c_2)(\|X^k-W^{k-1}\|+\|Y^k-H^{k-1}\|+\|W^{k-1}-X^{k-1}\|+\|H^{k-1}-Y^{k-1}\|)} \tag{55}$$

Due to the concavity of φ, we have:

$$\varphi\left(F\left(X^k,Y^k\right)-F\left(X^*,Y^*\right)\right)-\varphi\left(F\left(X^{k+1},Y^{k+1}\right)-F\left(X^*,Y^*\right)\right) \tag{56}$$

$$\geq \varphi'\left(F\left(X^k,Y^k\right)-F\left(X^*,Y^*\right)\right)\left(F\left(X^k,Y^k\right)-F\left(X^{k+1},Y^{k+1}\right)\right). \tag{57}$$

For $p \leq q \in \mathbb{N}$, we denote $\Delta_{p,q}$ as:

$$\Delta_{p,q} := \varphi\left(F\left(X^p,Y^p\right)-F\left(X^*,Y^*\right)\right)-\varphi\left(F\left(X^q,Y^q\right)-F\left(X^*,Y^*\right)\right) \tag{58}$$

and the constant C as:

$$C := \frac{2\left(2M+c_2\right)}{c_1} \tag{59}$$

Using (55), (57) and Lemma 1, we obtain:

$$\Delta_{k,k+1} \geq \frac{\left\|X^{k+1}-W^k\right\|^2+\left\|Y^{k+1}-H^k\right\|^2+\left\|W^k-X^k\right\|^2+\left\|H^k-Y^k\right\|^2}{C(\|X^k-W^{k-1}\|+\|Y^k-H^{k-1}\|+\|W^{k-1}-X^{k-1}\|+\|H^{k-1}-Y^{k-1}\|)}, \tag{60}$$

which implies:

$$\left\|X^{k+1}-W^k\right\|^2+\left\|Y^{k+1}-H^k\right\|^2+\left\|W^k-X^k\right\|^2+\left\|H^k-Y^k\right\|^2 \tag{61}$$

$$\leq C\Delta_{k,k+1}(\left\|X^k-W^{k-1}\right\|+\left\|Y^k-H^{k-1}\right\|+\left\|W^{k-1}-X^{k-1}\right\|+\left\|H^{k-1}-Y^{k-1}\right\|) \tag{62}$$

By further applying the Cauchy inequality, we get:

$$2(\left\|X^{k+1}-W^k\right\|+\left\|Y^{k+1}-H^k\right\|+\left\|W^k-X^k\right\|+\left\|H^k-Y^k\right\|) \tag{63}$$

$$\leq 4\sqrt{\|X^{k+1}-W^k\|^2+\|Y^{k+1}-H^k\|^2+\|W^k-X^k\|^2+\|H^k-Y^k\|^2} \tag{64}$$

$$\leq 2\sqrt{4C\Delta_{k,k+1}(\|X^k-W^{k-1}\|+\|Y^k-H^{k-1}\|+\|W^{k-1}-X^{k-1}\|+\|H^{k-1}-Y^{k-1}\|)} \tag{65}$$

$$\leq 4C\Delta_{k,k+1}+(\left\|X^k-W^{k-1}\right\|+\left\|Y^k-H^{k-1}\right\|+\left\|W^{k-1}-X^{k-1}\right\|+\left\|H^{k-1}-Y^{k-1}\right\|) \tag{66}$$

Summing both sides from k_0 to any $l > k_0$, we obtain:

$$2\sum_{k=k_0}^{l}(\left\|X^{k+1}-W^k\right\|+\left\|Y^{k+1}-H^k\right\|+\left\|W^k-X^k\right\|+\left\|H^k-Y^k\right\|) \tag{67}$$

$$\leq \sum_{k=k_0}^{l}(\left\|X^k-W^{k-1}\right\|+\left\|Y^k-H^{k-1}\right\|+\left\|W^{k-1}-X^{k-1}\right\|+\left\|H^{k-1}-Y^{k-1}\right\|)+4C\sum_{k=k_0}^{l}\Delta_{k,k+1} \tag{68}$$

$$\leq \sum_{k=k_0}^{l}(\left\|X^{k+1}-W^k\right\|+\left\|Y^{k+1}-H^k\right\|+\left\|W^k-X^k\right\|+\left\|H^k-Y^k\right\|) \tag{69}$$

$$+(\left\|X^{k_0+1}-W^{k_0}\right\|+\left\|Y^{k_0+1}-H^{k_0}\right\|+\left\|W^{k_0}-X^{k_0}\right\|+\left\|H^{k_0}-Y^{k_0}\right\|)+4C\Delta_{k_0,l+1} \tag{70}$$

Thus, it follows that:

$$\sum_{i=k_0}^{l}(\left\|X^{k+1}-W^k\right\|+\left\|Y^{k+1}-H^k\right\|+\left\|W^k-X^k\right\|+\left\|H^k-Y^k\right\|) \tag{71}$$

$$\leq(\left\|X^{k_0+1}-W^{k_0}\right\|+\left\|Y^{k_0+1}-H^{k_0}\right\|+\left\|W^{k_0}-X^{k_0}\right\|+\left\|H^{k_0}-Y^{k_0}\right\|) \tag{72}$$

$$+4C[\varphi\left(F\left(X^{k_0+1},Y^{k_0+1}\right)-F\left(X^*,Y^*\right)\right)-\varphi\left(F\left(X^{l+1},Y^{l+1}\right)-F\left(X^*,Y^*\right)\right)]. \tag{73}$$

From the above, conclusion (i) is proven.
For (ii), we only need to show that $\{Z^k\}$ is a Cauchy sequence, which is straightforward. For any $p \leq q$:

$$\|Z^q - Z^p\| = \left\| \sum_{k=p}^{q-1} (Z^{k+1} - Z^k) \right\| \tag{74}$$

$$\leq \sum_{k=p}^{q-1} \|Z^{k+1} - Z^k\| \tag{75}$$

Thus, by (i), $\{Z^k\}$ is a Cauchy sequence and therefore converges. Combined with Lemma 3, we conclude that it converges to a stationary point. □

7.2. local convergence

We first introduce some notations used throughout. We denote $\{(X^k, Y^k)\}_k$ as the sequence generated by PALM-NA, and (X^*, Y^*) as its limit point. Our goal is to prove that the KL inequality with an exponent of $\frac{1}{2}$ holds around the point (X^*, Y^*), thereby obtaining local linear convergence guarantee for the iterative sequence. In this section, we consider only the problem with ℓ_0 regularization term, specifically:

$$\min_{X,Y} F(X, Y) = f(X, Y) + R_1(X) + R_2(Y) \tag{76}$$

$$= \frac{1}{2} \|A - XY\|_F^2 + \sum_{i=1}^{r} \left[\delta_{\mathbb{R}_+^n}(x_i) + \delta_{\mathbb{B}_0^{\alpha_i}}(x_i) \right] + \sum_{i=1}^{r} \left[\delta_{\mathbb{R}_+^m}(y_i) + \delta_{\mathbb{B}_0^{\beta_i}}(y_i) \right]. \tag{77}$$

$$\tag{78}$$

According to the relevant results of manifold identification theory[30], under certain conditions, after a finite number of steps, the iterative sequence will lie in a manifold related to the limit point, which will be useful for our subsequent proof. Specifically, we have the following theorem

Theorem 5 (identification for functions)**.** *Let the function F be $\mathcal{C}^p$-partly smooth($p \geq 2$) at the point $\bar{x}$ relative to the manifold $\mathcal{M}$, and prox-regular there, with $0 \in$ rint $\partial F(\bar{x})$. Suppose $x_k \to \bar{x}$ and $F(x_k) \to F(\bar{x})$. Then*

$$x_k \in \mathcal{M} \text{ for all large } k \Leftrightarrow \operatorname{dist}(0, \partial F(x_k)) \to 0. \tag{79}$$

The specific definitions of $\mathcal{C}^p$-partly smooth and prox-regular are omitted, and further details may be found in the relevant articles[30].

Before proceeding to the proof of Theorem 2, we first present some definitions that will be used. By viewing matrices as vectors and adopting the standard notation, we define the support set S_M of a matrix M as follows:

Definition 4. *For matrix $M \in \mathbb{R}^{p \times q}$, we define S_M as*

$$S_M \triangleq \{(i, j) \in [p] \times [q] \mid M_{ij} \neq 0\} \tag{80}$$

We denote the manifold spanned by the non-zero components of (X^*, Y^*) as $\mathcal{M}_*$, specifically as follows

Definition 5. *Let S_{X^*} and S_{Y^*} be the support sets of X^* and Y^*, respectively, and define $\mathcal{M}_*$ as*

$$\mathcal{M}_* \triangleq \{(X, Y) \in \mathbb{R}_+^{n \times r} \times \mathbb{R}_+^{r \times m} | S_X = S_{X^*}, Y = S_{Y^*}\} \tag{81}$$

Proof of Thm 2

Proof. For smooth terms $f(X, Y)$, conditions $\mathcal{C}^p$-partly smooth and prox-regular hold naturally, we need only consider the ℓ_0 regularization term. Without loss of generality, we consider the vector case $R(x) = \delta_C(x)$, where $C = \{x \in \mathbb{R}^d : x \geq 0, \|x\|_0 \leq k\}$, $k \leq d$. First, For any point $x \in C$, we denote $S_x := \{i : x_i > 0\}$, $s(x) = |S_x|$. For a point $\bar{x} \in C$, we can get Fréchet subdifferential of δ_C at point $\bar{x}$:

$$\hat{\partial}\delta_C(\bar{x}) = \begin{cases} \{v \in \mathbb{R}^d : v_i = 0,\ i \in S_{\bar{x}}.\ v_j \leq 0,\ j \notin S_{\bar{x}}\} & s(\bar{x}) < k \\ \{v \in \mathbb{R}^d : v_i = 0,\ i \in S_{\bar{x}}\} & s(\bar{x}) = k \end{cases} \tag{82}$$

Therefore, we obtain the subdifferential:

$$\begin{aligned} \partial\delta_C(\bar{x}) &= \limsup_{x \to \bar{x}, \delta_C(x) \to \delta_C(\bar{x})} \hat{\partial}\delta_C(x) \\ &= \begin{cases} \{v \in \mathbb{R}^d : v_i = 0,\ i \in S_{\bar{x}}.\ v_j \leq 0,\ j \notin S_{\bar{x}}\} \cup \{v \in \mathbb{R}^d : v_i = 0,\ i \in S_{\bar{x}},\ |Z_v| \geq k\} & s(\bar{x}) < k \\ \{v \in \mathbb{R}^d : v_i = 0,\ i \in S_{\bar{x}}\} & s(\bar{x}) = k \end{cases} \end{aligned} \tag{83}$$

where $Z_v = \{i : v_i = 0\}$. It is not Clarke regular at x when $s < k$. Consequently, we base our analysis on the assumption that $s = k$, a condition that is both practical and, as shown above, sufficient for regularity. The rationality of this condition is further confirmed through corresponding numerical simulations.

For the case $s = k$, it is regular as established above, and it is straightforward to demonstrate its prox-regularity. Regarding partly smooth, we define the manifold $\mathcal{M}_{\bar{x}}$ as

$$\mathcal{M}_{\bar{x}} \triangleq \{x \in \mathbb{R}_+^d | S_x = S_{\bar{x}}\} \tag{84}$$

the smoothness of δ_C and the continuity of $\partial\delta_C$ on the manifold $\mathcal{M}_{\bar{x}}$ are immediate. As for sharpness, we first have:

$$\text{par}(\partial\delta_C(\bar{x})) = \{v \in \mathbb{R}^d : v_i = 0,\ i \in S_{\bar{x}}\} \tag{85}$$

For the manifold $\mathcal{M}_{\bar{x}}$, we obtain:

$$T_{\bar{x}}\mathcal{M}_{\bar{x}} = \{v \in \mathbb{R}^d : v_i = 0,\ i \notin S_{\bar{x}}\} \tag{86}$$

Hence, we get:

$$N_{\bar{x}}\mathcal{M}_{\bar{x}} = \{v \in \mathbb{R}^d : v_i = 0,\ i \in S_{\bar{x}}\} \tag{87}$$

$$= \text{par}(\partial\delta_C(\bar{x})) \tag{88}$$

At this stage, we have verified that the function is both prox-regular and partly smooth with respect to the manifold $\mathcal{M}_*$. Condition $\text{dist}(0, \partial F(x_k)) \to 0$ is guaranteed by lemma 2 and 1, we obtain the result from theorem 5. □

Let us recall the MB condition, we need to verify the condition for the MB property:

1. The function F is constant on the local solution set $\mathcal{S}$
2. $\mathcal{S}$ is a smooth embedded submanifold of $\mathcal{M}_*$.
3. The Riemannian Hessian of F is zero along the tangent space $T_{(X,Y)}\mathcal{S}$, and positive definite on the normal space $N_{(X,Y)}\mathcal{S}$.

Note that the original problem is a non-smooth problem, and the MB condition requires that the objective function be at least C^2 on the manifold $\mathcal{M}$. From Thm 2, the sequence $\{(X^k, Y^k)\}_k$ generated by the PALM-NA satisfies the requirements of the previous manifold identification theorem 5. Therefore, in order to prove the linear convergence rate of sequence $\{(X^k, Y^k)\}_k$, we can naturally select the manifold in the MB property as $\mathcal{M}_*$, and the objective function Φ is the restriction $F|_{\mathcal{M}_*}$ which is smooth around (X^*, Y^*) on $\mathcal{M}_*$.

7.2.1. General case

In this subsection, we analyse the general case under assumption [2, 3, 4], we will show that conditions 1 and 3 of the MB property are satisfied around (X^*, Y^*) on $\mathcal{M}_*$.

Lemma 4. *Suppose that point $Z^* = (X^*, Y^*)$ is a global optimum of problem 76, i.e. $\|X^*Y^* = A\|_F = 0$, then*

$$\mathcal{S} = \{(X, Y) \in \mathcal{M}_* | XY = A\} \tag{89}$$

Proof. We denote the set of critical points of F as $\mathcal{X}$. First, F is subanalytic, it follows that F is constant on every connected component of $\mathcal{X}$ from [32]. Furthermore, F is constant around (X^*, Y^*) on $\mathcal{X}$. Next, since (X^*, Y^*) is a global optimal solution, there exists a neighbourhood U of (X^*, Y^*), for any $(X, Y) \in \mathcal{X} \cap U$, we have

$$\begin{aligned} \|A - XY\|_F^2 &= \|A - X^*Y^*\|_F^2 \\ &= 0 \end{aligned} \tag{90}$$

This completes the proof. □

Under Assumption 3, We analyze the tangent and normal directions of the manifold $\mathcal{S}$.

Lemma 5. *Consider the map:*

$$h : \mathcal{M}_* \to \mathbb{R}_+^{n\times m}$$
$$(X, Y) \to XY$$

Then local solution set $\mathcal{S}$ is level set $h^{-1}(A)$,and we consider differential of map h at point (X, Y):

$$dh = XdY + dXY$$

We denote the kernel of dh by $\mathcal{N}(X, Y)$ at point (X, Y),then we have:

$$\mathcal{N}(X, Y) = \{(U, V) \in T_{(X,Y)}\mathcal{M}_* : XV + UY = 0\}$$

and orthogonal complement of $\mathcal{N}$:

$$\mathcal{N}^\perp(X, Y) = \{(\Lambda Y^T, X^T\Lambda) \in T_{(X,Y)}\mathcal{M}_*, \Lambda \in \mathbb{R}^{n\times m}\}$$

Proof. The computation of the tangent space is routine, so we restrict our attention to the derivation of the norm space. Note that $\mathcal{N}$ is the kernel of a linear map. Specifically, the map $L : T_{(X,Y)}\mathcal{M}_* \to \mathbb{R}^{n\times m}$ is defined by

$$L(U, V) = XV + UY \tag{91}$$

Recall that the adjoint mapping of L,denoted by L^*. Then $\mathcal{N}^\perp = (Ker(L))^\perp = Im(L^*)$

And for any $\Lambda \in \mathbb{R}^{n\times m}$,we have

$$\begin{aligned}\langle L(U, V), \Lambda\rangle &= \langle XV + UY, \Lambda\rangle \\ &= tr(XV\Lambda^T) + tr(UY\Lambda^T) \\ &= tr(V\Lambda^T X) + tr(UY\Lambda^T) \\ &= \langle U, \Lambda Y^T\rangle + \langle V, X^T\Lambda\rangle \\ &= \langle (U, V), (\Lambda Y^T, X^T\Lambda)\rangle\end{aligned} \tag{92}$$

From the uniqueness of the adjoint mapping, it follows that $L^*(\Lambda) = (\Lambda Y^T, X^T\Lambda)$. □

Proposition 4. *Suppose that $(X, Y) \in \mathcal{S}$, then the restriction of $\nabla^2 F(X, Y)$ to $\mathcal{N}^\perp$ is positive definite.*

Proof. For the sake of convenience, we continue to use matrix form to express the variable of F. We can see that the Riemannian Hessian of F is given by

$$Hess\ F(X, Y)[U, V] = \begin{bmatrix} P_{X^*}(W_X) \\ P_{Y^*}(W_Y) \end{bmatrix} \tag{93}$$

where $P_{X^*}(W_X), P_{Y^*}(W_Y)$ are the corresponding X and Y components of $[W_X, W_Y]$ projected onto $\mathcal{M}_*$. Specifically, for $R \in \mathbb{R}^{n\times r}$, $(i, j) \in [n] \times [r]$

$$P_{X^*}(R)_{ij} = \begin{cases} R_{ij} & \text{if } (i, j) \in S_{X^*} \\ 0 & otherwise \end{cases} \tag{94}$$

and $P_{Y^*}(\cdot)$, similarly defined as above. We now turn our attention to the variables W_X and W_Y in equation 93, which are

$$\begin{aligned} W_X &= (XY - A)V^T + UYY^T + XVY^T \\ W_Y &= U^T(XY - A) + X^T XV + X^T UY \end{aligned} \tag{95}$$

This equation is obtained by taking the second-order derivative of the function $f(X, Y)$.

Assume that $[U, V] \in \mathcal{N}^\perp$, $Hess\ F(X, Y)[U, V] = \mathbf{0}$. Then

$$\begin{aligned} 0 &= \langle (U, V), Hess\ F(X, Y)[U, V]\rangle \\ &= tr(U^T P_{X^*}(W_X)) + tr(V^T P_{Y^*}(W_Y)) \end{aligned} \tag{96}$$

Since $P_{X^*}(U) = U$ and $P_{Y^*}(V) = V$, this implies that $tr(U^T P_{X^*}(W_X)) = tr(U^T W_X)$ and $tr(V^T P_{Y^*}(W_Y)) = tr(V^T W_Y)$. Substituting the explicit expression for $[U, V]$ from Lemma 5 yields that

$$\begin{aligned} tr(U^T W_X) + tr(V^T W_Y) &= tr(Y\Lambda^T(XY - A)\Lambda^T X)+ \\ & tr(Y\Lambda^T\Lambda Y^T YY^T) + tr(Y\Lambda^T XX^T\Lambda Y^T)+ \\ & tr(\Lambda^T XX^T\Lambda Y^T Y) + tr(\Lambda^T XX^T XX^T\Lambda)+ \\ & tr(\Lambda^T XY\Lambda^T(XY - A)) \\ &= tr(Y\Lambda^T\Lambda Y^T YY^T) + tr(Y\Lambda^T XX^T\Lambda Y^T)+ \\ & tr(\Lambda^T XX^T\Lambda Y^T Y) + tr(\Lambda^T XX^T XX^T\Lambda) \\ &= 0 \end{aligned} \tag{97}$$

The second equation follows from $(X, Y) \in \mathcal{S}$. And note that each term on the right-hand side of the equation is nonnegative. From the first item, we have

$$\begin{aligned} tr(Y\Lambda^T\Lambda Y^T YY^T) = 0 &\Rightarrow \Lambda Y^T Y = 0 \\ &\Rightarrow \Lambda Y^T Y\Lambda^T = 0 \\ &\Rightarrow \Lambda Y^T = 0 \end{aligned} \tag{98}$$

Similarly, we derive $X^T\Lambda = 0$ from the last item, Thus $[U, V] = \mathbf{0}$, the proposition holds. □

Proof of Thm 3

Proof. With the above results established, the conclusion of Theorem 3 follows naturally. First, we analyse the critical point conditions:

$$0 \in \partial F(X,Y) \Leftrightarrow -\nabla_X f(X,Y) \in \partial R_1(X) \text{ and } -\nabla_Y f(X,Y) \in \partial R_2(Y) \tag{99}$$

Without loss of generality, we focus on variable X. We are using the Fréchet subdifferential as the generalized subdifferential. Then from equation 82, we get

$$-\nabla_X f(X,Y) \in \partial R_1(X) \Leftrightarrow \begin{cases} (\nabla_X f(X,Y))_{ij} = 0 & X_{ij} > 0 \\ (\nabla_X f(X,Y))_{ij} \geq 0 & X_{ij} = 0 \end{cases} \tag{100}$$

By the smoothness of f and Assumption 2, the critical point set of F coincides with that of $F|\mathcal{M}$ near $(X^{,}Y^*)$. The original problem is equivalent to minimizing the function F restricted to the manifold $\mathcal{M}_*$. By Lemma 4 and Proposition 4, we establish that the MB property holds at point (X^*, Y^*) under assumption 3. Finally, by the equivalence of the MB property and the PL inequality [19], the proof is complete. □

Remark. *Under Assumption 4, Assumption 2 is actually redundant here, since $\nabla f(X,Y) = 0$ on $\mathcal{S}$ can be readily deduced from Assumption 4. It thus follows that minimizing F locally naturally coincides with minimizing its restriction $F|_{\mathcal{M}_*}$. We remark that Assumption 2 is conceptually more fundamental and applies in greater generality. To maintain the comprehensiveness of our framework, we deliberately retain it as a stated assumption.*

7.2.2. *KL exponent is $\frac{1}{2}$ for $r = 1$*

In the previous section, we presented the main theoretical results under the assumption that the set of local solutions $\mathcal{S}$ forms a submanifold, without providing a formal theoretical justification. The primary difficulty arises from the significant challenges involved in constructing a direct proof. A natural approach is to apply the constant rank theorem; however, this requires analyzing the distribution of non-zero components of (X^*, Y^*), which introduces considerable complexity into the theoretical framework. In this section, we conduct a detailed theoretical treatment of the simpler case where $r = 1$

Proposition 5. *If $r = 1$ and $A \neq \mathbf{0}$, then $\mathcal{S}$ is a 1-dimensional submanifold of $\mathcal{M}_*$.*

Proof. First, there exists $a \in \mathbb{R}_+^{n\times 1}, b \in \mathbb{R}_+^{1\times m}$ such that $A = ab$. Let $(x^*, y^*) \in \mathcal{S}$ is a global optimum, then $x^* y^* = ab$, for $1 \leq i \leq n, 1 \leq j \leq m$, we have

$$x_i^* y_j^* = a_i b_j \tag{101}$$

Since $A \neq \mathbf{0}$, there exist $1 \leq p \leq n, 1 \leq q \leq m$, $A_{pq} > 0$, s.t. $x_p^* > 0, y_q^* > 0$. Then, setting $i = p$ for all j, we have

$$x_p^* y_j^* = a_p b_j \tag{102}$$

First take $j = q$, then $a_p > 0$ and $b_q > 0$. Furthermore, if $y_j^* = 0$, then $b_j = 0$ and vice versa. Therefore, we conclude that $S_{y^*} = S_b$ and similarly $S_{x^*} = S_a$. Furthermore, a simple rearrangement of the equation (101) yields the equivalent form:

$$\frac{x_i^*}{a_i} = \frac{b_j}{y_j^*} = M_{ij} > 0 \tag{103}$$

where $i \in S_a, j \in S_b$. It is not difficult to conclude that $M_{ij} = M > 0$. Therefore, we get

$$\mathcal{S} = \{(Ma, \frac{1}{M} b), M > 0\} \tag{104}$$

which is a 1-dimensional submanifold of $\mathcal{M}_*$.Therefore, we establish the following corresponding conclusions. □

7.3. Experimental Supplement

7.3.1. *Implementation Details*

For completeness and ease of reference, the algorithms used by Algorithm 1 are listed below,which can be found in [20].

Algorithm 2 Trust-Region Method on Manifold

Input: Current iterate X_k, initial trust-region radius $0 < \Delta_0 \leq \bar{\Delta}$, and $0 < \rho' < 1/4$
$k = 0$
repeat
  $\eta_k = \text{tCG}(X_k, \Delta_k, H_k, P_k)$
  $X_k^+ = \text{Retr}_{X_k}(\eta_k)$
  $\rho_1 = f(X_k, Y_k) - f(X_k^+, Y_k)$
  $\rho_2 = m_k(0) - m_k(\eta_k)$
  if $\rho_1/\rho_2 < 1/4$ **then**
    $\Delta_{k+1} = \Delta_k/4$
  else if $\rho_1/\rho_2 > 3/4$ and the tCG solution reaches the trust-region boundary **then**
    $\Delta_{k+1} = \min(2\Delta_k, \bar{\Delta})$
  else
    $\Delta_{k+1} = \Delta_k$
  end if
  if $\rho_1/\rho_2 > \rho'$ **then**
    $X_{k+1} = X_k^+$
  else
    $X_{k+1} = X_k$
  end if
  $k = k + 1$
until stopping criterion is satisfied

Algorithm 3 Truncated Conjugate Gradient (tCG)

Input: $X \in \mathcal{M}_X$, $\Delta, \theta, \kappa > 0$, $H, P : \mathcal{T}_X\mathcal{M}_X \to \mathcal{T}_X\mathcal{M}_X$, $N > 0$
$\eta^0 = 0 \in \mathcal{T}_X\mathcal{M}_X$, $r_0 = \text{grad}\, f(X)$, $z_0 = P[r_0]$, $\delta_0 = -z_0$
for $j = 0, \cdots, N$ **do**
  $\kappa_j = \langle \delta_j, H[\delta_j] \rangle_X$
  $\alpha_j = \langle z_j, r_j \rangle_X / \kappa_j$
  if $\kappa_j \leq 0$ **or** $\|\eta^j + \alpha_j \delta_j\|_{P^{-1}} \geq \Delta$ **then**
    Find τ_j ¿ 0 s.t. $\|\eta^j + \tau_j \delta_j\|_{P^{-1}}^2 = \Delta^2$
    $\eta^{j+1} = \eta^j + \tau_j \delta_j$
    if $m(\eta^{j+1}) \geq m(\eta^j)$ **then**
      return η^j
    end if
    return η^{j+1}
  end if
  $\eta^{j+1} = \eta^j + \alpha_j \delta_j$
  if $m(\eta^{j+1}) \geq m(\eta^j)$ **then**
    return η^j
  end if
  $r_{j+1} = r_j + \alpha_j H[\delta_j]$
  if $\|r_{j+1}\|_X \leq \|r_0\|_X \cdot \min(\|r_0\|_X^\theta, \kappa)$ **then**
    return η^{j+1}
  end if
  $z_{j+1} = P[r_{j+1}]$
  $\beta_j = \langle z_{j+1}, r_{j+1} \rangle_X / \langle z_j, r_j \rangle_X$
  $\delta_{j+1} = -z_{j+1} + \beta_j \delta_j$
end for
return η^N

Fig. 1. ManAcc algorithm and its subprocedure.

The above is the detailed pseudocode of the ManAcc algorithm used in this paper, where $\text{Retr}_{X_k}(\eta_k)$ denotes the retraction on the manifold $\mathcal{M}_{X_k}$. A simple choice for the retraction is to directly perform an orthogonal projection onto the manifold; it is only necessary to control the trust region radius to ensure its validity.

7.3.2. Numerical Evidence for Local Linear Convergence

In this subsection, we conduct numerical experiments to validate the local linear convergence of the algorithm. The objective function, dataset, and relevant settings remain consistent with those described in the main text. In the experiments, we first run the algorithm until the stopping condition $\log 10(F(X^k, Y^k) - F(X^{k+1} - Y^{k+1})) < 10^{-12}$ is met, and take the resulting iterate as the optimal solution. Then, starting from the same initial point, we perform a second run and obtain the experimental results shown in the following figure.

7.3.3. Verification of the Condition s = k

This section is devoted to verifying the assumption $s = k$, under which the iterative sequence lies on the boundary of the constraint set. The objective function, dataset, and parameter settings remain consistent with those in the main text. We conduct tests with three random initial points, and the experimental results are summarized in the table below.