

Supplemental Materials for: "Dropping Symmetry for Fast Symmetric Nonnegative Matrix Factorization"

1 Problem Statement and Transferring Symmetric NMF to Nonsymmetric NMF

Problem Statement. General nonnegative matrix factorization (NMF) is referred to the following problem: Given a matrix $\mathbf{Y} \in \mathbb{R}^{n \times m}$ and a factorization rank r , solve

$$\min_{\mathbf{U} \in \mathbb{R}^{n \times r}, \mathbf{V} \in \mathbb{R}^{m \times r}} \frac{1}{2} \|\mathbf{Y} - \mathbf{UV}^T\|_F^2, \quad \text{subject to } \mathbf{U} \geq \mathbf{0}, \mathbf{V} \geq \mathbf{0}, \quad (1)$$

where $\mathbf{U} \geq \mathbf{0}$ means each element in \mathbf{U} is nonnegative.

When the two factors \mathbf{U} and \mathbf{V} are required identical, (2) becomes the following symmetric NMF: given a PSD matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$

$$\min_{\mathbf{U}} \frac{1}{2} \|\mathbf{X} - \mathbf{UU}^T\|_F^2, \quad \text{subject to } \mathbf{U} \geq \mathbf{0}. \quad (2)$$

In the following, we call (1) nonsymmetric NMF.

Problem Reformulation: Dropping Symmetry. We transfer the symmetric NMF problem in (2) to

$$\min_{\mathbf{U}, \mathbf{V}} f(\mathbf{U}, \mathbf{V}) = \frac{1}{2} \|\mathbf{X} - \mathbf{UV}^T\|_F^2 + \frac{\lambda}{2} \|\mathbf{U} - \mathbf{V}\|_F^2, \quad \text{subject to } \mathbf{U} \geq \mathbf{0}, \mathbf{V} \geq \mathbf{0}. \quad (3)$$

To theoretically establish this transformation, we restate Theorem 1, Lemma 1, and Theorem 2 in the original paper and give their proofs.

Theorem 1 (Restatement of Theorem 1 in the original paper). Suppose $(\mathbf{U}^*, \mathbf{V}^*)$ be any critical point of (3) satisfying $\|\mathbf{U}^* \mathbf{V}^{*\top}\| < 2\lambda + \sigma_n(\mathbf{X})$, where $\sigma_n(\cdot)$ denotes the n -th largest singular value. Then $\mathbf{U}^* = \mathbf{V}^*$ and \mathbf{U}^* is a critical point of (2).

Proof of Theorem 1. We first preset the following useful result, which generalizes the classical result for two PSD matrices.

Lemma 1. For any symmetric $\mathbf{A} \in \mathbb{R}^{n \times n}$ and PSD matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$, we have

$$\sigma_n(\mathbf{A}) \text{trace}(\mathbf{B}) \leq \text{trace}(\mathbf{AB}) \leq \sigma_1(\mathbf{A}) \text{trace}(\mathbf{B}),$$

where $\sigma_i(\mathbf{A})$ is the i -th largest eigenvalue of \mathbf{A} .

Proof of Lemma 1. Let $\mathbf{A} = \Phi_1 \Lambda_1 \Phi_1^T$ and $\mathbf{B} = \Phi_2 \Lambda_2 \Phi_2^T$ be the eigendecompositions of \mathbf{A} and \mathbf{B} , respectively. Here Λ_1 (Λ_2) is a diagonal matrix with the eigenvalues of \mathbf{A} (\mathbf{B}) along its diagonal. We first rewrite $\text{trace}(\mathbf{AB})$ as

$$\text{trace}(\mathbf{AB}) = \text{trace}(\Lambda_1 \Phi_1^T \Phi_2 \Lambda_2 \Phi_2^T \Phi_1).$$

Noting that Λ_1 is a diagonal matrix and $\Phi_1^T \Phi_2 \Lambda_2 \Phi_2^T \Phi_1 \succeq \mathbf{0}$ since $\Lambda_2 \succeq \mathbf{0}$, we have

$$\text{trace}(\Lambda_1 \Phi_1^T \Phi_2 \Lambda_2 \Phi_2^T \Phi_1) \leq \max_i \Lambda_1[i, i] \cdot \text{trace}(\Phi_1^T \Phi_2 \Lambda_2 \Phi_2^T \Phi_1) = \sigma_1(\mathbf{A}) \text{trace}(\mathbf{B}).$$

The other direction follows similarly. □

We now turn to proving Theorem 1. The subdifferential of f is given as follows

$$\partial_U f(\mathbf{U}, \mathbf{V}) = (\mathbf{U}\mathbf{V}^T - \mathbf{X})\mathbf{V} + \lambda(\mathbf{U} - \mathbf{V}) + \partial\delta_+(\mathbf{U}), \quad (4)$$

$$\partial_V f(\mathbf{U}, \mathbf{V}) = (\mathbf{U}\mathbf{V}^T - \mathbf{X})^T\mathbf{U} - \lambda(\mathbf{U} - \mathbf{V}) + \partial\delta_+(\mathbf{V}), \quad (5)$$

where $\partial\delta_+(\mathbf{U}) = \{\mathbf{G} \in \mathbb{R}^{n \times r} : \mathbf{G} \circ \mathbf{U} = \mathbf{0}, \mathbf{G} \leq \mathbf{0}\}$ when $\mathbf{U} \geq \mathbf{0}$ and otherwise $\partial\delta_+(\mathbf{U}) = \emptyset$. Since $(\mathbf{U}^*, \mathbf{V}^*)$ is a critical point of (3), it satisfies

$$(\mathbf{U}^*\mathbf{V}^{*\top} - \mathbf{X})\mathbf{V}^* + \lambda(\mathbf{U}^* - \mathbf{V}^*) + \mathbf{G} = \mathbf{0}, \quad (6)$$

$$(\mathbf{U}^*\mathbf{V}^{*\top} - \mathbf{X})^T\mathbf{U}^* - \lambda(\mathbf{U}^* - \mathbf{V}^*) + \mathbf{H} = \mathbf{0}, \quad (7)$$

where $\mathbf{G} \in \partial\delta_+(\mathbf{U}^*)$ and $\mathbf{H} \in \partial\delta_+(\mathbf{V}^*)$. Subtracting (7) from (6), we have

$$(2\lambda\mathbf{I} + \mathbf{X})(\mathbf{U}^* - \mathbf{V}^*) = \mathbf{V}^*\mathbf{U}^{*\top}\mathbf{U}^* - \mathbf{U}^*\mathbf{V}^{*\top}\mathbf{V}^* - \mathbf{G} + \mathbf{H}. \quad (8)$$

where we utilize the fact that \mathbf{X} is symmetric, i.e., $\mathbf{X} = \mathbf{X}^T$. Taking the inner product of $\mathbf{U}^* - \mathbf{V}^*$ with both sides of the above equation gives

$$\langle (2\lambda\mathbf{I} + \mathbf{X}), (\mathbf{U}^* - \mathbf{V}^*)(\mathbf{U}^* - \mathbf{V}^*)^T \rangle = \langle \mathbf{V}^*\mathbf{U}^{*\top}\mathbf{U}^* - \mathbf{U}^*\mathbf{V}^{*\top}\mathbf{V}^* - \mathbf{G} + \mathbf{H}, \mathbf{U}^* - \mathbf{V}^* \rangle. \quad (9)$$

In what follows, by choosing sufficiently large λ , we show that $(\mathbf{U}^*, \mathbf{V}^*)$ satisfying (9) must satisfy $\mathbf{U}^* = \mathbf{V}^*$. To that end, we first provide the lower bound and the upper bound for the LHS and RHS of (9), respectively. Specifically,

$$\langle (2\lambda\mathbf{I} + \mathbf{X}), (\mathbf{U}^* - \mathbf{V}^*)(\mathbf{U}^* - \mathbf{V}^*)^T \rangle \geq \sigma_n((2\lambda\mathbf{I} + \mathbf{X}))\|\mathbf{U}^* - \mathbf{V}^*\|_F^2 = ((2\lambda + \sigma_n(\mathbf{X}))\|\mathbf{U}^* - \mathbf{V}^*\|_F^2, \quad (10)$$

where the inequality follows from Lemma 1. On the other hand,

$$\begin{aligned} & \langle \mathbf{V}^*\mathbf{U}^{*\top}\mathbf{U}^* - \mathbf{U}^*\mathbf{V}^{*\top}\mathbf{V}^* - \mathbf{G} + \mathbf{H}, \mathbf{U}^* - \mathbf{V}^* \rangle \\ & \leq \langle \mathbf{V}^*\mathbf{U}^{*\top}\mathbf{U}^* - \mathbf{U}^*\mathbf{V}^{*\top}\mathbf{V}^*, \mathbf{U}^* - \mathbf{V}^* \rangle \\ & = \left\langle \frac{\mathbf{V}^*\mathbf{U}^{*\top} + \mathbf{U}^*\mathbf{V}^{*\top}}{2}, (\mathbf{U}^* - \mathbf{V}^*)(\mathbf{U}^* - \mathbf{V}^*)^T \right\rangle - \frac{1}{2} \left\| \mathbf{U}^*\mathbf{V}^{*\top} - \mathbf{V}^*\mathbf{U}^{*\top} \right\|_F^2 \\ & \leq \left\langle \frac{\mathbf{V}^*\mathbf{U}^{*\top} + \mathbf{U}^*\mathbf{V}^{*\top}}{2}, (\mathbf{U}^* - \mathbf{V}^*)(\mathbf{U}^* - \mathbf{V}^*)^T \right\rangle \\ & \leq \sigma_1 \left(\frac{\mathbf{V}^*\mathbf{U}^{*\top} + \mathbf{U}^*\mathbf{V}^{*\top}}{2} \right) \|\mathbf{U}^* - \mathbf{V}^*\|_F^2 \end{aligned} \quad (11)$$

where the last inequality utilizes Lemma 1 and the first inequality follows because $\mathbf{V}^*, \mathbf{U}^* \geq \mathbf{0}$ indicating that

$$-\langle \mathbf{G}, \mathbf{U}^* - \mathbf{V}^* \rangle \leq 0, \quad \langle \mathbf{H}, \mathbf{U}^* - \mathbf{V}^* \rangle \leq 0$$

Now plugging (10) and (11) back into (9) and utilizing the assumption that $\|\mathbf{U}^*\mathbf{V}^{*\top}\|_F \leq \alpha$, we have

$$((2\lambda + \sigma_n(\mathbf{X}))\|\mathbf{U}^* - \mathbf{V}^*\|_F^2 \leq \sigma_1 \left(\frac{\mathbf{V}^*\mathbf{U}^{*\top} + \mathbf{U}^*\mathbf{V}^{*\top}}{2} \right) \|\mathbf{U}^* - \mathbf{V}^*\|_F^2 \leq \alpha\|\mathbf{U}^* - \mathbf{V}^*\|_F^2,$$

which implies that if we choose $2\lambda > \alpha - \sigma_n(\mathbf{X})$, then $\mathbf{U}^* = \mathbf{V}^*$ must hold. Plugging it into (4) gives

$$\mathbf{0} \in (\mathbf{U}^*(\mathbf{U}^*)^T - \mathbf{X})\mathbf{U}^* + \partial\delta_+(\mathbf{U}^*).$$

which implies \mathbf{U}^* is a critical point of (2). □

Lemma 2 (Restatement of Lemma 1 in the original paper). *For any local search algorithm solving (3) with initialization $\mathbf{V}_0 = \mathbf{U}_0$, suppose it sequentially decreases the objective value. Then, for any $k \geq 0$, the iterate $(\mathbf{U}_k, \mathbf{V}_k)$ generated by this algorithm satisfies*

$$\begin{aligned} \|\mathbf{U}_k\|_F^2 + \|\mathbf{V}_k\|_F^2 &\leq \left(\frac{1}{\lambda} + 2\sqrt{r}\right) \|\mathbf{X} - \mathbf{U}_0 \mathbf{U}_0^T\|_F^2 + 2\sqrt{r} \|\mathbf{X}\|_F := B_0, \\ \|\mathbf{U}_k \mathbf{V}_k^T\|_F &\leq \|\mathbf{X} - \mathbf{U}_0 \mathbf{V}_0^T\|_F + \|\mathbf{X}\|_F. \end{aligned} \quad (12)$$

Proof of Lemma 2. By the assumption that the algorithm decreases the objective function, we have

$$\frac{1}{2} \left\| \mathbf{X} - \mathbf{U}_k \mathbf{V}_k^T \right\|_F^2 + \frac{\lambda}{2} \|\mathbf{U}_k - \mathbf{V}_k\|_F^2 \leq \frac{1}{2} \left\| \mathbf{X} - \mathbf{U}_0 \mathbf{U}_0^T \right\|_F^2$$

which further implies that

$$\begin{cases} \left\| \mathbf{X} - \mathbf{U}_k \mathbf{V}_k^T \right\|_F \leq \left\| \mathbf{X} - \mathbf{U}_0 \mathbf{U}_0^T \right\|_F \\ \frac{\lambda}{2} \left(\|\mathbf{U}_k\|_F^2 + \|\mathbf{V}_k\|_F^2 - 2|\langle \mathbf{U}_k \mathbf{V}_k^T, \mathbf{I}_r \rangle| \right) \leq \frac{\lambda}{2} \|\mathbf{U}_k - \mathbf{V}_k\|_F^2 \leq \frac{1}{2} \left\| \mathbf{X} - \mathbf{U}_0 \mathbf{U}_0^T \right\|_F^2 \end{cases}$$

where the first line further gives that

$$\|\mathbf{U}_k \mathbf{V}_k^T\|_F \leq \|\mathbf{X} - \mathbf{U}_0 \mathbf{V}_0^T\|_F + \|\mathbf{X}\|_F,$$

while the second line leads to

$$\begin{aligned} \|\mathbf{U}_k\|_F^2 + \|\mathbf{V}_k\|_F^2 &\leq \frac{1}{\lambda} \|\mathbf{X} - \mathbf{U}_0 \mathbf{U}_0^T\|_F^2 + 2\|\mathbf{U}_k \mathbf{V}_k^T\|_F \|\mathbf{I}_r\|_F \\ &= \frac{1}{\lambda} \|\mathbf{X} - \mathbf{U}_0 \mathbf{U}_0^T\|_F^2 + 2\sqrt{r} \|\mathbf{U}_k \mathbf{V}_k^T\|_F \\ &\leq \left(\frac{1}{\lambda} + 2\sqrt{r}\right) \|\mathbf{X} - \mathbf{U}_0 \mathbf{U}_0^T\|_F^2 + 2\sqrt{r} \|\mathbf{X}\|_F =: B_0 \end{aligned}$$

□

Theorem 2 (Restatement of Theorem 2 in the original paper). *Choose $\lambda > \frac{1}{2}(\|\mathbf{X}\|_2 + \|\mathbf{X} - \mathbf{U}_0 \mathbf{U}_0^T\|_F - \sigma_n(\mathbf{X}))$ for (3). For any local search algorithm solving (3) with initialization $\mathbf{V}_0 = \mathbf{U}_0$, if it sequentially decreases the objective value, is convergent and converges to a critical point $(\mathbf{U}^*, \mathbf{V}^*)$ of (3), then we have $\mathbf{U}^* = \mathbf{V}^*$ and that \mathbf{U}^* is also a critical point of (2).*

Proof of Theorem 2. This theorem is a direct consequence of Theorem 1 and Lemma 2. □

2 Main Convergence Results

2.1 ANLS for symmetric NMF(SymANLS)

Algorithm 1 SymANLS

Initialization: $k = 1$ and $\mathbf{U}_0 = \mathbf{V}_0$.

- 1: **while** stop criterion not meet **do**
- 2: $\mathbf{U}_k = \arg \min_{\mathbf{V} \geq 0} \frac{1}{2} \|\mathbf{X} - \mathbf{U} \mathbf{V}_{k-1}^T\|_F^2 + \frac{\lambda}{2} \|\mathbf{U} - \mathbf{V}_{k-1}\|_F^2$;
- 3: $\mathbf{V}_k = \arg \min_{\mathbf{U} \geq 0} \frac{1}{2} \|\mathbf{X} - \mathbf{U}_k \mathbf{V}^T\|_F^2 + \frac{\lambda}{2} \|\mathbf{U}_k - \mathbf{V}\|_F^2$;
- 4: $k = k + 1$.
- 5: **end while**

Output: factorization $(\mathbf{U}_k, \mathbf{V}_k)$.

We restate Lemma 2, Theorem 3 and Corollary 1 in the original paper as follows, and provide their proof in Section 3.2.

Lemma 3 (Restatement of Lemma 2 in the original paper). *Let $\{(U_k, V_k)\}$ be the iterates sequence generated by Algorithm 1. Then we have*

$$f(U_k, V_k) - f(U_{k+1}, V_{k+1}) \geq \frac{\lambda}{2} (\|U_{k+1} - U_k\|_F^2 + \|V_{k+1} - V_k\|_F^2). \quad (13)$$

We now give the following main convergence guarantee for Algorithm 1.

Theorem 3 (Restatement of Theorem 3 in the original paper). *Let $\{(U_k, V_k)\}$ be the sequence generated by Algorithm 1. Then*

$$\lim_{k \rightarrow \infty} (U_k, V_k) = (U^*, V^*),$$

where (U^*, V^*) is a critical point of (3). Furthermore the convergence rate is at least sublinear.

Corollary 1 (Restatement of Corollary 1 in the original paper). *Suppose Algorithm 1 is initialized with $V_0 = U_0$. Choose*

$$\lambda > \frac{1}{2} \left(\|X\|_2 + \left\| X - U_0 U_0^T \right\|_F - \sigma_n(X) \right).$$

Let $\{(U_k, V_k)\}$ be the sequence generated by Algorithm 1. Then $\{(U_k, V_k)\}$ is convergent and converges to (U^, V^*) with $U^* = V^*$ and U^* a critical point of (2). Furthermore, the convergence rate is at least sublinear.*

Proof of Corollary 1. This corollary is a direct consequence of Theorem 2, Lemma 3 and Theorem 3. \square

2.2 HALS for symmetric NMF (SymHALS)

Algorithm 2 SymHALS

Initialization: U_0, V_0 , iteration $k = 1$.

```

1: while stop criterion not meet do
2:   for  $i = 1 : r$  do
3:      $X_i^k = X - \sum_{j=1}^{i-1} u_j^k (v_j^k)^T + \sum_{j=i+1}^r u_j^{k-1} (v_j^{k-1})^T$ ;
4:      $u_i^k = \arg \min_{u_i \geq 0} \frac{1}{2} \|X_i^k - u_i (v_i^{k-1})^T\|_F^2 + \frac{\lambda}{2} \|u_i - v_i^{k-1}\|_F^2 = \max \left( \frac{(X_i^k + \lambda I) v_i^{k-1}}{\|v_i^{k-1}\|_2^2 + \lambda}, 0 \right)$ ;
5:      $v_i^k = \arg \min_{v_i \geq 0} \frac{1}{2} \|X_i^k - u_i^k v_i^T\|_F^2 + \frac{\lambda}{2} \|u_i^k - v_i\|_F^2 = \max \left( \frac{(X_i^k + \lambda I) u_i^k}{\|u_i^k\|_2^2 + \lambda}, 0 \right)$ ;
6:   end for
7:    $k = k + 1$ .
8: end while
```

Output: factorization (U_k, V_k) .

Algorithm SymHALS has similar descend property and convergence guarantee to algorithm SymANLS, three theoretical results are displayed in the following and their proof can be found in Section 3.3.

Lemma 4. *Suppose the iterates sequence $\{(U_k, V_k)\}$ is generated by Algorithm 2, then we have*

$$f(U_k, V_k) - f(U_{k+1}, V_{k+1}) \geq \frac{\lambda}{2} (\|U_{k+1} - U_k\|_F^2 + \|V_{k+1} - V_k\|_F^2).$$

Theorem 4 (Sequence convergence of Algorithm 2). *For any $\lambda > 0$, let $\{(U_k, V_k)\}$ be the sequence generated by Algorithm 2. Then*

$$\lim_{k \rightarrow \infty} (U_k, V_k) = (U^*, V^*)$$

where (U^*, V^*) is a critical point of (3). Furthermore the convergence rate is at least sublinear.

Corollary 2 (Restatement of Corollary 2 in the original paper). *Suppose it is initialized with $V_0 = U_0$. Choose*

$$\lambda > \frac{1}{2} \left(\|X\|_2 + \left\| X - U_0 U_0^T \right\|_F - \sigma_n(X) \right).$$

Let $\{(\mathbf{U}_k, \mathbf{V}_k)\}$ be the sequence generated by Algorithm 2. Then $\{(\mathbf{U}_k, \mathbf{V}_k)\}$ is convergent and converges to $(\mathbf{U}^*, \mathbf{V}^*)$ with

$$\mathbf{U}^* = \mathbf{V}^*$$

and \mathbf{U}^* a critical point of (2). Furthermore, the convergence rate is at least sublinear.

Proof of Corollary 2. This corollary is a direct consequence of Theorem 2, Lemma 4 and Theorem 4. \square

3 Proof of Convergence Results

Proving Theorem 3 and Theorem 4 has very similar structure, and since the former is slightly easier to prove, hence we will give the detailed proof for Theorem 3 and provide a sketch to Theorem 4.

Notation: Define the iterates $\mathbf{W}_k = (\mathbf{U}_k, \mathbf{V}_k)$, in this remaining proofs, we may constantly change from $(\mathbf{U}_k, \mathbf{V}_k)$ to \mathbf{W}_k for notational convenience.

We first give out the high level proof sketch.

Proof sketch: We prove sequence convergence by following the framework developed in [1,2]. The main contents are to establish sufficient decrease, safeguard, and a uniform Kurdyka-Lojasiewicz inequality for our proposed algorithms, which are stated in the following:

- *sufficient decrease.* $\exists C' > 0$, s.t.

$$f(\mathbf{W}_k) - f(\mathbf{W}_{k+1}) \geq C' \|\mathbf{W}_{k+1} - \mathbf{W}_k\|_F^2.$$

- *and safeguard.* $\exists C'' > 0$, s.t.

$$\exists \mathbf{B}_{k+1} \in \partial f(\mathbf{W}_{k+1}), \quad \text{s.t.} \quad \|\mathbf{B}_{k+1}\|_F \leq C'' \|\mathbf{W}_{k+1} - \mathbf{W}_k\|_F, \quad \forall$$

- *uniform Kurdyka-Lojasiewicz inequality* in Definition 3.

then the sequence convergence of $\{\mathbf{W}_k\}$ can be obtained by taking similar arguments as in [1,2].

Before going to the main convergence proof, we firstly introduce some supporting materials.

3.1 Definitions and basic ingredients

Definition 1. The indicator function $\delta_+(\mathbf{x})$ of nonnegative constraint is defined as

$$\delta_+(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \geq 0 \\ \infty, & \text{otherwise} \end{cases}$$

Since problem (3) is nonsmooth and nonconvex, we need some generalized differential to characterize its optimality.

Definition 2. [1,3] Let $h : \mathbb{R}^d \rightarrow (-\infty, \infty]$ be a proper and lower semi-continuous function

- (i) the effective domain is defined as

$$\text{dom } h := \{\mathbf{u} \in \mathbb{R}^d : h(\mathbf{u}) < \infty\}.$$

- (ii) The (Fréchet) subdifferential ∂h of h at \mathbf{u} is defined by

$$\partial h(\mathbf{u}) = \left\{ \mathbf{z} : \liminf_{\mathbf{v} \rightarrow \mathbf{u}, \mathbf{v} \neq \mathbf{u}} \frac{h(\mathbf{v}) - h(\mathbf{u}) - \langle \mathbf{z}, \mathbf{v} - \mathbf{u} \rangle}{\|\mathbf{u} - \mathbf{v}\|} \geq 0 \right\}$$

for any $\mathbf{u} \in \text{dom } h$ and $\partial h(\mathbf{u}) = \emptyset$ if $\mathbf{u} \notin \text{dom } h$.

Remark. First, when $h(\mathbf{u})$ is differentiable at \mathbf{u} , the (Fréchet) subdifferential reduces to the simple derivative or gradient in multiple dimension. In this paper, we main consider the NMF problem (2), where the nonsmooth indicator function of the nonnegative constraint $\delta_+(\cdot)$ is subdifferentialble everywhere in its effective domain. Therefore, the subdifferential of the objective in (3) is simply given by: gradient of its smooth part + the subdifferential of the nonsmooth part $\delta_+(\cdot)$, where the ‘+’ represents Minkowski summation of sets. Finally, the (Fréchet) subdifferential is commonly used to measure the optimality. In particular, a necessary condition for optimality is $0 \in \partial h(\bar{\mathbf{u}})$, and such a point is called critical point of $h(\mathbf{u})$.

By Definition 2, the subdifferential of the objective function $f(\mathbf{U}, \mathbf{V})$ is given in the following lemma.

Lemma 5. *The subdifferential of f in Equation (3) is given by*

$$\begin{aligned}\partial_{\mathbf{U}} f(\mathbf{U}, \mathbf{V}) &= \nabla_{\mathbf{U}} g(\mathbf{U}, \mathbf{V}) + \partial \delta_+(\mathbf{U}), \\ \partial_{\mathbf{V}} f(\mathbf{U}, \mathbf{V}) &= \nabla_{\mathbf{V}} g(\mathbf{U}, \mathbf{V}) + \partial \delta_+(\mathbf{V}),\end{aligned}\tag{14}$$

where

$$\begin{aligned}\partial \delta_+(\mathbf{U}) &= \{\mathbf{S} \in \mathbb{R}^{n \times r} : \delta_+(\mathbf{U}') - \delta_+(\mathbf{U}) \geq \langle \mathbf{S}, \mathbf{U}' - \mathbf{U} \rangle, \forall \mathbf{U}' \in \mathbb{R}^{n \times r}\} \\ &= \{\mathbf{S} \in \mathbb{R}^{n \times r} : \mathbf{S} \leq \mathbf{0}, \mathbf{S} \odot \mathbf{U} = \mathbf{0}\}\end{aligned}\tag{15}$$

The following property states the geometry of objective function(including its constraints) around its critical points, which plays a key role in our sequel analysis.

Definition 3 (Kurdyka-Lojasiewicz (KL) property). [2, 4] *We say a proper semi-continuous function $h(\mathbf{u})$ satisfies Kurdyka-Lojasiewicz (KL) property, if $\bar{\mathbf{u}}$ is a critical point of $h(\mathbf{u})$, then there exist $\delta > 0$, $\theta \in [0, 1)$, $C_1 > 0$, s.t.*

$$|h(\mathbf{u}) - h(\bar{\mathbf{u}})|^\theta \leq C_1 \text{dist}(0, \partial h(\mathbf{u})), \quad \forall \mathbf{u} \in B(\bar{\mathbf{u}}, \delta)$$

Remark. We mention that the above KL property(also known as KL inequality) states the regularity of $h(\mathbf{u})$ around its critical point \mathbf{u} and the KL inequality trivially holds at non-critical point. A very large set of functions satisfy the KL inequality, as stated in [1, Theorem 5.1], for a proper lower semi-continuous function, it has KL property once it is semi-algebraic. And the semi-algebraic property of sets and functions is sufficiently general, including but never limited to any polynomials, any norm, quasi norm, ℓ_0 norm, smooth manifold, etc. For more discussions and examples, see [1, 5].

3.2 Proof of Theorem 3

Lemma 6 (Lipschitz continuous gradient). *The function*

$$g(\mathbf{U}, \mathbf{V}) = \frac{1}{2} \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_F^2 + \frac{\lambda}{2} \|\mathbf{U} - \mathbf{V}\|_F^2$$

has Lipschitz continuous gradient with the Lipschitz constant as $3B + 2\lambda + \|\mathbf{X}\|_F$ in any bounded ℓ_2 -norm ball $\{(\mathbf{U}, \mathbf{V}) : \|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 \leq B\}$ for any $B > 0$.

Proof of Lemma 6. Denote $\mathbf{W}_k := (\mathbf{U}_k, \mathbf{V}_k)$ and similar notations apply to \mathbf{W} and $g(\mathbf{W})$. To obtain the Lipschitz constant, it is equivalently to bound the spectral norm of the quadrature form of the Hessian $[\nabla^2 g(\mathbf{W})](\mathbf{D}, \mathbf{D})$ for any $\mathbf{D} := (\mathbf{D}_U, \mathbf{D}_V)$:

$$\begin{aligned}[\nabla^2 g(\mathbf{W})](\mathbf{D}, \mathbf{D}) &= \|\mathbf{U}\mathbf{D}_V^T + \mathbf{D}_U\mathbf{V}^T\|_F^2 - 2\langle \mathbf{X} - \mathbf{U}\mathbf{V}^T, \mathbf{D}_U\mathbf{D}_V^T \rangle + \lambda \|\mathbf{D}_V - \mathbf{D}_U\|_F^2 \\ &\leq 2\|\mathbf{U}\|_F^2 \|\mathbf{D}_V\|_F^2 + 2\|\mathbf{V}\|_F^2 \|\mathbf{D}_U\|_F^2 + 2(\|\mathbf{X}\|_F + \|\mathbf{U}\mathbf{V}^T\|_F) \underbrace{\|\mathbf{D}_U\mathbf{D}_V^T\|_F}_{\leq \|\mathbf{D}\|_F^2/2} + 2 \underbrace{(\lambda \|\mathbf{D}_U\|_F^2 + \lambda \|\mathbf{D}_V\|_F^2)}_{=\lambda \|\mathbf{D}\|_F^2} \\ &\leq (3\|\mathbf{U}\|_F^2 + 3\|\mathbf{V}\|_F^2 + \|\mathbf{X}\|_F + 2\lambda) \|\mathbf{D}\|_F^2 \leq (3B + \|\mathbf{X}\|_F + 2\lambda) \|\mathbf{D}\|_F^2.\end{aligned}$$

□

Remark. As each iterate $(\mathbf{U}_k, \mathbf{V}_k)$ lives in the ℓ_2 -norm ball with the radius as $\sqrt{B_0}$ by (12). We immediately have the function $g(\mathbf{W})$ has Lipschitz continuous gradient with the Lipschitz constant being $2B_0 + \lambda + \|\mathbf{X}\|_F$ around each $\mathbf{W}_k = (\mathbf{U}_k, \mathbf{V}_k)$.

We now show the descend property in the following lemma.

Proof of Lemma 3 Since all $\{(\mathbf{U}_k, \mathbf{V}_k)\}$ generated by Algorithm 1 are nonnegative, we identify that

$$\begin{aligned} f(\mathbf{U}_k, \mathbf{V}_k) - f(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}) &= g(\mathbf{U}_k, \mathbf{V}_k) - g(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}) \\ &= g(\mathbf{U}_k, \mathbf{V}_k) - g(\mathbf{U}_k, \mathbf{V}_{k+1}) + g(\mathbf{U}_k, \mathbf{V}_{k+1}) - g(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}) \end{aligned} \quad (16)$$

with g denoting the smooth differentiable part of f .

The main proof would consist of bounding $(g(\mathbf{U}_k, \mathbf{V}_k) - g(\mathbf{U}_k, \mathbf{V}_{k+1}))$ and $(g(\mathbf{U}_k, \mathbf{V}_{k+1}) - g(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}))$, respectively. We now bound $(g(\mathbf{U}_k, \mathbf{V}_k) - g(\mathbf{U}_k, \mathbf{V}_{k+1}))$. For that end, we identify that

$$\mathbf{V}_{k+1} = \arg \min_{\mathbf{V} \geq 0} g(\mathbf{U}_k, \mathbf{V})$$

As the indicator function of nonnegative constraint $\sigma_+(\mathbf{V})$ convex subdifferentiable for all \mathbf{V} in its effective domain including relative boundary, hence the following set valued subdifferential

$$\partial\sigma_+(\tilde{\mathbf{V}}) = \{\mathbf{S} \in \mathbb{R}^{n \times r} : \sigma_+(\mathbf{V}) \geq \sigma_+(\tilde{\mathbf{V}}) + \langle \mathbf{S}, \mathbf{V} - \tilde{\mathbf{V}} \rangle, \quad \forall \mathbf{V} \in \mathbb{R}^{n \times r}\}$$

is nonempty for all $\mathbf{V} \geq 0$.

Utilizing the nonnegativity of iterates $\mathbf{V} = \mathbf{V}_k$, $\tilde{\mathbf{V}} = \mathbf{V}_{k+1}$ gives

$$0 \geq \langle \mathbf{S}_{k+1}, \mathbf{V}_k - \mathbf{V}_{k+1} \rangle, \quad \forall \mathbf{S}_{k+1} \in \partial\delta_+(\mathbf{V}_{k+1})$$

Since the update means $\mathbf{V}_{k+1} = \arg \min_{\mathbf{V}} g(\mathbf{U}_k, \mathbf{V}) + \sigma_+(\mathbf{V})$, it can be seen from the first order optimality $\mathbf{0} \in \nabla_{\mathbf{V}} g(\mathbf{U}_k, \mathbf{V}_{k+1}) + \partial\sigma_+(\mathbf{V}_{k+1})$ that

$$\nabla_{\mathbf{V}} g(\mathbf{U}_k, \mathbf{V}_{k+1}) + \mathbf{S}_{k+1} = \mathbf{0}$$

multiplying $\mathbf{V}_k - \mathbf{V}_{k+1}$ on both sides in the above equation providing

$$\langle \nabla_{\mathbf{V}} g(\mathbf{U}_k, \mathbf{V}_{k+1}), \mathbf{V}_k - \mathbf{V}_{k+1} \rangle + \langle \mathbf{S}_{k+1}, \mathbf{V}_k - \mathbf{V}_{k+1} \rangle = \mathbf{0}$$

hence

$$\langle \nabla_{\mathbf{V}} g(\mathbf{U}_k, \mathbf{V}_{k+1}), \mathbf{V}_k - \mathbf{V}_{k+1} \rangle \geq 0 \quad (17)$$

Now combining Equation (17) and the Taylor expansion, we arrive at the desired result

$$\begin{aligned} g(\mathbf{U}_k, \mathbf{V}_k) &= g(\mathbf{U}_k, \mathbf{V}_{k+1}) + \langle \nabla_{\mathbf{V}} g(\mathbf{U}_k, \mathbf{V}_{k+1}), \mathbf{V}_k - \mathbf{V}_{k+1} \rangle \\ &\quad + \frac{1}{2} \int_0^1 \nabla_{\mathbf{V}\mathbf{V}}^2 g(\mathbf{U}_k, t\mathbf{V}_k + (1-t)\mathbf{V}_{k+1}) [\mathbf{V}_k - \mathbf{V}_{k+1}, \mathbf{V}_k - \mathbf{V}_{k+1}] dt \\ &\geq g(\mathbf{U}_k, \mathbf{V}_{k+1}) + \frac{\lambda}{2} \|\mathbf{V}_k - \mathbf{V}_{k+1}\|_F^2 \end{aligned}$$

where the last inequality is from the fact that $g(\mathbf{U}, \mathbf{V})$ is strongly convex on variable \mathbf{V} with modulus at least λ . This further implies

$$g(\mathbf{U}_k, \mathbf{V}_k) - g(\mathbf{U}_k, \mathbf{V}_{k+1}) \geq \frac{\lambda}{2} \|\mathbf{V}_k - \mathbf{V}_{k+1}\|_F^2.$$

Using similar argument, we have

$$g(\mathbf{U}_k, \mathbf{V}_{k+1}) - g(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}) \geq \frac{\lambda}{2} \|\mathbf{U}_k - \mathbf{U}_{k+1}\|_F^2.$$

We end the proof by plugging the above two inequalities back to Equation (16).

Lemma 7 (Convergence of objective function values). *Suppose the iterates sequence $\{(\mathbf{U}_k, \mathbf{V}_k)\}$ is generated by Algorithm 1, the following holds*

(a) *The objective function value sequence $\{f(\mathbf{U}_k, \mathbf{V}_k)\}$ is nonincreasing and it converges to some finite value:*

$$\lim_{k \rightarrow \infty} f(\mathbf{U}_k, \mathbf{V}_k) = f^* \quad (18)$$

for some nonnegative f^ depending on $(\mathbf{U}_0, \mathbf{V}_0)$.*

(b) *The difference between iterates sequence is convergent to zero, i.e.,*

$$\lim_{k \rightarrow \infty} \|\mathbf{U}_{k+1} - \mathbf{U}_k\|_F = 0, \quad \lim_{k \rightarrow \infty} \|\mathbf{V}_{k+1} - \mathbf{V}_k\|_F = 0. \quad (19)$$

Proof of Lemma 7. Denote $\mathbf{W}_k := (\mathbf{U}_k, \mathbf{V}_k)$ and similar notations apply to \mathbf{W} and $f(\mathbf{W})$. Then it follows from (13) that

$$\sum_{k=0}^{\infty} f(\mathbf{W}_k) - f(\mathbf{W}_{k+1}) \geq \frac{\lambda}{2} \sum_{k=0}^{\infty} \|\mathbf{W}_{k+1} - \mathbf{W}_k\|_F^2 \implies \sum_{k=0}^{\infty} \|\mathbf{W}_{k+1} - \mathbf{W}_k\|_F^2 \leq \frac{2f(\mathbf{W}_0)}{\lambda}$$

Now, we conclude the proof of (a) by identifying that the sequence $\{f(\mathbf{W}_k)\}$ is non-increasing and lower-bounded by zero. For proving (b), we recognize the series $\{\sum_{k=n}^{\infty} \|\mathbf{W}_{k+1} - \mathbf{W}_k\|_F^2\}_n$ is convergent, so $\lim_{k \rightarrow \infty} \|\mathbf{W}_{k+1} - \mathbf{W}_k\|_F = 0$. \square

Lemma 8 (Bounded iterates). *Suppose the iterates sequence $\{(\mathbf{U}_k, \mathbf{V}_k)\}$ is generated by Algorithm 1, then it lies in a bounded subset, i.e.,*

$$\|\mathbf{U}_k\|_F^2 + \|\mathbf{V}_k\|_F^2 \leq B_0$$

with B_0 defined in Equation (12).

Proof of Lemma 8. The proof directly follows from Equation (12) in Lemma 2 and the sufficient decrease property in Lemma 3. \square

Lemma 9 (Bounded subdifferential). *Suppose the iterates sequence $\{\mathbf{W}_k = (\mathbf{U}_k, \mathbf{V}_k)\}$ is generated by Algorithm 1, then there exist $\mathbf{S}_{k+1} \in \partial_{\mathbf{U}} f(\mathbf{U}_{k+1}, \mathbf{V}_{k+1})$ and $\mathbf{D}_{k+1} \in \partial_{\mathbf{V}} f(\mathbf{U}_{k+1}, \mathbf{V}_{k+1})$ such that*

$$\left\| \begin{bmatrix} \mathbf{S}_{k+1} \\ \mathbf{D}_{k+1} \end{bmatrix} \right\|_F \leq (2B_0 + \lambda + \|\mathbf{X}\|_F) \|\mathbf{W}_{k+1} - \mathbf{W}_k\|_F. \quad (20)$$

Proof of Lemma 9. First from the optimality of \mathbf{V}_{k+1} , we have

$$\mathbf{0} \in \nabla_{\mathbf{V}} g(\mathbf{U}_k, \mathbf{V}_{k+1}) + \partial \sigma_+(\mathbf{V}_{k+1})$$

and noting that

$$\partial_{\mathbf{V}} f(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}) = \nabla_{\mathbf{V}} g(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}) + \partial \sigma_+(\mathbf{V}_{k+1}),$$

we hence can choose \mathbf{D}_{k+1} as follows

$$\mathbf{D}_{k+1} = \nabla_{\mathbf{V}} g(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}) - \nabla_{\mathbf{V}} g(\mathbf{U}_k, \mathbf{V}_{k+1}).$$

Then by the Lipschitz property of g in Lemma 6 and the boundedness property $\|\mathbf{U}_k\|_F^2 + \|\mathbf{V}_k\|_F^2 \leq B_0$ in Lemma 8, we have

$$\|\mathbf{D}_{k+1}\|_F \leq (2B_0 + \lambda + \|\mathbf{X}\|_F) \|\mathbf{U}_{k+1} - \mathbf{U}_k\|_F.$$

On the other hand, we actually can choose $\mathbf{S}_{k+1} = \mathbf{0}$ by recognizing that

$$\mathbf{U}_{k+1} = \arg \min_{\mathbf{U}} f(\mathbf{U}, \mathbf{V}_{k+1}).$$

Thus, we have $\|(\mathbf{S}_{k+1}, \mathbf{D}_{k+1})\|_F \leq (2B_0 + \lambda + \|\mathbf{X}\|_F) \|\mathbf{W}_{k+1} - \mathbf{W}_k\|_F$. \square

We denote $\mathcal{C}(\mathbf{W}_0)$ as the collection of the limit points of sequence $\{\mathbf{W}_k\}$ (which may depend on the initialization \mathbf{W}_0). The following two results demonstrate some useful properties and optimality of $\mathcal{C}(\mathbf{W}_0)$.

Lemma 10. *Suppose the iterates sequence $\{(\mathbf{U}_k, \mathbf{V}_k)\}$ is generated by Algorithm 1, then*

$$f(\mathbf{U}^*, \mathbf{V}^*) = f^*, \quad \forall (\mathbf{U}^*, \mathbf{V}^*) \in \mathcal{C}(\mathbf{U}_0, \mathbf{V}_0).$$

In other words, all the limiting points of $\mathcal{C}(\mathbf{U}_0, \mathbf{V}_0)$ share the same function value, that is equal to f^ , the limit of the objective function value sequence of Algorithm 1.*

Proof of Lemma 10. We first extract an arbitrary convergent subsequence $\{\mathbf{W}_{k_m}\}_m$ with limit \mathbf{W}^* . By the definition of the algorithm we have

$$\mathbf{U}_k \geq 0, \quad \mathbf{V}_k \geq 0, \quad \forall k \geq 0.$$

Furthermore,

$$\mathbf{U}^* \geq 0, \quad \mathbf{V}^* \geq 0.$$

Hence,

$$\lim_{m \rightarrow \infty} \delta_+(\mathbf{U}_{k_m}) = 0, \quad \lim_{m \rightarrow \infty} \delta_+(\mathbf{V}_{k_m}) = 0.$$

Taking limit on the sequence of the function values $\{f(\mathbf{W}_{k_m})\}$, we have

$$f^* = \lim_{m \rightarrow \infty} f(\mathbf{W}_{k_m}) = \lim_{m \rightarrow \infty} g(\mathbf{W}_{k_m}) + \lim_{m \rightarrow \infty} (\delta_+(\mathbf{U}_{k_m}) + \delta_+(\mathbf{V}_{k_m})) = g(\lim_{m \rightarrow \infty} \mathbf{W}_{k_m}) = g(\mathbf{W}^*) = f(\mathbf{U}^*, \mathbf{V}^*).$$

Here the first equality comes from Equation (18), the third equality is due to the continuity of the smooth part $g(\mathbf{W})$ in (3) and the last equality is because $\mathbf{U}^* \geq 0, \mathbf{V}^* \geq 0$. Therefore, the proof is concluded by noting that \mathbf{W}^* is an arbitrary limiting point in $\mathcal{C}(\mathbf{W}_0)$. \square

Lemma 11. *Suppose the iterates sequence $\{\mathbf{W}_k\}$ is generated by Algorithm 1, then each element $\mathbf{W}^* := (\mathbf{U}^*, \mathbf{V}^*) \in \mathcal{C}(\mathbf{W}^0)$ is a critical point of (3) and $\mathcal{C}(\mathbf{W}^0)$ is a nonempty, compact and connected set, and satisfies*

$$\lim_{k \rightarrow \infty} \text{dist}(\mathbf{W}_k, \mathcal{C}(\mathbf{W}_0)) = 0.$$

Proof of Lemma 11. It follows from Lemma 7 and Lemma 9 that then there exist $\mathbf{S}_{k+1} \in \partial_{\mathbf{U}} f(\mathbf{U}_{k+1}, \mathbf{V}_{k+1})$ and $\mathbf{D}_{k+1} \in \partial_{\mathbf{V}} f(\mathbf{U}_{k+1}, \mathbf{V}_{k+1})$ such that $\|(\mathbf{S}_{k+1}, \mathbf{D}_{k+1})\|_F \leq (2B_0 + \lambda + \|\mathbf{X}\|_F) \|\mathbf{U}_{k+1} - \mathbf{U}_k\|_F$ with the right hand side converging to zero as k goes to infinity, implying that

$$\lim_{k \rightarrow \infty} (\mathbf{S}_k, \mathbf{D}_k) = \mathbf{0}.$$

Since $\{\mathbf{W}_k\}$ is bounded by Lemma 8, by the Bolzano-Weierstrass theorem, we now take an arbitrary convergent subsequence $\{\mathbf{W}_{k_m}\}_m$ that converges to a point $\mathbf{W}^* \in \mathcal{C}(\mathbf{W}^0)$. By Lemma 5, we have

$$\mathbf{S}_{k_m} = \nabla_{\mathbf{U}} g(\mathbf{U}_{k_m}, \mathbf{V}_{k_m}) + \bar{\mathbf{S}}_{k_m}, \quad \bar{\mathbf{S}}_{k_m} \in \partial \delta_+(\mathbf{U}_{k_m}).$$

Since $\lim_{m \rightarrow \infty} \mathbf{S}_{k_m} = \mathbf{0}$, $\lim_{m \rightarrow \infty} \mathbf{W}_{k_m} = \mathbf{W}^*$, and $\nabla_{\mathbf{U}} g$ is continuous, we have $\{\bar{\mathbf{S}}_{k_m}\}$ is convergent. Denote by $\bar{\mathbf{S}}^* = \lim_{m \rightarrow \infty} \bar{\mathbf{S}}_{k_m}$. By the definition of $\bar{\mathbf{S}}_{k_m} \in \partial \delta_+(\mathbf{U}_{k_m})$, for any $\mathbf{U}' \in \mathbb{R}^{n \times r}$, we have

$$\delta_+(\mathbf{U}') - \delta_+(\mathbf{U}_{k_m}) \geq \langle \bar{\mathbf{S}}_{k_m}, \mathbf{U}' - \mathbf{U}_{k_m} \rangle.$$

Since $\lim_{m \rightarrow \infty} \delta_+(\mathbf{U}_{k_m}) = \delta_+(\mathbf{U}^*) = 0$, taking $m \rightarrow \infty$ for both sides of the above equation gives

$$\delta_+(\mathbf{U}') - \delta_+(\mathbf{U}^*) \geq \langle \bar{\mathbf{S}}^*, \mathbf{U}' - \mathbf{U}^* \rangle.$$

Since the above equation holds for any $\mathbf{U}' \in \mathbb{R}^{n \times r}$, we have $\bar{\mathbf{S}}^* \in \partial \delta_+(\mathbf{U}^*)$ and thus $\mathbf{0} = \nabla_{\mathbf{U}} g(\mathbf{U}^*, \mathbf{V}^*) + \bar{\mathbf{S}}^* \in \partial_{\mathbf{U}} f(\mathbf{W}^*)$. With similar argument, we get $\mathbf{0} \in \partial_{\mathbf{V}} f(\mathbf{W}^*)$ and thus

$$\mathbf{0} \in \partial f(\mathbf{W}^*).$$

since \mathbf{W}^* is an arbitrary elements in $\mathcal{C}(\mathbf{W}_0)$, implies each limiting point $\mathbf{W}^* = (\mathbf{U}^*, \mathbf{V}^*) \in \mathcal{C}(\mathbf{W}_0)$ is a critical point of (3).

Finally, by [1, Lemma 3.5] and identifying the sequence $\{\mathbf{W}_k\}$ is bounded and regular (i.e. $\lim_{k \rightarrow \infty} \|\mathbf{W}_{k+1} - \mathbf{W}_k\|_F = 0$), we obtain that the set of accumulation points $\mathcal{C}(\mathbf{W}_0)$ is a nonempty compact and connect set satisfying

$$\lim_{k \rightarrow \infty} \text{dist}(\mathbf{W}_k, \mathcal{C}(\mathbf{W}_0)) = 0.$$

□

Lemma 12 (Uniform KL property of f). *For arbitrary $(\mathbf{U}, \mathbf{V}) \in \mathcal{C}(\mathbf{U}_0, \mathbf{V}_0)$, we can uniformly find a set of constants $C_2 > 0, \delta > 0, \theta \in [0, 1)$ such that*

$$[f(\mathbf{W}) - f(\mathbf{W}^*)]^\theta \leq C_2 \text{dist}(0, \partial f(\mathbf{W}))$$

for all \mathbf{W} such that $\text{dist}(\mathbf{W}, \mathcal{C}(\mathbf{W}_0)) \leq \delta$.

Proof of Lemma 12. It is easy and straightforward to identify that $f(\mathbf{U}, \mathbf{V})$ satisfies the KL inequality at every point in its effective domain. From Lemma 11 we have the set $\mathcal{C}(\mathbf{W}_0)$ is a compact and connected set. Hence we can find finitely many balls $B(\mathbf{W}_i, r_i)$ and their union to cover

$$\mathcal{D} = \{(\mathbf{W}) : \text{dist}(\mathbf{W}, \mathcal{C}(\mathbf{W}_0)) \leq \delta\}$$

where each r_i is chosen such that the KL inequality holds true at each center and we can choose $c_i > 0, \theta_i \in [0, 1)$ that

$$[f(\mathbf{W}) - f(\mathbf{W}_i)]^{\theta_i} \leq c_i \text{dist}(0, \partial f(\mathbf{W})), \quad \forall \mathbf{W} \in B(\mathbf{W}_i, r_i).$$

Hence it is straightforward to verify

$$[f(\mathbf{W}) - f(\mathbf{W}^*)]^\theta \leq C_2 \text{dist}(0, \partial f(\mathbf{W})).$$

for all (\mathbf{W}) such that $\text{dist}(\mathbf{W}, \mathcal{C}(\mathbf{W}_0)) \leq \delta$, where $C_2 = \max\{c_i\}$ and $\theta = \max\{\theta_i\}$.

□

We now turn to the remaining proof of Theorem 3.

Global Convergence. We first assume there exists a finite \bar{k} for which $f(\mathbf{W}_{\bar{k}}) = f(\mathbf{W}^*)$. Then by the convergence of the function values $\{f(\mathbf{W}_k)\}$ with limit $f(\mathbf{W}^*)$ in Lemma 10, we have

$$f(\mathbf{W}^*) = f(\mathbf{W}_{\bar{k}}) = f(\mathbf{W}_{\bar{k}+1}) = \dots$$

Together with the sufficient decrease property in Lemma 3, this implies

$$\mathbf{W}_{\bar{k}} = \mathbf{W}_{\bar{k}+1} = \mathbf{W}_{\bar{k}+2} = \dots$$

establishing the convergence of the iterates $\{\mathbf{W}_k\}$ in finite steps.

Thus in the following, we assume $f(\mathbf{W}_k) > f(\mathbf{W}^*)$ for all $k > 0$. First of all, since $\lim_{k \rightarrow \infty} \text{dist}(\mathbf{W}_k, \mathcal{C}(\mathbf{W}_0)) = 0$ by Lemma 11, there exists a k_0 for which $\text{dist}(\mathbf{W}_k, \mathcal{C}(\mathbf{W}_0)) \leq \delta$ for all $k \geq k_0$ and any fixed $\delta > 0$. Hence

$$[f(\mathbf{W}_k) - f(\mathbf{W}^*)]^\theta \leq C_2 \text{dist}(0, \partial f(\mathbf{W}_k)), \quad \forall k \geq k_0. \quad (21)$$

In the subsequent analysis, we restrict to $k \geq k_0$. Construct a concave function $x^{1-\theta}$ for some $\theta \in [0, 1)$ with domain $x > 0$. Obviously, by the concavity, we have

$$x_2^{1-\theta} - x_1^{1-\theta} \geq (1-\theta)x_2^{-\theta}(x_2 - x_1), \quad \forall x_1 > 0, x_2 > 0$$

by replacing x_1 by $f(\mathbf{W}_{k+1}) - f(\mathbf{W}^*)$ and x_2 by $f(\mathbf{W}_k) - f(\mathbf{W}^*)$ and using the sufficient decrease property in Lemma 3, we have

$$\begin{aligned}
& [f(\mathbf{W}_k) - f(\mathbf{W}^*)]^{1-\theta} - [f(\mathbf{W}_{k+1}) - f(\mathbf{W}^*)]^{1-\theta} \\
& \geq (1-\theta) \frac{f(\mathbf{W}_k) - f(\mathbf{W}_{k+1})}{[f(\mathbf{W}_k) - f(\mathbf{W}^*)]^\theta} \\
& \geq \frac{\lambda(1-\theta)}{2C_2} \frac{\|\mathbf{W}_k - \mathbf{W}_{k+1}\|_F^2}{\text{dist}(0, \partial f(\mathbf{W}_k))}, \\
& \geq \frac{\lambda(1-\theta)}{2C_2(2B_0 + \lambda + \|\mathbf{X}\|_F)} \frac{\|\mathbf{W}_k - \mathbf{W}_{k+1}\|_F^2}{\|\mathbf{W}_k - \mathbf{W}_{k-1}\|_F} \\
& = \frac{\lambda(1-\theta)}{2C_2(2B_0 + \lambda + \|\mathbf{X}\|_F)} \left(\frac{\|\mathbf{W}_k - \mathbf{W}_{k+1}\|_F^2}{\|\mathbf{W}_k - \mathbf{W}_{k-1}\|_F} + \|\mathbf{W}_k - \mathbf{W}_{k-1}\|_F - \|\mathbf{W}_k - \mathbf{W}_{k+1}\|_F \right) \\
& \geq \frac{\lambda(1-\theta)}{2C_2(2B_0 + \lambda + \|\mathbf{X}\|_F)} (2\|\mathbf{W}_k - \mathbf{W}_{k+1}\|_F - \|\mathbf{W}_k - \mathbf{W}_{k-1}\|_F)
\end{aligned}$$

where the second inequality follows from Equation (21) and the third inequality is by Lemma 9. So,

$$2\|\mathbf{W}_k - \mathbf{W}_{k+1}\|_F - \|\mathbf{W}_k - \mathbf{W}_{k-1}\|_F \leq \beta ([f(\mathbf{W}_k) - f(\mathbf{W}^*)]^{1-\theta} - [f(\mathbf{W}_{k+1}) - f(\mathbf{W}^*)]^{1-\theta})$$

with $\beta := \left(\frac{\lambda(1-\theta)}{2C_2(2B_0 + \lambda + \|\mathbf{X}\|_F)} \right)^{-1}$.

Summing the above inequalities up from some $\tilde{k} > k_0$ to infinity yields

$$\sum_{k=\tilde{k}}^{\infty} \|\mathbf{W}_k - \mathbf{W}_{k+1}\|_F \leq \|\mathbf{W}_{\tilde{k}} - \mathbf{W}_{\tilde{k}-1}\|_F + \beta [f(\mathbf{W}_{\tilde{k}}) - f(\mathbf{W}^*)]^{1-\theta} \quad (22)$$

implying

$$\sum_{k=\tilde{k}}^{\infty} \|\mathbf{W}_k - \mathbf{W}_{k+1}\|_F < \infty.$$

Following some standard arguments one can see that

$$\limsup_{t \rightarrow \infty, t_1, t_2 \geq t} \|\mathbf{W}_{t_1} - \mathbf{W}_{t_2}\|_F = 0$$

which implies that the sequence $\{\mathbf{W}_k\}$ is Cauchy, and hence convergent. Hence, the limit point set $\mathcal{C}(\mathbf{W}_0)$ is singleton \mathbf{W}^* , which is also a critical point of (3) by Lemma 11.

Convergence Rate. Towards that end, we first know from the above argument that $\{\mathbf{W}_k\}$ converges to some point \mathbf{W}^* , i.e., $\lim_{k \rightarrow \infty} \mathbf{W}^k = \mathbf{W}^*$. Then using Equation (22) and the triangle inequality, we obtain

$$\|\mathbf{W}_{\tilde{k}} - \mathbf{W}^*\|_F \leq \sum_{k=\tilde{k}}^{\infty} \|\mathbf{W}_k - \mathbf{W}_{k+1}\|_F \leq \|\mathbf{W}_{\tilde{k}} - \mathbf{W}_{\tilde{k}-1}\|_F + \beta [f(\mathbf{W}_{\tilde{k}}) - f(\mathbf{W}^*)]^{1-\theta} \quad (23)$$

which indicates the convergence rate of $\mathbf{W}_{\tilde{k}} \rightarrow \mathbf{W}^*$ is at least as fast as the speed that $\|\mathbf{W}_{\tilde{k}} - \mathbf{W}_{\tilde{k}-1}\|_F + \beta [f(\mathbf{W}_{\tilde{k}}) - f(\mathbf{W}^*)]^{1-\theta}$ tends to 0. In particular, the second term $\beta [f(\mathbf{W}_{\tilde{k}}) - f(\mathbf{W}^*)]^{1-\theta}$ can be controlled by combining Lemma 12 and Lemma 9:

$$\begin{aligned}
\beta [f(\mathbf{W}_{\tilde{k}}) - f(\mathbf{W}^*)]^{1-\theta} & \leq \beta [C_2 \text{dist}(0, \partial f(\mathbf{W}_{\tilde{k}}))]^{\frac{1-\theta}{\theta}} \\
& \leq \underbrace{\beta [C_2(2B_0 + \lambda + \|\mathbf{X}\|_F)]^{\frac{1-\theta}{\theta}}}_{:=\alpha} \|\mathbf{W}_{\tilde{k}} - \mathbf{W}_{\tilde{k}-1}\|_F^{\frac{1-\theta}{\theta}}
\end{aligned} \quad (24)$$

Plugging (24) back to (23), we then have

$$\sum_{k=\tilde{k}}^{\infty} \|\mathbf{W}_k - \mathbf{W}_{k+1}\|_F \leq \|\mathbf{W}_{\tilde{k}} - \mathbf{W}_{\tilde{k}-1}\|_F + \alpha \|\mathbf{W}_{\tilde{k}} - \mathbf{W}_{\tilde{k}-1}\|_F^{\frac{1-\theta}{\theta}}.$$

We divide the following analysis into two cases based on the value of the KL exponent θ .

- *Case I: $\theta \in [0, \frac{1}{2}]$.* This case means $\frac{1-\theta}{\theta} \geq 1$. We define $P_{\tilde{k}} = \sum_{i=\tilde{k}}^{\infty} \|\mathbf{W}_{i+1} - \mathbf{W}_i\|_F$,

$$P_{\tilde{k}} \leq P_{\tilde{k}-1} - P_{\tilde{k}} + \alpha \left[P_{\tilde{k}-1} - P_{\tilde{k}} \right]^{\frac{1-\theta}{\theta}}. \quad (25)$$

Since $P_{\tilde{k}-1} - P_{\tilde{k}} \rightarrow 0$, there exists a positive integer k_1 such that $P_{\tilde{k}-1} - P_{\tilde{k}} < 1$, $\forall \tilde{k} \geq k_1$. Thus,

$$P_{\tilde{k}} \leq (1 + \alpha) (P_{\tilde{k}-1} - P_{\tilde{k}}), \quad \forall \tilde{k} \geq \max\{k_0, k_1\},$$

which implies that

$$P_{\tilde{k}} \leq \rho \cdot P_{\tilde{k}-1}, \quad \forall \tilde{k} \geq \max\{k_0, k_1\}, \quad (26)$$

where $\rho = \frac{1+\alpha}{2+\alpha} \in (0, 1)$. This together with (23) gives the linear convergence rate

$$\|\mathbf{W}_k - \mathbf{W}^*\|_F \leq \mathcal{O}(\rho^{k-\bar{k}}), \quad \forall k \geq \bar{k}. \quad (27)$$

where $\bar{k} = \max\{k_0, k_1\}$.

- *Case II: $\theta \in (1/2, 1)$.* This case means $\frac{1-\theta}{\theta} \leq 1$. Based on the former results, we have

$$P_{\tilde{k}} \leq (1 + \alpha) \left[P_{\tilde{k}-1} - P_{\tilde{k}} \right]^{\frac{1-\theta}{\theta}}, \quad \forall \tilde{k} \geq \max\{k_0, k_1\}.$$

We now run into the same situation as in [2, Theorem 2](after equation (13)) and [6, Theorem 2](after equation (30)), hence following a similar argument gives

$$P_{\tilde{k}}^{\frac{1-2\theta}{1-\theta}} - P_{\tilde{k}-1}^{\frac{1-2\theta}{1-\theta}} \geq \zeta, \quad \forall k \geq \bar{k}$$

for some $\zeta > 0$. Then repeating and summing up the above inequality from $\bar{k} = \max\{k_0, k_1\}$ to any $k > \bar{k}$, we can conclude

$$P_{\tilde{k}} \leq \left[P_{\tilde{k}}^{\frac{1-2\theta}{1-\theta}} + \zeta(\tilde{k} - \bar{k}) \right]^{-\frac{1-\theta}{2\theta-1}} = \mathcal{O} \left((\tilde{k} - \bar{k})^{-\frac{1-\theta}{2\theta-1}} \right), \quad \forall \tilde{k} > \bar{k}.$$

Finally, the following sublinear convergence holds

$$\|\mathbf{W}_k - \mathbf{W}^*\|_F \leq \mathcal{O} \left((k - \bar{k})^{-\frac{1-\theta}{2\theta-1}} \right), \quad \forall k > \bar{k}. \quad (28)$$

We end this proof by commenting that both linear and sublinear convergence rate are closely related to the KL exponent θ at the critical point \mathbf{W}^* .

3.3 Proof sketch of Lemma 4 and Theorem 4

The proof of Lemma 4 and Theorem 4 is highly similar to that of Lemma 3 and Theorem 3, respectively, hence we just outline the sketch here. The main difference is to update one column by one column in Algorithm 2, but the key feature that the objective function at each update is strongly convex doesn't change. Hence we can employ a similar argument as in Lemma 3 to obtain the sufficient decreasing property of HALS.

- *Case I*: update \mathbf{u}_i , i.e. $(\mathbf{u}_1^{k+1}, \dots, \mathbf{u}_{i-1}^{k+1}, \mathbf{u}_i^k, \dots, \mathbf{u}_r^k, \mathbf{V}_k) \rightarrow (\mathbf{u}_1^{k+1}, \dots, \mathbf{u}_i^{k+1}, \mathbf{u}_{i+1}^k, \dots, \mathbf{u}_r^k, \mathbf{V}_k)$, we have

$$f(\mathbf{u}_1^{k+1}, \dots, \mathbf{u}_i^k, \dots, \mathbf{u}_r^k, \mathbf{V}_k) - f(\mathbf{u}_1^{k+1}, \dots, \mathbf{u}_i^{k+1}, \dots, \mathbf{u}_r^k, \mathbf{V}_{k+1}) \geq \frac{\lambda}{2} \|\mathbf{u}_i^{k+1} - \mathbf{u}_i^k\|_2^2$$

- *Case II*: update \mathbf{v}_i , i.e. $(\mathbf{U}_{k+1}, \mathbf{v}_1^{k+1}, \dots, \mathbf{v}_{i-1}^{k+1}, \mathbf{v}_i^k, \dots, \mathbf{v}_r^k) \rightarrow (\mathbf{U}_{k+1}, \mathbf{v}_1^{k+1}, \dots, \mathbf{v}_i^{k+1}, \mathbf{v}_{i+1}^k, \dots, \mathbf{v}_r^k)$, we have

$$f(\mathbf{U}_{k+1}, \mathbf{v}_1^{k+1}, \dots, \mathbf{v}_{i-1}^{k+1}, \mathbf{v}_i^k, \dots, \mathbf{v}_r^k) - f(\mathbf{U}_{k+1}, \mathbf{v}_1^{k+1}, \dots, \mathbf{v}_i^{k+1}, \mathbf{v}_{i+1}^k, \dots, \mathbf{v}_r^k) \geq \frac{\lambda}{2} \|\mathbf{v}_i^{k+1} - \mathbf{v}_i^k\|_2^2$$

Unrolling the update from \mathbf{u}_1 to \mathbf{v}_r and summing them up, we can get the same sufficient decreasing inequality shown in Lemma 3. And the remaining proofs can be done following a similar way as in Theorem 3.

4 Additional Experimental Results

In Figure 1 and Figure 2, we show similar results as Figure 3 in the paper for larger truncated dataset MNIST_{train} , MNIST_{test} . Similarly, in Figure 3, we run ADMM 5000 iterations on ORL dataset.

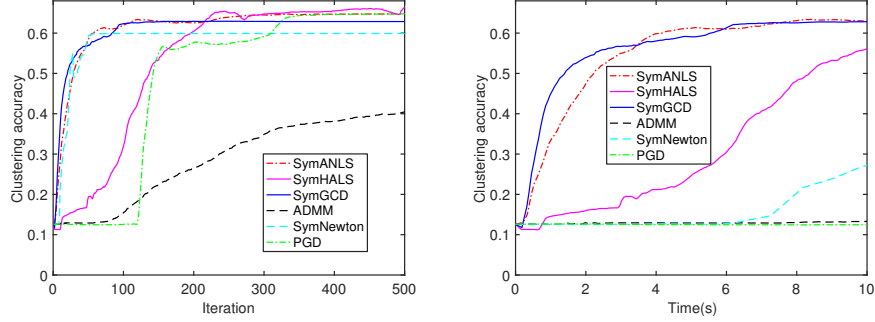


Figure 1: Image clustering quality on MNIST_{train} dataset, here $n = 3147, r = 10$.

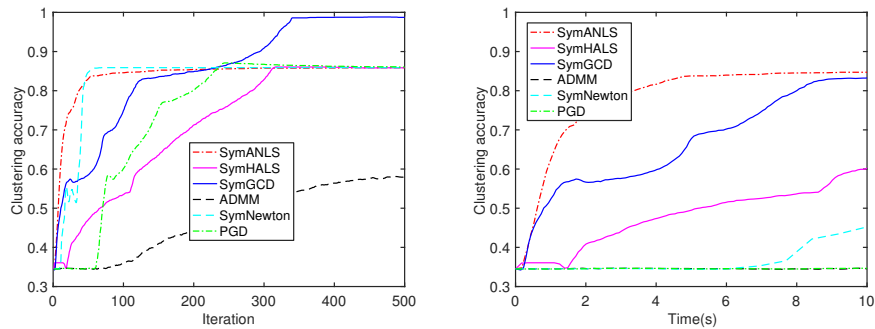


Figure 2: Image clustering quality on MNIST_{test} dataset, here $n = 3147, r = 3$.

5 Efficient Implementation of SymHALS

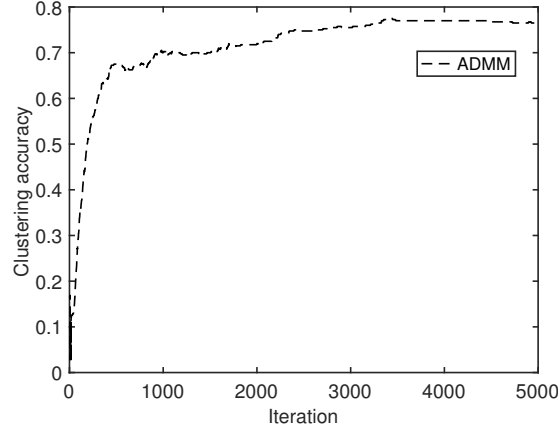


Figure 3: Image clustering quality of ADMM on ORL dataset, here $n = 400, r = 40$. ADMM needs roughly 3500 iterations to reach its maximum clustering rate!

Algorithm 3 SymHALS

Initialization: U_0, V_0 , iteration $k = 1$.

- 1: precompute residual $\mathbf{X}_1^k = \mathbf{X} - \mathbf{U}^{k-1}(\mathbf{V}^{k-1})^T$.
- 2: **while** stop criterion not meet **do**
- 3: **for** $i = 1 : r$ **do**
- 4: $\mathbf{X}_i^k = \mathbf{X}_i^k + \mathbf{u}_i^{k-1}(\mathbf{v}_i^{k-1})^T$
- 5: $\mathbf{u}_i^k = \arg \min_{\mathbf{u}_i \geq 0} \frac{1}{2} \|\mathbf{X}_i^k - \mathbf{u}_i(\mathbf{v}_i^{k-1})^T\|_F^2 + \frac{\lambda}{2} \|\mathbf{u}_i - \mathbf{v}_i^{k-1}\|_F^2 = \max \left(\frac{(\mathbf{X}_i^k + \lambda \mathbf{I}) \mathbf{v}_i^{k-1}}{\|\mathbf{v}_i^{k-1}\|_2^2 + \lambda}, 0 \right)$;
- 6: $\mathbf{v}_i^k = \arg \min_{\mathbf{v}_i \geq 0} \frac{1}{2} \|\mathbf{X}_i^k - \mathbf{u}_i^k \mathbf{v}_i^T\|_F^2 + \frac{\lambda}{2} \|\mathbf{u}_i^k - \mathbf{v}_i\|_F^2 = \max \left(\frac{(\mathbf{X}_i^k + \lambda \mathbf{I}) \mathbf{u}_i^k}{\|\mathbf{u}_i^k\|_2^2 + \lambda}, 0 \right)$;
- 7: update residual as $\mathbf{X}_i^k = \mathbf{X}_i^k - \mathbf{u}_i^k(\mathbf{v}_i^k)^T$.
- 8: **end for**
- 9: $\mathbf{X}_i^{k+1} = \mathbf{X}_i^k, k = k + 1$.
- 10: **end while**

Output: factorization $(\mathbf{U}_k, \mathbf{V}_k)$.

References

- [1] J. Bolte, S. Sabach, and M. Teboulle, “Proximal alternating linearized minimization for nonconvex and nonsmooth problems,” *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.
- [2] H. Attouch and J. Bolte, “On the convergence of the proximal algorithm for nonsmooth functions involving analytic features,” *Mathematical Programming*, vol. 116, no. 1, pp. 5–16, 2009.
- [3] R. T. Rockafellar, *Convex analysis*. Princeton university press, 2015.
- [4] J. Bolte, A. Daniilidis, and A. Lewis, “The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems,” *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 1205–1223, 2007.
- [5] H. Attouch, J. Bolte, and B. F. Svaiter, “Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gauss-seidel methods,” *Mathematical Programming*, vol. 137, no. 1-2, pp. 91–129, 2013.
- [6] Z. Zhu and X. Li, “Convergence analysis of alternating nonconvex projections,” *arXiv preprint arXiv:1802.03889*, 2018.

- [7] C.-J. Hsieh and I. S. Dhillon, “Fast coordinate descent methods with variable selection for non-negative matrix factorization,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1064–1072, ACM, 2011.