ICIC Express Letters
Volume **14**, Number **4**, April **2020**

ICIC International ©2020 ISSN 1881-803X
pp. **399–407**

A NOVEL NON-SMOOTH NON-NEGATIVE MATRIX FACTORIZATION ALGORITHM BASED ON ACCELERATED MIRROR PROJECTED METHODS

Xiangguang Dai, Yingyin Tao, Jiang Xiong and Yuming Feng*

School of Three Gorges Artificial Intelligence
Chongqing Three Gorges University
Wanzhou District, Chongqing 404100, P. R. China
*Corresponding author: yumingfeng25928@163.com

Received October 2019; accepted January 2020

Abstract. *This paper presents a novel non-smooth non-negative matrix factorization (nsNMF) algorithm for dimensionality reduction. Because the objective function of nsNMF is non-convex, we transform it into two symmetric convex problems and solve them iteratively. The optimal solution of each subproblem is obtained by solving a construed estimate sequence with accelerated mirror projected methods. We demonstrate that each subproblem has a fast convergence rate at $O(1/k^2)$. Examples on the image data demonstrate that our proposed algorithm produces the smaller factorization errors and the parser representations.*

Keywords: Non-smooth non-negative matrix factorization, Non-convex, Accelerated mirror projected methods, Dimensionality reduction, Sparse representation

1. **Introduction.** With the advent of big data, it is urgent to find some effective techniques for dimensionality reduction. Therefore, matrix factorization methods have attracted more interest as basic tools for data processing. The famous methods are principal component analysis [1], singular value decomposition [2] and vector quantization [3]. Their goals are finding several low-dimensional matrices whose product approximates the data matrix. Besides above mentioned methods, non-negative matrix factorization (NMF) can decompose a non-negative data matrix into two low-dimensional non-negative matrices. Suppose that a data matrix $V \in R^{m\times n}$ and $r << \min(m, n)$, NMF finds two non-negative matrices $A \in R^{m\times r}$ and $S \in R^{r\times n}$ to approximate V. Generally, Euclidean distance is used to measure the approximation error between V and $A \times S$ as follows:

$$F(A, S) = \frac{1}{2} \parallel V - AS \parallel_F^2, \quad \text{s.t. } A \geq 0, \quad S \geq 0, \tag{1}$$

where $\| \cdot \|_F$ denotes the Frobenius norm. The learned subspace S is very useful for clustering and classification [4, 5, 6]. It is obvious that an effective subspace should include the following aspects. Firstly, the potential structure or hidden features of the high-dimensional data should be captured. Secondly, the subspace should be sparse. Therefore, constraints or penalty terms imposed on A or S or both are proposed to solve mentioned problems [7, 8, 9]. Recently, non-smooth non-negative matrix factorization [10] performs satisfactorily in learning useful features. The achievement of nsNMF is to add a positive symmetric smooth matrix W into NMF instead of constraints on A or S. The nsNMF model can be described as

$$V = AWS, \tag{2}$$

DOI: 10.24507/icicel.14.04.399

where $W = (1-\theta)I + \frac{\theta}{n}\mathbf{1}\mathbf{1}^T$, $I \in R^{r\times r}$ is an identity matrix, $\mathbf{1} \in R^r$ is a vector of all ones, and $0 \le \theta < 1$. As θ becomes larger, A and S are sparser. Mathematically, nsNMF can be measured by Euclidean distance as follows:

$$F(A,S) = \frac{1}{2} \parallel V - AWS \parallel_F^2, \quad \text{s.t. } A \ge 0, \quad S \ge 0. \tag{3}$$

In fact, NMF algorithms can be improved to optimize nsNMF [11, 12, 13, 14, 15, 16, 18, 19]. The familiar algorithm is the multiplicative update rule (MULT) [11], which has a simple structure and comes to good results. However, it converges slowly when the data dimension becomes higher. Lin [12] proposed the block coordinate descent (BCD) [17] method to transform NMF into two convex problems and alternately optimized them until convergence. For each problem, the Armijo line search was proposed to determine the step length, and the projected gradient method was utilized to optimize each problem. Unfortunately, the Armijo line search spends much time. Based on the BCD method, Guan et al. [13] proposed the Lipchitz constant as the step length and each problem can be solved by Nesterov's gradient method. Although the proposed algorithm has a fast convergence rate at $O\left(\frac{1}{k^2}\right)$ in optimizing each problem, it is inefficient in high-dimensional data reduction. Recently, [18] and [19] proposed neural networks to search the global solution of non-negative matrix factorization. However, algorithms based on neural networks are convergent slowly than traditional algorithms. Previous algorithms utilized the BCD scheme by optimizing alternately A and S. Actually, another optimization scheme called hierarchical alternating least squares (HALS) was proposed to update each column of A and each row of S sequentially [14]. Gillis and Glineur [15] proposed the BCD scheme in updating the rows of S and the columns of W many times. Recently, utilizing random shuffling updates of the rows of S and the columns of A achieves better performances [16].

In this paper, we propose a novel algorithm (nsAMD) to optimize nsNMF. Firstly, we also transform nsNMF into two subproblems by BCD. For any $S^1 \ge 0$, nsAMD alternately solves

$$A^{t+1} = \underset{A^t \ge 0}{\arg\min} F\left(A^t, S^t\right) = \frac{1}{2}\left\|V^T - W S^{t^T} A^{t^T}\right\|_F^2 \tag{4}$$

and

$$S^{t+1} = \underset{S^t \ge 0}{\arg\min} F\left(A^{t+1}, S^t\right) = \frac{1}{2}\left\|V - A^{t+1} W S^t\right\|_F^2 \tag{5}$$

until convergence, where t denotes the iterative number. Secondly, to optimize (5), an estimate sequence $\phi_{k+1}(S)$ is proposed and constructed to satisfy $\min_{S\in\mathcal{X}}\{\phi_{k+1}(S)\} \ge F\left(A^{t+1}, S\right)$. Similarly, (4) can be optimized by the same way. Thirdly, we utilize the accelerated mirror descent methods to optimize the estimate sequences.

The remainder of this paper is organized as follows. In Section II, basic mathematical knowledge, definitions and lemmas of Nesterov's optimization methods are introduced. Three optimization sequences are proposed to optimize each problem and we present lemmas and theorems to prove each problem with a fast convergence rate at $O\left(1/k^2\right)$. Based on the analysis of Section II, we propose a novel algorithm named nsAMD to optimize non-smooth non-negative matrix factorization. In Section III, we compare our proposed algorithm with other nsNMF algorithms in terms of the convergence rate and the sparse representation. Finally, Section IV presents the conclusion and future work.

2. Theoretical Analysis.

In this section, some definitions and basic mathematical knowledge are applied to solve (3). It is obvious that (4) and (5) have a similar form. Hence, we only consider to solve (5), and (4) can be solved accordingly. Problem (5) is re-written into the following equivalent form:

$$\min_{S\in\chi} F(S) = \frac{1}{2}\left\|V - A^tWS\right\|_F^2. \tag{6}$$

- Gradient descent method. Given a step length L, the update rule $S \leftarrow S - \frac{1}{L}\nabla F(S)$ is the most familiar method in optimization. However, this update rule is considered as a special form of

$$S \leftarrow \arg\min_{Y\in\chi}\left\{\frac{1}{2}L\left\|Y - S\right\|_F^2 + < \nabla F(S), Y - S >\right\}. \tag{7}$$

- Dual averaging method. According to history solutions $S_0, S_1, \ldots, S_{k-1} \in \chi$, the next solution is

$$S_k \leftarrow \arg\min_{Y\in\chi}\left\{P(Y) + \sum_{i=1}^{k-1}\alpha_i(F(S_i) + < \nabla F(S_i), Y - S_i >)\right\}, \tag{8}$$

where $P(Y) = \frac{1}{2} \parallel Y \parallel_F^2$ and $\alpha_i > 0$ is the weight of the ith point.

- Mirror descent method. Given previous point $S \in \chi$, the next point can be optimized by

$$S \leftarrow \arg\min_{Y\in\chi}\{D(S) + \alpha < \nabla F(S), Y - S >\}, \tag{9}$$

where $D(S)$ is defined as a prox-function. Generally, it is a continuous differentiable and strong convex with the convexity parameter $\sigma > 0$. We assume $D(S_0) = 0$. Thus, for any $S \in \chi$, we have

$$D(S) \geq \frac{1}{2}\sigma \parallel S - S_0 \parallel_F^2. \tag{10}$$

Definition 2.1. *Suppose that $T_\chi(S) \in \chi$ is the optimal solution of (7), we obtain*

$$F\left(T_\chi(S)\right) \leq F(S) + \arg\min_{Y\in\chi}\left\{\frac{1}{2}L \parallel Y - S \parallel_F^2 + < \nabla F(S), Y - S >\right\}. \tag{11}$$

Definition 2.2. *Suppose that $\{\phi_k(S)\}_{k=0}^{\infty}$ is an estimate sequence. Thus, for a sequence $\{\lambda_k\}_{k=0}^{\infty}$ we have*

$$\phi_k(S) \leq (1 - \lambda_k)F(S) + \lambda_k\phi_0(S). \tag{12}$$

Lemma 2.1. *The objective function of $F(S)$ is convex.*

Proof: Suppose that $0 \leq \lambda \leq 1$, for any $S_1, S_2 \in \chi$ we have

$$\begin{aligned}
& F\left(\lambda S_1 + (1-\lambda)S_2\right) - \left(\lambda F(S_1) + (1-\lambda)F(S_2)\right) \\
= {} & \frac{1}{2}tr\left(\left(V - A^tW(\lambda S_1 + (1-\lambda)S_2)\right)^T \times \left(V - A^tW(\lambda S_1 + (1-\lambda)S_2)\right)\right) \\
& -\frac{\lambda}{2}\left(V - A^tWS_1\right)^T\left(X - A^tWS_1\right) - \frac{1-\lambda}{2}tr\left(\left(V - A^tWS_2\right)^T\left(V - A^tWS_2\right)\right) \\
= {} & -\frac{\lambda(1-\lambda)}{2}tr\left(\left(A^tW(S_1 - S_2)\right)^T\left(A^tW(S_1 - S_2)\right)\right) \\
= {} & -\frac{\lambda(1-\lambda)}{2} \parallel A^tW(S_1 - S_2) \parallel_F^2 \leq 0.
\end{aligned}$$

□

Lemma 2.2. *The gradient of $F(S)$ is Lipshitz continuous and the Lipshitz constant is $\left\|(A^tW)^TA^tW\right\|_F$.*

Proof: According to Lemma 2.1, $F(S)$ is differentiable and convex. Given any $S_1, S_2 \in \chi$, we obtain

$$\begin{aligned}
& \| \nabla F(S_1) - \nabla F(S_2) \|_F \\
= & \left\| \left(A^t W\right)^T A^t W S_1 - \left(A^t W\right)^T A^t W S_2 \right\|_F \\
= & \sum_{i=1}^{r} \sum_{j=1}^{n} \left| \sum_{k=1}^{r} \left(\left(A^t W\right)^T A^t W \right)_{ik} (S_1 - S_2)_{kj} \right| \\
\leq & \sum_{i=1}^{r} \sum_{j=1}^{n} \left(\sum_{k=1}^{r} \left| \left(\left(A^t W\right)^T A^t W \right)_{ik} \right| |(S_1 - S_2)_{kj}| \right) \\
= & \sum_{i=1}^{r} \sum_{k=1}^{r} \left| \left(\left(A^t W\right)^T A^t W \right)_{ik} \right| \sum_{j=1}^{n} \sum_{k=1}^{r} |(S_1 - S_2)_{kj}| \\
= & \left\| \left(A^t W\right)^T A^t W \right\|_F \| S_1 - S_2 \|_F .
\end{aligned}$$

□

Lemma 2.3. *If $F(S)$ is differentiable and Lipshitz continuous with the Lipshitz constant $L = \left\| (A^t W)^T A^t W \right\|_F$. For any $S_1, S_2 \in \chi$, we have*

$$0 \leq F(S_1) - F(S_2) - < \nabla F(S_2), S_1 - S_2 > \leq \frac{1}{2} L \| S_1 - S_2 \|_F^2 .$$

Proof: According to Taylor's expansion, $F(S_1)$ can be expanded by

$$F(S_1) = F(S_2) + < \nabla F(S_2), S_1 - S_2 > + \frac{1}{2} < S_1 - S_2, \nabla^2 F(S_2)(S_1 - S_2) > .$$

Thus,

$$\begin{aligned}
& F(S_1) - F(S_2) - < \nabla F(S_2), S_1 - S_2 > \\
= & \frac{1}{2} < S_1 - S_2, \nabla_S^2 F(S_2)(S_1 - S_2) > = \frac{1}{2} \left\| A^t W (S_1 - S_2) \right\|_F^2 .
\end{aligned}$$

By Lemma 2.2, we can obtain $0 \leq \| A^t W (S_1 - S_2) \|_F^2 \leq \frac{1}{2} \left\| (A^t W)^T A^t W \right\|_F \| S_1 - S_2 \|_F^2$.□

Lemma 2.4. *Given a sequence $\{S_k\}_{k=0}^{\infty}$, if the following inequality*

$$F(S_k) \leq \phi_k^* = \min_{S \in \chi} \phi_k(S) \tag{13}$$

holds, then $F(S_k) - F(S^*) \leq \lambda_k (\phi_0(S^*) - F(S^*))$.

Proof:

$$F(S_k) \leq \phi_k^* = \min_{S \in \chi} \phi_k(S) \leq \min_{S \in \chi} \{(1 - \lambda_k) F(S) + \lambda_k \phi_0(S)\} \leq (1 - \lambda_k) F(S^*) + \lambda_k \phi_0(S^*).$$

According to (13), the convergence rate of (5) can be easily computed. Next, we present lemmas and theorems to construct ϕ_k and λ_k. □

Lemma 2.5. *We assume $\alpha_k \in (0,1)$, $\alpha_k = \frac{\tau_{k+1}}{C_{k+1}}$ and $C_k = \sum_{i=0}^{k} \tau_i$. The following sequence*

$$\begin{aligned}
\phi_{k+1}(S) &= (1 - \alpha_k) \phi_k(S) + \alpha_k [F(S_{k+1}) + < \nabla F(S_{k+1}), S - S_{k+1} >] \\
&= \frac{C_0}{C_{k+1}} \phi_0(S) + \frac{1}{C_{k+1}} \sum_{i=1}^{k} \tau_i (F(S_i) + < \nabla F(S_i), S - S_i >)
\end{aligned} \tag{14}$$

is an estimate sequence.

Proof: Let $\lambda_k = \frac{1}{C_k}$. According to assumptions, we can easily obtain $\lambda_{k+1} = \lambda_k(1-\alpha_k)$. By Lemma 2.4, we have

$$\begin{aligned}\phi_{k+1}(S) &\leq (1-\alpha_k)\phi_k(S) + \alpha_k F(S)\\ &= (1-(1-\alpha_k)\lambda_k)F(S) + (1-\alpha_k)(\phi_k(S) - (1-\lambda_k)F(S))\\ &\leq (1-(1-\alpha_k)\lambda_k)F(S) + (1-\alpha_k)(\phi_0(S))\\ &= (1-\lambda_{k+1})F(S) + \lambda_{k+1}\phi_0(S).\end{aligned}$$

□

Lemma 2.6. *Suppose that*

$$\alpha_0 \in (0,1],\ \lambda_{k+1} = \lambda_k(1-\alpha_k),\ \lambda_{k+1} = \frac{1}{C_{k+1}},\ \frac{C_k}{C_{k+1}} \geq \alpha_k^2,\ C_k = \sum_{i=0}^{k}\tau_i,$$

$$\alpha_k = \frac{\tau_{k+1}}{C_{k+1}} \tag{15}$$

$$\phi_0(S) = \frac{1}{C_0}\left\{\frac{L}{\sigma}D(S) + \tau_0[F(S_0) + < \nabla F(S_0), S - S_0 >]\right\} \tag{16}$$

$$Z_k = \min_{S\in\chi}\phi_k(S) \tag{17}$$

$$S_{k+1} = \alpha_k Z_k + (1-\alpha_k)Y_k \tag{18}$$

$$Y_{k+1} = T_\chi(S_k). \tag{19}$$

Then, $\min_{S\in\chi}\{\phi_{k+1}(S)\} \geq F(Y_{k+1})$ *holds.*

Proof: By (16), $\phi_{k+1}(S)$ is equivalent to the following form:

$$\phi_{k+1}(S) = \frac{1}{C_{k+1}}\left(\frac{L}{\sigma}D(S) + \sum_{i=0}^{k+1}\tau_i(F(S_i) + < \nabla F(S_i), S - S_i >)\right). \tag{20}$$

Let $\tau_0 \in (0,1]$ and $Y_0 = T_\chi(S_0)$. For $k = 0$, one obtains

$$\begin{aligned}\min_{S\in\chi}\{\phi_0(S)\} &= \min_{S\in\chi}\frac{1}{C_0}\left\{\frac{L}{\sigma}D(S) + \tau_0[F(S_0) + < \nabla F(S_0), S - S_0 >]\right\}\\ &\geq \min_{S\in\chi}\left\{\frac{L}{2\tau_0} \parallel S - S_0 \parallel_F^2 + [F(S_0) + < \nabla F(S_0), S - S_0 >]\right\}\\ &\geq F(Y_0).\end{aligned}$$

We suppose $\nabla\phi_k(Z_k) = 0$ and $L = \left\|(A^tW)^T A^tW\right\|_F$, and $\phi_k(S)$ can be transformed by Taylor's expansion as

$$\begin{aligned}\phi_k(S) &= \phi_k(Z_k) + < \nabla\phi_k(Z_k), S - Z_k > + \frac{1}{2} < S - Z_k, \nabla^2\phi_k(Z_k)(S - Z_k) >\\ &\geq \phi_k(Z_k) + \frac{1}{2C_k}L \parallel S - Z_k \parallel_F^2 .\end{aligned} \tag{21}$$

Suppose that $\min_{H\in\chi}\{\phi_k(S)\} \geq F(Y_k)$ holds. According to (15), (18) and (21), we can get

$$\begin{aligned}&\min_{S\in\chi}\phi_{k+1}(S)\\ =\ &\min_{S\in\chi}\left\{\frac{C_k}{C_{k+1}}\phi_k(S) + \frac{\tau_{k+1}}{C_{k+1}}[F(S_{k+1}) + < \nabla F(S_{k+1}), S - S_{k+1} >]\right\}\\ \geq\ &\min_{S\in\chi}\left\{\frac{C_k}{C_{k+1}}\left[\phi_k(Z_k) + \frac{1}{2C_k}L \parallel S - Z_k \parallel_F^2\right] + \frac{\tau_{k+1}}{C_{k+1}}[F(S_{k+1})\right.\end{aligned}$$

$$
\begin{aligned}
& \quad + < \nabla F(S_{k+1}), S - S_{k+1} >] \Big\} \\
\geq\ & \min_{S\in\chi} \left\{ \frac{C_k}{C_{k+1}} Y_k + \frac{\tau_{k+1}}{C_{k+1}} [F(S_{k+1}) + < \nabla F(S_{k+1}), S - S_{k+1} >] + \frac{1}{2C_{k+1}} L \parallel S - Z_k \parallel_F^2 \right\} \\
=\ & \min_{S\in\chi} \left\{ F(S_{k+1}) + \alpha_k < \nabla F(S_{k+1}), S - Z_k > + \frac{1}{2C_{k+1}} L \parallel S - Z_k \parallel_F^2 \right\} \\
\geq\ & F(S_{k+1}) + \min_{S\in\chi} \left\{ \alpha_k < \nabla F(S_{k+1}), S - Z_k > + \frac{1}{2} \alpha_k^2 L \parallel S - Z_k \parallel_F^2 \right\}. \qquad (22)
\end{aligned}
$$

Let $Y = \alpha_k S + (1 - \alpha_k) Y_k$. By (18), we can easily obtain $Y - S_{k+1} = \alpha_k (S - Z_k)$. Thus, (22) can be simplified as

$$
\begin{aligned}
\min_{S\in\chi} \phi_{k+1}(S) & \geq F(S_{k+1}) + \min_{S\in\chi} \left\{ < \nabla F(S_{k+1}, Y - S_{k+1}) > + \frac{1}{2} L \parallel Y - S_{k+1} \parallel_F^2 \right\} \\
& \geq F(S_{k+1}) + F(Y_{k+1}) - F(S_{k+1}) \geq F(Y_{k+1})
\end{aligned}
$$

□

Note that τ_k, α_k and C_k are unknown, but they can be constructed by (15). Let $\alpha_k = \frac{2}{k+3}$, $\tau_k = \frac{k+1}{2}$ and $C_k = \frac{(k+1)(k+2)}{4}$. With above analysis, an optimal scheme is presented to optimize S. For $k \geq 0$ and $S_0 \geq 0$, we have

$$
Y_k = \arg\min_{Y\geq 0} \left\{ F(S_k) + < \nabla F(S_k), Y - S_k > + \frac{1}{2} L \parallel Y - S_k \parallel_F^2 \right\} \qquad (23)
$$

$$
Z_k = \arg\min_{Z\geq 0} \frac{1}{C_{k+1}} \left(\frac{L}{\sigma} D(Z, Z_0) + \sum_{i=0}^{k+1} \tau_i (F(S_i) + < \nabla F(S_i), Z - S_i >) \right) \qquad (24)
$$

$$
S_{k+1} = \alpha_k Z_k + (1 - \alpha_k) Y_k. \qquad (25)
$$

In the following, we will demonstrate that this scheme has a fast convergence rate at $O(1/k^2)$.

Theorem 2.1. *Suppose that $\{Y_k\}_{k=0}^{\infty}$ and $\{S_k\}_{k=0}^{\infty}$ are generated by (23), (24) and (25), one obtains*

$$
F(S_k) - F(S^*) \leq \frac{2LD(S^*)}{(k+1)(k+2)},
$$

where S^* is an optimal solution for (5).

Proof: By Lemma 2.4 and $\lambda_k = \frac{1}{C_k}$, we can get

$$
\begin{aligned}
& F(S_k) - \left(1 - \frac{1}{C_k}\right) F(S^*) \leq \frac{1}{C_k} \phi_0(S^*) \\
=\ & \frac{1}{C_k} \frac{1}{C_0} \left\{ \frac{L}{\sigma} D(S^*) + \tau_0 [F(S_0) + < \nabla F(S_0), S^* - S_0 >] \right\} \\
\leq\ & \frac{1}{C_k} \frac{1}{C_0} \frac{L}{\sigma} D(S^*) + \frac{1}{C_k} \frac{1}{C_0} \tau_0 F(S^*). \qquad (26)
\end{aligned}
$$

According to definitions of C_k and τ_k, (26) is simplified to $F(S_k) - F(S^*) \leq \frac{2LD(S^*)}{(k+1)(k+2)}$.□

Based on above analysis, we solve problem (3) by Algorithm 1, where ϵ_S, ϵ_A and ϵ are small tolerances.

Algorithm 1 nsAMD

```
Require: $V$, $A^t$, $S^t$, $\epsilon$
Ensure: $A$, $S$
  Initialization: $A^1 \geq 0$, $S^1 \geq 0$, $\epsilon_A$, $\epsilon_S$, $t \leftarrow 1$, $W$
  repeat
      1. $A^{t+1} \leftarrow AMD\left(V^T, S^{t^T}, A^{t^T}, W^T, \epsilon_A\right)$
      2. $S^{t+1} \leftarrow AMD(V, A^{t+1}, S^t, W, \epsilon_S)$
      3. $t \leftarrow t+1$
  until $\left|\frac{F(A^{t+1},S^{t+1})-F(A^t,S^t)}{F(A^t,S^t)}\right| \leq \epsilon$
  $A \leftarrow A^{t+1}$, $S \leftarrow S^{t+1}$
  function AMD($V$, $A^t$, $S^t$, $W$, $\epsilon_S$)
  Initialization: $Z_{-1} \leftarrow S_0 \leftarrow S^t$, $k \leftarrow 0$, $L \leftarrow \left\|(A^tW)^T(A^tW)\right\|_F$, $k \leftarrow 0$
     repeat
        1. $\nabla F(S_k) \leftarrow (A^tW)^T(A^tW)S_k - (A^tW)^TV$
        2. $Y_k \leftarrow P\left(S_k - \frac{1}{L}\nabla F(S_k)\right)$
        3. $Z_k \leftarrow P\left(Z_{k-1} - \frac{\tau_k}{L}\nabla F(S_k)\right)$
        4. $S_{k+1} \leftarrow \alpha_k Z_k + (1-\alpha_k)Y_k$
        5. $k \leftarrow k+1$
     until $\left|\frac{F(S_{k+1})-F(S_k)}{F(S_k)}\right| \leq \epsilon_S$
     return $S^{t+1} \leftarrow S_{k+1}$
  end function
```

3. **Simulations.** Four algorithms are compared in this section and we refer to them as nsAMD, nsMULT [11], nsPG [12] and nsNeNMF [13]. We compare them by the ORL dataset. Two indices are proposed to estimate their performances including the factorization error e and the sparsity degree d. e is defined by

$$e = \frac{1}{2} \parallel V - AWS \parallel_F^2 . \tag{27}$$

For any $V \in R^{m\times n}$, the sparsity degree of V can be described by

$$d_V = \frac{\sqrt{mn} - \sum_{i=1}^m \sum_{j=1}^n V_{ij} \Big/ \sqrt{\sum_{i=1}^m \sum_{j=1}^n V_{ij}^2}}{\sqrt{mn}-1}. \tag{28}$$

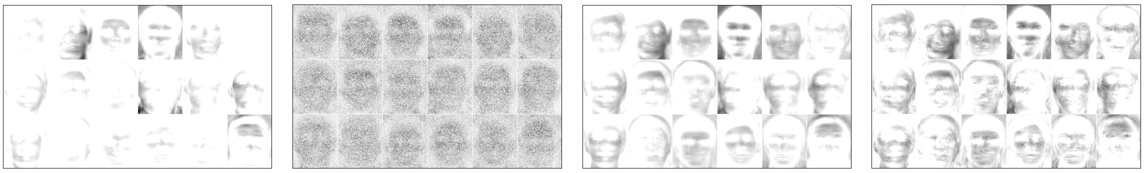
The ORL dataset can be downloaded from http://www.cl.cam.ac.uk/research/dtg/attarchive. To have fair comparisons, we do 10 experiments and report average results. Let $m = 10304$, $n = 400$, $r = 100$, $\epsilon_A = \epsilon_S = 10^{-4}$, $\theta = 0.9$ and a maximal number of the sub-iteration be 30.

Firstly, We present the factorization errors and the sparsity degrees obtained by all algorithms. Table 1 shows the average results in a short duration. Clearly, nsAMD achieves the best performances than other algorithms. However, this cannot reveal that our proposed algorithm is superior to other algorithms in learning bases.

Secondly, we mainly test whether each algorithm can learn bases in a short period. Figure 1(a)-1(d) show the learned bases by different algorithms. According to Table 1 and Figure 1, we observe that: 1) the larger factorization error leads to the worse performance in learning bases; 2) the smaller factorization error obtained by nsAMD means that the learned bases by nsAMD are sparser.

TABLE 1. Average results of $e(*10^9)$, S_{AS}, S_A and S_S on ORL

	nsAMD	**nsMULT**	**nsPG**	**nsNeNMF**
e	**0.1519**	0.2777	0.2034	0.2225
d_{AS}	**0.6704**	0.1884	0.4854	0.4517
d_A	**0.6666**	0.1791	0.4795	0.4456
d_S	**0.6717**	0.1514	0.4957	0.3725



(a) nsAMD (b) nsMULT (c) nsNeNMF (d) nsPG

FIGURE 1. Basis images on ORL under the time limit of 10 seconds

4. **Conclusion and Future Work.** This paper proposed a new efficient non-smooth non-negative matrix factorization algorithm called nsAMD. To optimize nsNMF, an iterative algorithm based on Nesterov's accelerated mirror descent methods is proposed and we demonstrate its convergence rate at $O(1/k^2)$. Experiments demonstrate that nsAMD is more effective to obtain the smaller factorization errors and the sparser bases.

Several topics should be discussed in the future work:

- The HALS optimization scheme and Nesterov's accelerated mirror descent methods should be considered to optimize nsNMF;
- A variable step length should be considered instead of the Lipshitz constant.

Acknowledgement. This work is supported by Foundation of Chongqing Municipal Key Laboratory of Institutions of Higher Education ([2017]3), Foundation of Chongqing Development and Reform Commission (2017[1007]), and Foundation of Chongqing Three Gorges University.

REFERENCES

[1] I. Jolliffe, *Principal Componente Analysis*, Springer-Verlag, 1986.

[2] A. Gersho and R. M. Gray, Vector quantization and signal compression, *Springer International*, vol.159, no.1, pp.407-485, 1992.

[3] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, 2nd Edition, Wiley-Interscience, 605 Third Avenue, New York, NY, United States, 2000.

[4] V. P. Pauca, F. Shahnaz, M. W. Berry and R. J. Plemmons, Text mining using non-negative matrix factorizations, *Siam International Conference on Data Mining*, Lake Buena Vista, Florida, USA, 2004.

[5] W. Xu, X. Liu and Y. H. Gong, Document clustering based on non-negative matrix factorization, *SIGIR 2003: Proc. of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, pp.267-273, 2003.

[6] D. Cai, X. F. He, J. W. Han and T. S. Huang, Graph regularized nonnegative matrix factorization for data representation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.33, no.8, pp.1548-1560, 2011.

[7] P. O. Hoyer, Non-negative sparse coding, *Proc. of the 2002 IEEE Workshop on Neural Networks for Signal*, pp.557-565, 2002.

[8] P. O. Hoyer, Non-negative matrix factorization with sparseness constraints, *Journal of Machine Learning Research*, pp.1457-1469, 2004.

[9] W. X. Liu, N. N. Zheng and X. F. Lu, Non-negative matrix factorization for visual coding, *Proc. of 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, vol.3, pp.III-293, IEEE, 2003.

[10] A. Pascualmontano, J. M. Carazo, K. Kochi, D. Lehmann and R. D. Pascualmarqui, Nonsmooth nonnegative matrix factorization (nsNMF), *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.28, no.3, pp.403-415, 2006.

[11] D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature*, vol.401, no.6755, pp.788-791, 1999.

[12] C. J. Lin, Projected gradient methods for nonnegative matrix factorization, *Neural Computation*, vol.19, no.10, pp.2756-2779, 2007.

[13] N. Guan, D. Tao, Z. Luo and B. Yuan, NeNMF: An optimal gradient method for nonnegative matrix factorization, *IEEE Trans. Signal Processing*, vol.60, no.6, pp.2882-2898, 2012.

[14] A. Cichocki and A. Phan, Fast local algorithms for large scale nonnegative matrix and tensor factorizations, *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol.92, no.3, pp.708-721, 2009.

[15] N. Gillis and F. Glineur, Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization, *Neural Computation*, vol.24, no.4, pp.1085-1105, 2012.

[16] S. W. N. Erichson, A. Mendible and J. Kutz, Randomized nonnegative matrix factorization, *Pattern Recognition Letters*, vol.104, pp.1-7, 2018.

[17] P. Tseng, Convergence of a block coordinate descent method for nondifferentiable minimization, *Journal of Optimization Theory and Applications*, vol.109, no.3, pp.475-494, 2001.

[18] H. Che and J. Wang, A nonnegative matrix factorization algorithm based on a discrete-time projection neural network, *Neural Networks*, vol.103, pp.63-71, 2018.

[19] J. Fan and J. Wang, A collective neurodynamic optimization approach to nonnegative matrix factorization, *IEEE Trans. Neural Networks and Learning Systems*, vol.28, no.10, pp.2344-2356, 2017.