Online Sparse Representation Clustering for Evolving Data Streams —Supplementary Document

Jie Chen, *Member, IEEE,* Shengxiang Yang, *Senior Member, IEEE,* Conor Fahy, Zhu Wang, Yi-nan Guo, and Yingke Chen

I. Proof of Theorem 1

In this section, we prove Theorem 1 in the paper regarding the optimization program

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{J},\mathbf{W}} \|\mathbf{Z}\|_0 + \frac{\lambda}{2}\left\|\mathbf{W}^T\mathbf{X} - \mathbf{W}^T\mathbf{D}\mathbf{J}\right\|_F^2 \\ &s.t. \quad \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}_m,\ \mathbf{J} = \mathbf{Z} - diag\left(\mathbf{Z}\right). \end{aligned} \tag{1}$$

Given the fixed $\mathbf{J}_{k+1}$ and $\mathbf{W}_{k+1}$, $\mathbf{Z}_{k+1}$ is updated by the following objective function:

$$\mathbf{Z}_{k+1} = \min_{\mathbf{Z}_{k+1}} \frac{1}{\mu_k}\|\mathbf{Z}_{k+1}\|_0 + \frac{1}{2}\left\|\mathbf{Z}_{k+1} - \left(\mathbf{J}_{k+1} + \frac{\mathbf{Y}_k}{\mu_k}\right)\right\|_F^2. \tag{2}$$

Given a positive number $\lambda > 0$, the hard thresholding operator $\mathcal{T}_{\sqrt{\lambda}}\left(\mathbf{Y}\right)$ is defined as follows [1]:

$$\mathcal{T}_{\sqrt{\lambda}}\left(x\right) = \begin{cases} 0, & if \quad |x| \le \sqrt{\lambda} \\ x, & if \quad |x| > \sqrt{\lambda} \end{cases} \tag{3}$$

where $\mathbf{Y} \in \mathbb{R}^{m\times n}$ is a matrix and x represents an element of $\mathbf{Y}$. The closed-form solution of (2) is obtained by using the operator $\mathcal{T}$:

$$\mathbf{Z}_{k+1} = \mathcal{T}_{\sqrt{\frac{1}{\mu_k}}}\left(\mathbf{J}_{k+1} + \frac{\mathbf{Y}_k}{\mu_k}\right). \tag{4}$$

Theorem 1 *The convergence condition $\|\mathbf{Z}_k - \mathbf{J}_k\|_{\max} < \varepsilon$ will eventually be satisfied as k increases if ρ and μ satisfy the following conditions:*

$$\rho > 2 \quad and \quad \mu > 0$$

where k represents the number of iterations and ε is a small positive number, e.g., $\varepsilon = 10^{-4}$.

J. Chen is with the College of Computer Science, Sichuan University, Chengdu 610065, China (E-mail: chenjie2010@scu.edu.cn).

S. Yang is with the School of Computer Science and Informatics, De Montfort University, Leicester LE1 9BH, U.K. (e-mail: syang@dmu.ac.uk).

C. Fahy are with the School of Computer Science and Informatics, De Montfort University, Leicester LE1 9BH, U.K. (e-mail: conor.fahy@dmu.ac.uk).

Z. Wang is with the Law School, Sichuan University, Chengdu 610065, China (E-mail: wangzhu@scu.edu.cn).

Y. Guo is with the School of Mechanical Electronic and Information Engineering, China University of Mining and Technology (Beijing), Beijing, 100083, China. E-mail: nanfly@126.com.

Y. Chen is with the Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, NE1 8ST, U. K. (E-mail: yke.chen@gmail.com).

Proof *According to* (3), $\mathbf{Z}_{k+1}$ *has a closed-form solution in* (2)*. Thus, we have:*

$$\|\mathbf{Z}_{k+1} - \mathbf{J}_{k+1}\|_{\max} = \left\|T_{\sqrt{\frac{1}{\mu_k}}}\left(\mathbf{J}_{k+1} + \frac{\mathbf{Y}_k}{\mu_k}\right) - \mathbf{J}_{k+1}\right\|_{\max}.$$

Suppose that $\rho > 2$ and $\mu > 0$, and we obtain $\mu_k \to \infty$ when $k \to \infty$ according to $\mu_k = \rho\mu_{k-1}$. This indicates that we will obtain

$$T_{\sqrt{\frac{1}{\mu_k}}}\left(\mathbf{J}_{k+1} + \frac{\mathbf{Y}_k}{\mu_k}\right) = \mathbf{J}_{k+1} + \frac{\mathbf{Y}_k}{\mu_k}$$

as k steadily increases. Thus, we have

$$\begin{aligned} &\|\mathbf{Z}_k - \mathbf{J}_k\|_{\max} - \|\mathbf{Z}_{k+1} - \mathbf{J}_{k+1}\|_{\max} \\ &= \left\|\frac{\mathbf{Y}_{k-1}}{\mu_{k-1}}\right\|_{\max} - \left\|\frac{\mathbf{Y}_k}{\mu_k}\right\|_{\max} \\ &= \frac{\rho\|\mathbf{Y}_{k-1}\|_{\max} - \|\mathbf{Y}_k\|_{\max}}{\mu_k} \\ &= \frac{\rho\|\mathbf{Y}_{k-1}\|_{\max} - \|\mathbf{Y}_{k-1} + \mu_{k-1}\left(\mathbf{Z}_k - \mathbf{J}_k\right)\|_{\max}}{\mu_k}. \end{aligned}$$

In addition, we obtain

$$\begin{aligned} &\rho\|\mathbf{Y}_{k-1}\|_{\max} - \left(\|\mathbf{Y}_{k-1}\|_{\max} + \|\mu_{k-1}\left(\mathbf{Z}_k - \mathbf{J}_k\right)\|_{\max}\right) \\ &= \left(\rho - 1\right)\|\mathbf{Y}_{k-1}\|_{\max} - \mu_{k-1}\cdot\left\|\frac{\mathbf{Y}_{k-1}}{\mu_{k-1}}\right\|_{\max} \\ &= \left(\rho - 2\right)\|\mathbf{Y}_{k-1}\|_{\max} \\ &> 0. \end{aligned}$$

It is easy to see that the following inequality holds:

$$\begin{aligned} &\|\mathbf{Y}_{k-1}\|_{\max} + \|\mu_{k-1}\left(\mathbf{Z}_k - \mathbf{J}_k\right)\|_{\max} \\ &\ge \|\mathbf{Y}_{k-1} + \mu_{k-1}\left(\mathbf{Z}_k - \mathbf{J}_k\right)\|_{\max}. \end{aligned}$$

Hence,

$$\|\mathbf{Z}_k - \mathbf{J}_k\|_{\max} - \|\mathbf{Z}_{k+1} - \mathbf{J}_{k+1}\|_{\max} > 0.$$

This means there exists a certain k with two conditions, i.e., $\rho > 2$ and $\mu_1 > 0$, such that the following inequality holds:

$$\|\mathbf{Z}_{k+1} - \mathbf{J}_{k+1}\|_{\max} \le \varepsilon$$

where $\mu = \mu_1$. Hence, convergence will eventually be achieved as k gradually increases if $\rho > 2$ and $\mu > 0$. □

II. Proof of Theorem 2

In this section, we prove Theorem 2 in the paper. We consider a general form of the $l_{2,1}$-norm optimization problem:

$$f\left(\mathbf{Z}^l\right) = \min_{\mathbf{Z}^l} \left\|\mathbf{Z}^l\right\|_{2,1} + \frac{\beta}{2}\left\|\mathbf{C}^l - \mathbf{C}^l\mathbf{Z}^l\right\|_F^2 \quad s.t. \quad \mathbf{diag}\left(\mathbf{Z}^l\right) = 0 \tag{5}$$

where $\beta > 0$ is a parameter. Problem (5) is a convex optimization problem. Let

$$\frac{\partial f\left(\mathbf{Z}^l\right)}{\partial \mathbf{W}} = 0 \tag{6}$$

and we have

$$\mathbf{Z}^l = \left(\frac{1}{\beta}\mathbf{\Sigma} + {\mathbf{C}^l}^T\mathbf{C}^l\right)^{-1}{\mathbf{C}^l}^T\mathbf{C}^l \tag{7}$$

where $\mathbf{Z}^l = \left[\mathbf{z}_1^r, \mathbf{z}_2^r, ..., \mathbf{z}_i^r, ..., \mathbf{z}_{n_1}^r\right]^T$ and $\mathbf{\Sigma} \in \mathbb{R}^{n_1 \times n_1}$ is a diagonal matrix whose diagonal entries are given by $\frac{1}{\left\|z_i^r\right\|_2}$.

To prove Theorem 2, we need to prove Lemma 1.

Lemma 1 *For two matrices $\mathbf{A} \in \mathbb{R}^{d\times m}$ and $\mathbf{B} \in \mathbb{R}^{m\times n}$, the following inequality holds:*

$$\left\|\mathbf{AB}\right\|_F^2 \le \left\|\mathbf{A}\right\|_F^2 \left\|\mathbf{B}\right\|_F^2.$$

Proof *First, we have*

$$\left\|\mathbf{AB}\right\|_F^2 = tr\left(\mathbf{A}^T\mathbf{ABB}^T\right)$$

by the definition of the trace function.

Second, we want to prove that

$$tr\left(\mathbf{A}^T\mathbf{ABB}^T\right) \le tr\left(\mathbf{A}^T\mathbf{A}\right) tr\left(\mathbf{BB}^T\right) \tag{8}$$

which implies that $\left\|\mathbf{AB}\right\|_F^2 \le \left\|\mathbf{A}\right\|_F^2 \left\|\mathbf{B}\right\|_F^2$. It is easy to see that $\mathbf{A}^T\mathbf{A}$ and $\mathbf{BB}^T$ are positive semidefinite and symmetric matrices. Using the singular value decomposition (SVD) results of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{BB}^T$, we obtain

$$\begin{aligned} \mathbf{A}^T\mathbf{A} &= \mathbf{U}_A\Sigma_{\mathbf{A}}\mathbf{U}_{\mathbf{A}}^T, \\ \mathbf{BB}^T &= \mathbf{U}_{\mathbf{B}}\Sigma_{\mathbf{B}}\mathbf{U}_{\mathbf{B}}^T, \end{aligned} \tag{9}$$

where $\mathbf{U}_A$ and $\mathbf{U}_B$ are unitary matrices, and $\Sigma_{\mathbf{A}}$ and $\Sigma_{\mathbf{B}}$ are diagonal matrices whose diagonal elements are singular values of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{BB}^T$, respectively. The singular values of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{BB}^T$ are all nonnegative. Then,

$$\begin{aligned} tr\left(\mathbf{A}^T\mathbf{ABB}^T\right) &= tr\left(\mathbf{U}_A\Sigma_{\mathbf{A}}\mathbf{U}_{\mathbf{A}}^T\mathbf{U}_{\mathbf{B}}\Sigma_{\mathbf{B}}\mathbf{U}_{\mathbf{B}}^T\right) \\ &\le tr\left(\mathbf{U}_{\mathbf{B}}^T\mathbf{U}_A\Sigma_{\mathbf{A}}\mathbf{U}_{\mathbf{A}}^T\mathbf{U}_{\mathbf{B}}\right) tr\left(\Sigma_{\mathbf{B}}\right) \\ &= tr\left(\mathbf{U}_A\Sigma_{\mathbf{A}}\mathbf{U}_{\mathbf{A}}^T\right) tr\left(\mathbf{U}_{\mathbf{B}}\Sigma_{\mathbf{B}}\mathbf{U}_{\mathbf{B}}^T\right) \\ &= tr\left(\mathbf{A}^T\mathbf{A}\right) tr\left(\mathbf{BB}^T\right). \end{aligned} \tag{10}$$

Hence,

$$\left\|\mathbf{AB}\right\|_F^2 \le \left\|\mathbf{A}\right\|_F^2 \left\|\mathbf{B}\right\|_F^2.$$

□

The proof of Theorem 2 is motivated by a feature selection method [2].

Theorem 2 *The objective value of Equation (7) will monotonically decrease until convergence to the global optimum of Problem (5).*

Proof *Suppose that $\mathbf{Z}_{t+1}^l$ is the global optimal solution to Problem (5), i.e.,*

$$\mathbf{Z}_{t+1}^l = \arg\min_{\mathbf{Z}^l \; diag\left(\mathbf{Z}^l\right)=0} \left\|\mathbf{Z}^l\right\|_{2,1} + \frac{\beta}{2}\left\|\mathbf{C}^l - \mathbf{C}^l\mathbf{Z}^l\right\|_F^2.$$

Problem (5) is a convex optimization problem, which indicates that

$$\begin{aligned} &\left\|\mathbf{Z}_{t+1}^l\right\|_{2,1} + \frac{1}{\beta}\left\|\mathbf{C}^l - \mathbf{C}^l\mathbf{Z}_{t+1}^l\right\|_F^2 \\ &\le \left\|\mathbf{Z}_{t+1}^l\right\|_{2,1} + \frac{1}{\beta}\left\|\mathbf{C}^l - \mathbf{C}^l\mathbf{Z}_t^l\right\|_F^2. \end{aligned}$$

Thus,

$$\left\|\mathbf{C}^l - \mathbf{C}^l\mathbf{Z}_{t+1}^l\right\|_F^2 \le \left\|\mathbf{C}^l - \mathbf{C}^l{\mathbf{Z}_t}^l\right\|_F^2.$$

According to Lemma 1, we have

$$\left\|\mathbf{I} - \mathbf{Z}_{t+1}^l\right\|_F^2 \le \left\|\mathbf{I} - \mathbf{C}^l{\mathbf{Z}_t}^l\right\|_F^2$$

where $\mathbf{I}$ is an identity of size $n_l \times n_l$. Then, we have the following inequality:

$$tr\left(\mathbf{Z}_{t+1}^l\left(\mathbf{Z}_{t+1}^l\right)^T - \mathbf{Z}_t^l\left(\mathbf{Z}_t^l\right)^T\right) \le tr\left(2\left(\mathbf{Z}_{t+1}^l - \mathbf{Z}_t^l\right)\right).$$

Using the constraint $diag\left(\mathbf{Z}^l\right) = 0$ in Problem (5), we obtain

$$tr\left(\mathbf{Z}_{t+1}^l\left(\mathbf{Z}_{t+1}^l\right)^T - \mathbf{Z}_t^l\left(\mathbf{Z}_t^l\right)^T\right) \le 0).$$

Then

$$\sum_{i=1}^n \left\|\left(\mathbf{z}^i\right)_{t+1}^l\right\|_2^2 \le \sum_{i=1}^n \left\|\left(\mathbf{z}^i\right)_t^l\right\|_2^2$$

where $\left(\mathbf{z}^i\right)_{t+1}^l$ and $\left(\mathbf{z}^i\right)_t^l$ are the i-th row vectors of $\mathbf{Z}_{t+1}^l$ and $\mathbf{Z}_t^l$, respectively. Hence,

$$\left\|\mathbf{Z}_{t+1}^l\right\|_{2,1} \le \left\|\mathbf{Z}_t^l\right\|_{2,1}.$$

This means that the objective value of Equation (7) will monotonically decrease at each iteration. At the $(t+1)$-th iteration, Equation (7) holds for given $\mathbf{Z}_{t+1}^l$ and $\mathbf{\Sigma}_{t+1}^l$. Consequently, the objective value of Equation (7) will converge to the global optimum of Problem (5). □

References

[1] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 629–654, Sept. 2008.

[2] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," in *Adv. Neural. Inf. Process. Syst.*, Vancouver, British Columbia, Canada, Dec. 2010, pp. 1813–1821.