Appendix For "One Pass Late Fusion Multi-view Clustering"

Xinwang Liu [1] **Li Liu** [1] **Qing Liao** [2] **Siwei Wang** [1] **Yi Zhang** [1] **Wenxuan Tu** [1] **Chang Tang** [3] **Jiyuan Liu** [1] **En Zhu** [1]

1. Summary of the Appendix

In this appendix, we provide the generalization analysis of the proposed algorithm and give the detailed proof.

2. The Theoretical Results

Generalization error for k-means clustering has been studied by fixing the centroids obtained in the training process and computing their generalization to unseen data (Maurer & Pontil, 2010; Liu et al., 2016). In this section, we study how the centroids obtained by the proposed OP-LFMVC generalizes onto test data by deriving its generalization bound.

We now define the error of OP-LFMVC. Let $\hat{\mathbf{C}} = [\hat{\mathbf{C}}_1, \cdots, \hat{\mathbf{C}}_k]$ be the learned matrix composed of the k centroids, $\hat{\boldsymbol{\beta}}$ the learned kernel weights and $\{\hat{\mathbf{W}}_p\}_{p=1}^m$ the transformation matrices learned by the proposed OP-LFMVC. By defining $\Theta = \{\mathbf{e}_1, \cdots, \mathbf{e}_k\}$, effective OP-LFMVC should make the following error small,

$$1 - \mathbb{E}_{\mathbf{x}}\left[\max_{\mathbf{y}\in\Theta}\left\langle\sum_{p=1}^m \hat{\beta}_p\hat{\mathbf{W}}_p^\top\mathbf{h}_p(\mathbf{x}), \hat{\mathbf{C}}\mathbf{y}\right\rangle\right], \quad (1)$$

where $\mathbf{h}_p(\mathbf{x})$ denotes the p-th partition vector corresponding to the p-th view of $\mathbf{x}$ with $\|\mathbf{h}_p(\mathbf{x})\| = 1$, and $\mathbf{e}_1, \cdots, \mathbf{e}_k$ form the orthogonal bases of $\mathbb{R}^k$. Intuitively, it says that the expected alignment between test points and their closest centroid should be high. In the following, we show how the proposed algorithm achieves this goal.

Let us define a function class first:

$$\mathcal{F} = \Big\{ f:\ \mathbf{x} \mapsto 1 - \max_{\mathbf{y}\in\Theta}\left\langle\sum_{p=1}^m \beta_p\mathbf{W}_p^\top\mathbf{h}_p(\mathbf{x}), \mathbf{C}\mathbf{y}\right\rangle \Big| \sum_{p=1}^m \beta_p^2 = 1,\ \beta_p \geq 0, \mathbf{C}\in\mathbb{R}^{k\times k},\ \mathbf{C}^\top\mathbf{C} = \mathbf{I}_k, \mathbf{W}_p \in \mathbb{R}^{k\times k},\ \mathbf{W}_p^\top\mathbf{W}_p = \mathbf{I}_k,\ \forall p \Big\}. \quad (2)$$

*Equal contribution [1]School of Computer, National University of Defense Technology. [2]Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen. [3]School of Computer Science, China University of Geosciences. Correspondence to: Xinwang Liu <xinwangliu@nudt.edu.cn>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

We have the following claim on the generalization error bound for the proposed OP-LFMVC.

Theorem 1. *For any $\delta > 0$, with probability at least $1-\delta$, the following inequality holds for all $f \in \mathcal{F}$,*

$$\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})] \leq \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i) + \frac{\sqrt{\pi/2}k}{\sqrt{n}} + 2\sqrt{\frac{\log 1/\delta}{2n}}. \quad (3)$$

3. Proof of Theorem 1

In the following, we give the detailed proof of Theorem 1. For i.i.d. given samples $\{\mathbf{x}_i\}_{i=1}^n$, OP-LFMVC minimizes the empirical error, i.e.,

$$1 - \frac{1}{n}\sum_{i=1}^n \max_{\mathbf{y}_i\in\Theta}\left\langle\sum_{p=1}^m \beta_p\mathbf{W}_p^\top\mathbf{h}_p(\mathbf{x}_i), \mathbf{C}\mathbf{y}_i\right\rangle, \quad (4)$$

where $\Theta = \{\mathbf{e}_1, \cdots, \mathbf{e}_k\}$ and $\mathbf{e}_1, \cdots, \mathbf{e}_k$ form the orthogonal bases of $\mathbb{R}^k$.

Let

$$\hat{R}(\mathbf{C}, \boldsymbol{\beta}, \{\mathbf{W}_p\}_{p=1}^m) = 1 - \frac{1}{n}\sum_{i=1}^n \max_{\mathbf{y}_i\in\Theta}\left\langle\sum_{p=1}^m \beta_p\mathbf{W}_p^\top\mathbf{h}_p(\mathbf{x}_i), \mathbf{C}\mathbf{y}_i\right\rangle. \quad (5)$$

Our proof idea is to upper the following bound

$$\sup_{\mathbf{C},\boldsymbol{\beta},\{\mathbf{W}_p\}_{p=1}^m}\left(\mathbb{E}\left[\hat{R}(\mathbf{C}, \boldsymbol{\beta}, \{\mathbf{K}_p\}_{p=1}^m)\right] - \hat{R}(\mathbf{C}, \boldsymbol{\beta}, \{\mathbf{W}_p\}_{p=1}^m)\right), \quad (6)$$

and then upper bound the term $\hat{R}(\mathbf{C}, \boldsymbol{\beta}, \{\mathbf{W}_p\}_{p=1}^m)$ by the proposed objective.

Let us define a function class $\mathcal{F}$ first:

$$\mathcal{F} = \Big\{ f:\ \mathbf{x} \mapsto 1 - \max_{\mathbf{y}\in\Theta}\left\langle\sum_{p=1}^m \beta_p\mathbf{W}_p^\top\mathbf{h}_p(\mathbf{x}), \mathbf{C}\mathbf{y}\right\rangle \Big| \sum_{p=1}^m \beta_p^2 = 1,\ \beta_p \geq 0, \mathbf{C}\in\mathbb{R}^{k\times k},\ \mathbf{C}^\top\mathbf{C} = \mathbf{I}_k, \mathbf{W}_p \in \mathbb{R}^{k\times k},\ \mathbf{W}_p^\top\mathbf{W}_p = \mathbf{I}_k,\ \forall p \Big\}. \quad (7)$$

Then, Eq. (6) can be rewritten as,

$$\sup_{f\in\mathcal{F}}\left(\mathbb{E}\left[f(\mathbf{x})\right] - \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)\right). \quad (8)$$

It is not difficult to check that

$$\begin{aligned}
&\left\langle \sum_{p=1}^{m} \beta_p \mathbf{W}_p^\top \mathbf{h}_p(\mathbf{x}), \mathbf{C}\mathbf{y} \right\rangle \\
&= \mathbf{y}^\top \mathbf{C}^\top \left(\sum_{p=1}^{m} \beta_p \mathbf{W}_p^\top \mathbf{h}_p(\mathbf{x}) \right), \\
&= \left(\frac{1}{|\mathbf{C}_l|} \sum_{\mathbf{x}_j \in \mathbf{C}_l} \sum_{p=1}^{m} \beta_p \mathbf{W}_p^\top \mathbf{h}_p(\mathbf{x}_j) \right)^\top \left(\sum_{p=1}^{m} \beta_p \mathbf{W}_p^\top \mathbf{h}_p(\mathbf{x}) \right), \\
&= \frac{1}{|\mathbf{C}_l|} \sum_{\mathbf{x}_j \in \mathbf{C}_l} \sum_{p,q=1}^{m} \beta_p \beta_q \mathbf{h}_p^\top(\mathbf{x}_j) \mathbf{W}_p \mathbf{W}_q^\top \mathbf{h}_q(\mathbf{x}), \\
&\leq \frac{1}{|\mathbf{C}_l|} \sum_{\mathbf{x}_j \in \mathbf{C}_l} \sum_{p,q=1}^{m} \beta_p \beta_q \\
&\leq \sum_{p,q=1}^{m} \frac{1}{2}(\beta_p^2 + \beta_q^2) = 1.
\end{aligned} \tag{9}$$

By the same way, one can readily prove that $-1 \leq \left\langle \sum_{p=1}^{m} \beta_p \mathbf{W}_p^\top \mathbf{h}_p(\mathbf{x}), \mathbf{C}\mathbf{y} \right\rangle \leq 1$. As a result, we have $f(\mathbf{x}) \leq 2$.

By exploiting McDiarmid's concentration inequality, we have the following theorem.

Theorem 2. *For any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $f \in \mathcal{F}$:*

$$\mathbb{E}\left[f(\mathbf{x})\right] - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) \leq 2\mathfrak{R}_n(\mathcal{F}) + 2\sqrt{\frac{\log 1/\delta}{2n}}, \tag{10}$$

where

$$\mathfrak{R}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum\nolimits_{i=1}^{n} \gamma_i f(\mathbf{x}_i)\right] \tag{11}$$

and $\gamma_1, \ldots, \gamma_n$ are i.i.d. Rademacher random variables uniformly distributed from $\{-1, 1\}$.

Now, we are going to upper bound $\mathfrak{R}_n(\mathcal{F})$. Since there is a maximization function in f, it is not easy to directly upper $\mathfrak{R}_n(\mathcal{F})$. Similar to the proof method in (Maurer & Pontil, 2010), we upper bound it by introducing Gaussian complexities:

$$\mathfrak{G}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum\nolimits_{i=1}^{n} \gamma_i f(\mathbf{x}_i)\right], \tag{12}$$

where $\gamma_1, \ldots, \gamma_n$ are i.i.d. Gaussian random variables with zero mean and unit standard deviation.

The following two lemmas (Maurer & Pontil, 2010) will be used in our proof.

Lemma 1.

$$\mathfrak{R}_n(\mathcal{F}) \leq \sqrt{\pi/2}\mathfrak{G}_n(\mathcal{F}). \tag{13}$$

Lemma 2. *Let $G_f = \sum_{i=1}^{n} \gamma_i G(\mathbf{x}_i, f)$ and $H_f = \sum_{i=1}^{n} \gamma_i H(\mathbf{x}_i, f)$ be two zero mean, separable Gaussian processes. If for all $f_1, f_2 \in \mathcal{F}$,*

$$\mathbb{E}\left[(G_{f_1} - G_{f_2})^2\right] \leq \mathbb{E}[(H_{f_1} - H_{f_2})^2]. \tag{14}$$

Then,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} G_f\right] \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} H_f\right]. \tag{15}$$

In our case, let

$$G_f = G_{\boldsymbol{\beta},\mathbf{C}} = \sum_{i=1}^{n} \gamma_i \left(1 - \max_{\mathbf{y}_i \in \Theta} h_{\boldsymbol{\beta}}^\top(\mathbf{x}_i)\mathbf{C}\mathbf{y}_i\right) \tag{16}$$

and

$$H_f = H_{\boldsymbol{\beta},\mathbf{C}} = \sum_{i=1}^{n} h_{\boldsymbol{\beta}}^\top(\mathbf{x}_i) \sum_{l=1}^{k} \gamma_{il} \mathbf{C}\mathbf{e}_l. \tag{17}$$

where $h_{\boldsymbol{\beta}}(\mathbf{x}) = \sum_{p=1}^{m} \beta_p \mathbf{W}_p^\top \mathbf{h}_p(\mathbf{x})$.

we are going to prove that

$$\mathbb{E}_\gamma\left[(G_{\boldsymbol{\beta}_1,\mathbf{C}_1} - G_{\boldsymbol{\beta}_2,\mathbf{C}_2})^2\right] \leq \mathbb{E}_\gamma\left[(H_{\boldsymbol{\beta}_1,\mathbf{C}_1} - H_{\boldsymbol{\beta}_2,\mathbf{C}_2})^2\right]. \tag{18}$$

Specifically, for any $f_1, f_2 \in \mathcal{F}$, we have

$$\begin{aligned}
&\left[\left(1 - \max_{\mathbf{y}\in\Theta} h_{\boldsymbol{\beta}_1}^\top(\mathbf{x})\mathbf{C}_1\mathbf{y}\right) - \left(1 - \max_{\mathbf{y}\in\Theta} h_{\boldsymbol{\beta}_2}^\top(\mathbf{x})\mathbf{C}_2\mathbf{y}\right)\right]^2 \\
&= \left(\max_{\mathbf{y}\in\Theta} h_{\boldsymbol{\beta}_1}^\top(\mathbf{x})\mathbf{C}_1\mathbf{y} - \max_{\mathbf{y}\in\Theta} h_{\boldsymbol{\beta}_2}^\top(\mathbf{x})\mathbf{C}_2\mathbf{y}\right)^2 \\
&\leq \left(\max_{\mathbf{y}\in\Theta} \left(h_{\boldsymbol{\beta}_1}^\top(\mathbf{x})\mathbf{C}_1\mathbf{y} - h_{\boldsymbol{\beta}_2}^\top(\mathbf{x})\mathbf{C}_2\mathbf{y}\right)\right)^2 \\
&= \left(\max_{\mathbf{y}\in\Theta} \left(h_{\boldsymbol{\beta}_1}^\top(\mathbf{x})\mathbf{C}_1 - h_{\boldsymbol{\beta}_2}^\top(\mathbf{x})\mathbf{C}_2\right)\mathbf{y}\right)^2 \\
&= \max_{\mathbf{y}\in\Theta} \left(\sum_{l=1}^{k} y_l \left(h_{\boldsymbol{\beta}_1}^\top(\mathbf{x})\mathbf{C}_1 - h_{\boldsymbol{\beta}_2}^\top(\mathbf{x})\mathbf{C}_2\right)\mathbf{e}_l\right)^2 \\
&\leq \sum\nolimits_{l=1}^{k} \left(\left(h_{\boldsymbol{\beta}_1}^\top(\mathbf{x})\mathbf{C}_1 - h_{\boldsymbol{\beta}_2}^\top(\mathbf{x})\mathbf{C}_2\right)\mathbf{e}_l\right)^2,
\end{aligned} \tag{19}$$

where the last inequality holds because $\sum_{l=1}^{k} y_l = 1$.

Thus, we have

$$\begin{aligned}
&\mathbb{E}_\gamma\left[\left(G_{\boldsymbol{\beta}_1,\mathbf{C}_1}-G_{\boldsymbol{\beta}_2,\mathbf{C}_2}\right)^2\right]\\
=&\mathbb{E}_\gamma\left[\left(\sum_{i=1}^n\gamma_i\left[\left(1-\max_{\mathbf{y}_i\in\Theta}h_{\boldsymbol{\beta}_1}^\top(\mathbf{x}_i)\mathbf{C}_1\mathbf{y}_i\right)\right.\right.\right.\\
&\left.\left.\left.-\left(1-\max_{\mathbf{y}_i\in\Theta}h_{\boldsymbol{\beta}_2}^\top(\mathbf{x}_i)\mathbf{C}_2\mathbf{y}_i\right)\right]\right)^2\right]\\
=&\sum_{i=1}^n\left(\max_{\mathbf{y}_i\in\Theta}h_{\boldsymbol{\beta}_1}^\top(\mathbf{x}_i)\mathbf{C}_1\mathbf{y}_i-\max_{\mathbf{y}_i\in\Theta}h_{\boldsymbol{\beta}_2}^\top(\mathbf{x}_i)\mathbf{C}_2\mathbf{y}_i\right)^2\\
\leq&\sum_{i=1}^n\sum_{l=1}^k\left(\left(h_{\boldsymbol{\beta}_1}^\top(\mathbf{x}_i)\mathbf{C}_1-h_{\boldsymbol{\beta}_2}^\top(\mathbf{x}_i)\mathbf{C}_2\right)\mathbf{e}_l\right)^2\\
=&\mathbb{E}_\gamma\left[\left(H_{\boldsymbol{\beta}_1,\mathbf{C}_1}-H_{\boldsymbol{\beta}_2,\mathbf{C}_2}\right)^2\right].
\end{aligned}\tag{20}$$

Using Hölder's inequality and Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}}H_f\right]&=\mathbb{E}_\gamma\left[\sup_{\boldsymbol{\beta},\mathbf{C}}\sum_{i=1}^n\sum_{l=1}^k\gamma_{il}h_{\boldsymbol{\beta}}^\top(\mathbf{x}_i)\mathbf{C}\mathbf{e}_l\right]\\
&\leq\mathbb{E}_\gamma\left[\sum_{l=1}^k\left|\sum_{i=1}^n\gamma_{il}\right|\right]\\
&\leq k\sqrt{n}.
\end{aligned}\tag{21}$$

Combining Lemmas 1 and 2, Eqs. (12), (20), and (21), we have

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{F})&\leq\frac{1}{n}\sqrt{\pi/2}\mathbb{E}[\sup_{f\in\mathcal{F}}G_{\boldsymbol{\beta},\mathbf{C}}]\\
&\leq\frac{1}{n}\sqrt{\pi/2}\mathbb{E}\left[\sup_{f\in\mathcal{F}}H_{\boldsymbol{\beta},\mathbf{C}}\right]\\
&\leq\frac{1}{n}\sqrt{\pi/2}k\sqrt{n}\\
&=\frac{\sqrt{\pi/2}k}{\sqrt{n}}.
\end{aligned}$$

Putting the above inequality into Theorem 2, with probability at least $1-\delta$, the following holds for all $f\in\mathcal{F}$:

$$\mathbb{E}\left[f(\mathbf{x})\right]\leq\frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)+\frac{\sqrt{\pi/2}k}{\sqrt{n}}+2\sqrt{\frac{\log 1/\delta}{2n}}.\tag{22}$$

This completes the proof.

References

Liu, T., Tao, D., and Xu, D. Dimensionality-dependent generalization bounds for k-dimensional coding schemes. *Neural computation*, 28(10):2213–2249, 2016.

Maurer, A. and Pontil, M. k-dimensional coding schemes in Hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.