Appendix of “SimpleMKKM: Simple Multiple Kernel K-means”

Xinwang Liu, *Senior Member, IEEE*

1 Summary of the Appendix

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

2 The Generalization Analysis

Let $\hat{\mathbf{C}} = [\hat{\mathbf{C}}_1, \cdots, \hat{\mathbf{C}}_k]$ be the learned matrix composed of the k centroids and $\hat{\boldsymbol{\gamma}}$ the learned kernel weights by the proposed SimpleMKKM, where $\hat{\mathbf{C}}_v = \frac{1}{|\hat{\mathbf{C}}_v|}\sum_{j\in\hat{\mathbf{C}}_v}\phi_{\hat{\boldsymbol{\gamma}}}(\mathbf{x}_j), 1 \leq c \leq k$. By defining $\Theta = \{\mathbf{e}_1, \cdots, \mathbf{e}_k\}$, effective SimpleMKKM clustering should make the following error small

$$1 - \mathbb{E}_{\mathbf{x}}\left[\max_{\mathbf{y}\in\Theta}\langle\phi_{\hat{\boldsymbol{\gamma}}}(\mathbf{x}), \hat{\mathbf{C}}\mathbf{y}\rangle_{\mathcal{H}^k}\right], \tag{1}$$

where $\phi_{\hat{\boldsymbol{\gamma}}}(\mathbf{x}) = [\hat{\boldsymbol{\gamma}}_1\phi_1^\top(\mathbf{x}), \cdot, \hat{\boldsymbol{\gamma}}_m\phi_m^\top(\mathbf{x})]^\top$ is the learned feature map associated with the kernel function $K_{\hat{\boldsymbol{\gamma}}}(\cdot,\cdot)$ and $\mathbf{e}_1, \cdots, \mathbf{e}_k$ form the orthogonal bases of $\mathbb{R}^k$. Intuitively, it says the expected alignment between test points and their closest centroid should be high. We show how the proposed algorithm achieves this goal.

Let us define a function class first:

$$\begin{aligned}\mathcal{F} =\Big\{ f:\ & \mathbf{x} \mapsto 1 - \max_{\mathbf{y}\in\Theta}\langle\phi_{\boldsymbol{\gamma}}(\mathbf{x}), \mathbf{C}\mathbf{y}\rangle_{\mathcal{H}^k}\Big|\boldsymbol{\gamma}^\top\mathbf{1}_m = 1,\\ & \gamma_p \geq 0, \mathbf{C} \in \mathcal{H}^k,\ |K_p(\mathbf{x}, \tilde{\mathbf{x}})| \leq b,\ \forall p, \forall \mathbf{x} \in \mathcal{X}\Big\},\end{aligned} \tag{2}$$

where $\mathcal{H}^k$ stands for the multiple kernel Hilbert space.

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

$$\mathbb{E}\left[f(\mathbf{x})\right] \leq \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) + \frac{\sqrt{\pi/2}bk}{\sqrt{n}} + (1+b)\sqrt{\frac{\log 1/\delta}{2n}}. \tag{3}$$

3 Proof of Theorem ??

In the following, we give the detailed proof of Theorem **??**. For an i.i.d. given sample $\{\mathbf{x}_i\}_{i=1}^n$, SimpleMKKM algorithm is to minimize an empirical error, i.e.,

$$1 - \frac{1}{n}\sum_{i=1}^{n}\max_{\mathbf{y}_i\in\Theta}\langle\phi_{\boldsymbol{\gamma}}(\mathbf{x}_i), \mathbf{C}\mathbf{y}_i\rangle_{\mathcal{H}^k}, \tag{4}$$

where $\phi_{\boldsymbol{\gamma}}(\mathbf{x}) = [\gamma_1\phi_1^\top(\mathbf{x}), \cdots, \gamma_m\phi_m^\top(\mathbf{x})]^\top$ is the feature map associated with the kernel function $K_{\boldsymbol{\gamma}}(\cdot,\cdot)$ and $\Theta = \{\mathbf{e}_1, \cdots, \mathbf{e}_k\}$ in which $\mathbf{e}_1, \cdots, \mathbf{e}_k$ form the orthogonal bases of $\mathbb{R}^k$.

Let

$$\hat{R}(\mathbf{C}, \boldsymbol{\gamma}, \{\mathbf{K}_p\}_{p=1}^m) = 1 - \frac{1}{n}\sum_{i=1}^{n}\max_{\mathbf{y}_i\in\Theta}\langle\phi_{\boldsymbol{\gamma}}(\mathbf{x}_i), \mathbf{C}\mathbf{y}_i\rangle_{\mathcal{H}^k}. \tag{5}$$

Our proof idea is to upper bound

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

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

We assume that the kernel mapping of each kernel is upper bounded, i.e., every entry of $\mathbf{K}_p$ $(p \in \{1, \cdots, m\})$, is no larger than b. Let us define a function class first:

$$\begin{aligned}\mathcal{F} =\Big\{ f:\ & \mathbf{x} \mapsto 1 - \max_{\mathbf{y}\in\Theta}\langle\phi_{\boldsymbol{\gamma}}(\mathbf{x}), \mathbf{C}\mathbf{y}\rangle_{\mathcal{H}^k}\Big|\boldsymbol{\gamma}^\top\mathbf{1}_m = 1, \gamma_p \geq 0,\\ & \mathbf{C} \in \mathcal{H}^k,\ |K_p(\mathbf{x}, \tilde{\mathbf{x}})| \leq b,\ \forall p, \forall \mathbf{x},\ \tilde{\mathbf{x}} \in \mathcal{X}\Big\},\end{aligned} \tag{7}$$

where $\mathcal{H}^k$ stands for the multiple kernel Hilbert space.

Then, Eq. (**??**) becomes

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

It is obvious that

$$\begin{aligned}\phi_{\boldsymbol{\gamma}}^\top(\mathbf{x})\phi_{\boldsymbol{\gamma}}(\tilde{\mathbf{x}}) &= \sum\nolimits_{p=1}^{m}\gamma_p^2\phi_p^\top(\mathbf{x}^{(p)})\phi_p(\tilde{\mathbf{x}}^{(p)})\\ &= \sum\nolimits_{p=1}^{m}\gamma_p^2 K_p(\mathbf{x}^{(p)}, \tilde{\mathbf{x}}^{(p)})\\ &\geq -b\sum\nolimits_{p=1}^{m}\gamma_p^2 \geq -b\sum\nolimits_{p=1}^{m}\gamma_p\\ &= -b.\end{aligned} \tag{9}$$

In the same way, it is easy to prove $-b \leq \phi_{\boldsymbol{\gamma}}^\top(\mathbf{x})\phi_{\boldsymbol{\gamma}}(\tilde{\mathbf{x}}) \leq b$. For $\mathbf{x}$ in v-th cluster,

$$\begin{aligned}&\langle\phi_{\boldsymbol{\gamma}}(\mathbf{x}), \mathbf{C}\mathbf{y}\rangle_{\mathcal{H}}\\ &= \phi_{\boldsymbol{\gamma}}^\top(\mathbf{x})\left(\frac{1}{|\mathbf{C}_v|}\sum\nolimits_{i\in\mathbf{C}_v}\phi_{\boldsymbol{\gamma}}(\mathbf{x}_i)\right)\\ &= \sum_{p=1}^{m}\gamma_p^2\left(\frac{1}{|\mathbf{C}_v|}\sum\nolimits_{i\in\mathbf{C}_v}\phi_p^\top(\mathbf{x}_i)\phi_p(\mathbf{x})\right)\\ &\geq -b\sum_{p=1}^{m}\gamma_p^2\ \geq\ -b\sum_{p=1}^{m}\gamma_p\ \geq -b.\end{aligned} \tag{10}$$

- *X. Liu is with College of Computer, National University of Defense Technology, Changsha, 410073, China. E-mail: xinwangliu@nudt.edu.cn.*

Manuscript received February 19, 2022.

As a result, we have $f(\mathbf{x}, \tilde{\mathbf{x}}) \leq 1 + b$.

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 $f \in \mathcal{F}$:*

$$\mathbb{E}\left[f(\mathbf{x})\right] - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) \leq 2\mathfrak{R}_n(\mathcal{F}) + (1+b)\sqrt{\frac{\log 1/\delta}{2n}}, \tag{11}$$

where

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

and $\sigma_1, \ldots, \sigma_n$ 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 [**?**], we upper bound it by introducing Gaussian complexities:

$$\mathfrak{G}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}\left[\sup\nolimits_{f\in\mathcal{F}} \sum\nolimits_{i=1}^{n} \beta_i f(\mathbf{x}_i)\right], \tag{13}$$

where $\beta_1, \ldots, \beta_n$ are i.i.d. Gaussian random variables with zero mean and unit standard deviation.

The following two lemmas [**?**] will be used in our proof.

Lemma 1.

$$\mathfrak{R}_n(\mathcal{F}) \leq \sqrt{\pi/2}\,\mathfrak{G}_n(\mathcal{F}). \tag{14}$$

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

$$\mathbb{E}\left[(G_{f_1} - G_{f_2})^2\right] \leq \mathbb{E}[(H_{f_1} - H_{f_2})^2]. \tag{15}$$

Then,

$$\mathbb{E}\left[\sup\nolimits_{f\in\mathcal{F}} G_f\right] \leq \mathbb{E}\left[\sup\nolimits_{f\in\mathcal{F}} H_f\right]. \tag{16}$$

In our case, let

$$G_{\boldsymbol{\gamma},\mathbf{C}} = \sum_{i=1}^{n} \beta_i \left(1 - \max_{\mathbf{y}_i\in\Theta} \langle \phi_{\boldsymbol{\gamma}}(\mathbf{x}_i), \mathbf{C}\mathbf{y}_i \rangle_{\mathcal{H}^k}\right) \tag{17}$$

and

$$H_{\boldsymbol{\gamma},\mathbf{C}} = \sum_{i=1}^{n} \phi_{\boldsymbol{\gamma}}^{\top}(\mathbf{x}_i) \sum_{v=1}^{k} \beta_{iv} \mathbf{C}\mathbf{e}_v. \tag{18}$$

we are going to prove that

$$\mathbb{E}_{\beta}\left[(G_{\boldsymbol{\gamma}_1,\mathbf{C}_1} - G_{\boldsymbol{\gamma}_2,\mathbf{C}_2})^2\right] \leq \mathbb{E}_{\beta}\left[(H_{\boldsymbol{\gamma}_1,\mathbf{C}_1} - H_{\boldsymbol{\gamma}_2,\mathbf{C}_2})^2\right]. \tag{19}$$

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

$$\begin{aligned}
&\left[\left(1 - \max_{\mathbf{y}\in\Theta} \langle \phi_{\boldsymbol{\gamma}_1}(\mathbf{x}), \mathbf{C}_1\mathbf{y} \rangle_{\mathcal{H}^k}\right) - \left(1 - \max_{\mathbf{y}\in\Theta} \langle \phi_{\boldsymbol{\gamma}_2}(\mathbf{x}), \mathbf{C}_2\mathbf{y} \rangle_{\mathcal{H}^k}\right)\right]^2 \\
&= \left(\max_{\mathbf{y}\in\Theta} \langle \phi_{\boldsymbol{\gamma}_1}(\mathbf{x}), \mathbf{C}_1\mathbf{y} \rangle_{\mathcal{H}^k} - \max_{\mathbf{y}\in\Theta} \langle \phi_{\boldsymbol{\gamma}_2}(\mathbf{x}), \mathbf{C}_2\mathbf{y} \rangle_{\mathcal{H}^k}\right)^2 \\
&\leq \left(\max_{\mathbf{y}\in\Theta} \left(\phi_{\boldsymbol{\gamma}_1}^{\top}(\mathbf{x})\mathbf{C}_1\mathbf{y} - \phi_{\boldsymbol{\gamma}_2}^{\top}(\mathbf{x})\mathbf{C}_2\mathbf{y}\right)\right)^2 \\
&= \left(\max_{\mathbf{y}\in\Theta} \left(\phi_{\boldsymbol{\gamma}_1}^{\top}(\mathbf{x})\mathbf{C}_1 - \phi_{\boldsymbol{\gamma}_2}^{\top}(\mathbf{x})\mathbf{C}_2\right)\mathbf{y}\right)^2 \\
&= \max_{\mathbf{y}\in\Theta} \left(\sum_{v=1}^{k} y_v \left(\phi_{\boldsymbol{\gamma}_1}^{\top}(\mathbf{x})\mathbf{C}_1 - \phi_{\boldsymbol{\gamma}_2}^{\top}(\mathbf{x})\mathbf{C}_2\right)\mathbf{e}_v\right)^2 \\
&\leq \sum\nolimits_{v=1}^{k} \left(\left(\phi_{\boldsymbol{\gamma}_1}^{\top}(\mathbf{x})\mathbf{C}_1 - \phi_{\boldsymbol{\gamma}_2}^{\top}(\mathbf{x})\mathbf{C}_2\right)\mathbf{e}_v\right)^2,
\end{aligned} \tag{20}$$

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

Thus, we have

$$\begin{aligned}
&\mathbb{E}_{\beta}\left[(G_{\boldsymbol{\gamma}_1,\mathbf{C}_1} - G_{\boldsymbol{\gamma}_2,\mathbf{C}_2})^2\right] \\
=&\mathbb{E}_{\beta}\left[\left(\sum\nolimits_{i=1}^{n} \beta_i \left[\left(1 - \max\nolimits_{\mathbf{y}_i\in\Theta} \langle \phi_{\boldsymbol{\gamma}_1}(\mathbf{x}_i), \mathbf{C}_1\mathbf{y}_i \rangle_{\mathcal{H}^k}\right)\right.\right.\right. \\
&\qquad \left.\left.\left. - \left(1 - \max\nolimits_{\mathbf{y}_i\in\Theta} \langle \phi_{\boldsymbol{\gamma}_2}(\mathbf{x}_i), \mathbf{C}_2\mathbf{y}_i \rangle_{\mathcal{H}^k}\right)\right]\right)^2\right] \\
=&\sum_{i=1}^{n} \left(\max_{\mathbf{y}_i\in\Theta} \langle \phi_{\boldsymbol{\gamma}_1}(\mathbf{x}_i), \mathbf{C}_1\mathbf{y}_i \rangle_{\mathcal{H}^k} - \max_{\mathbf{y}_i\in\Theta} \langle \phi_{\boldsymbol{\gamma}_2}(\mathbf{x}_i), \mathbf{C}_2\mathbf{y}_i \rangle_{\mathcal{H}^k}\right)^2 \\
\leq&\sum_{i=1}^{n}\sum_{v=1}^{k} \left(\left(\phi_{\boldsymbol{\gamma}_1}^{\top}(\mathbf{x}_i)\mathbf{C}_1 - \phi_{\boldsymbol{\gamma}_2}^{\top}(\mathbf{x}_i)\mathbf{C}_2\right)\mathbf{e}_v\right)^2 \\
=&\mathbb{E}_{\beta}\left[(H_{\boldsymbol{\gamma}_1,\mathbf{C}_1} - H_{\boldsymbol{\gamma}_2,\mathbf{C}_2})^2\right].
\end{aligned} \tag{21}$$

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

$$\begin{aligned}
\mathbb{E}\left[\sup\nolimits_{f\in\mathcal{F}} H_f\right] &= \mathbb{E}_{\beta}\left[\sup_{\mathbf{C},\boldsymbol{\gamma}} \sum_{i=1}^{n}\sum_{v=1}^{k} \beta_{iv}\phi_{\boldsymbol{\gamma}}^{\top}(\mathbf{x}_i)\mathbf{C}\mathbf{e}_v\right] \\
&\leq \mathbb{E}_{\beta}\left[b\sum_{v=1}^{k}\left|\sum\nolimits_{i=1}^{n}\beta_{iv}\right|\right] \\
&\leq bk\sqrt{n}.
\end{aligned} \tag{22}$$

Combining Lemmas **??** and **??**, Eqs. (**??**) (**??**), and (**??**), we have

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

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

$$\mathbb{E}\left[f(\mathbf{x})\right] \leq \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) + \frac{\sqrt{\pi/2}\,bk}{\sqrt{n}} + (1+b)\sqrt{\frac{\log 1/\delta}{2n}}. \tag{23}$$

This completes the proof.