

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

Xinwang Liu, Li Liu, Jian Xiong, En Zhu, Junwei Han, Meng Wang, Dinggang Shen, and Wen Gao

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

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

where  $\phi_{\hat{\gamma}}(\mathbf{x}) = [\hat{\gamma}_1 \phi_1^\top(\mathbf{x}), \dots, \hat{\gamma}_m \phi_m^\top(\mathbf{x})]^\top$  is the learned feature map associated with the kernel function  $K_{\hat{\gamma}}(\cdot, \cdot)$  and  $\mathbf{e}_1, \dots, \mathbf{e}_k$  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:

$$\mathcal{F} = \left\{ f : \mathbf{x} \mapsto 1 - \max_{\mathbf{y} \in \Theta} \langle \phi_{\gamma}(\mathbf{x}), \mathbf{C}\mathbf{y} \rangle_{\mathcal{H}^k} \mid \gamma^\top \mathbf{1}_m = 1, \right. \\ \left. \gamma_p \geq 0, \mathbf{C} \in \mathcal{H}^k, |K_p(\mathbf{x}, \tilde{\mathbf{x}})| \leq b, \forall p, \forall \mathbf{x} \in \mathcal{X} \right\}, \quad (2)$$

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

- X. Liu and E. Zhu are with College of Computer, National University of Defense Technology, Changsha, 410073, China. E-mail: {xinwangliu, enzhu}@nudt.edu.cn.
- L. Liu is with College of System Engineering, National University of Defense Technology, Changsha, China, and also with the Center for Machine Vision and Signal Analysis, University of Oulu, 90014 Oulu, Finland (E-mail: li.liu@oulu.fi).
- J. Xiong is with School of Business Administration, Southwestern University of Finance and Economics, Chengdu, Sichuan, 611130, China (E-mail: xiongjian2017@swufe.edu.cn).
- J. Han is with School of Automation, Northwestern Polytechnical University, Xian, China (E-mail: junwei.han2010@gmail.com).
- M. Wang is with School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China (E-mail: eric.mengwang@gmail.com).
- D. Shen is with Department of Radiology and BRIC, University of North Carolina at Chapel Hill, North Carolina 27599, USA, and also with Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea (E-mail: dgshen@med.unc.edu).
- W. Gao is with School of Electronics Engineering and Computer Science, Peking University, Beijing, China, 100871 (E-mail: wgao@pku.edu.cn).

Manuscript received November 18, 2020.

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

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

## 3 PROOF OF THEOREM 1

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

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

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

Let

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

Our proof idea is to upper bound

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

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

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

$$\mathcal{F} = \left\{ f : \mathbf{x} \mapsto 1 - \max_{\mathbf{y} \in \Theta} \langle \phi_{\gamma}(\mathbf{x}), \mathbf{C}\mathbf{y} \rangle_{\mathcal{H}^k} \mid \gamma^\top \mathbf{1}_m = 1, \gamma_p \geq 0, \right. \\ \left. \mathbf{C} \in \mathcal{H}^k, |K_p(\mathbf{x}, \tilde{\mathbf{x}})| \leq b, \forall p, \forall \mathbf{x} \in \mathcal{X} \right\}, \quad (7)$$

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

Then, Eq. (6) becomes

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

It is obvious that

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

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

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

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 [1].

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

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

where

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

and  $\sigma_1, \dots, \sigma_n$  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 bound  $\mathfrak{R}_n(\mathcal{F})$ . Similar to the proof method in [2], we upper bound it by introducing Gaussian complexities:

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

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

The following two lemmas [2] 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  $f_1, f_2 \in \mathcal{F}$ ,

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

Then,

$$\mathbb{E}[\sup_{f \in \mathcal{F}} G_f] \leq \mathbb{E}[\sup_{f \in \mathcal{F}} H_f]. \tag{16}$$

In our case, let

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

and

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

we are going to prove that

$$\mathbb{E}_{\beta} [(G_{\gamma_1, \mathbf{C}_1} - G_{\gamma_2, \mathbf{C}_2})^2] \leq \mathbb{E}_{\beta} [(H_{\gamma_1, \mathbf{C}_1} - H_{\gamma_2, \mathbf{C}_2})^2]. \tag{19}$$

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

$$\begin{aligned}
 &\left[ \left( 1 - \max_{\mathbf{y} \in \Theta} \langle \phi_{\gamma_1}(\mathbf{x}), \mathbf{C}_1 \mathbf{y} \rangle_{\mathcal{H}^k} \right) - \left( 1 - \max_{\mathbf{y} \in \Theta} \langle \phi_{\gamma_2}(\mathbf{x}), \mathbf{C}_2 \mathbf{y} \rangle_{\mathcal{H}^k} \right) \right]^2 \\
 &= \left( \max_{\mathbf{y} \in \Theta} \langle \phi_{\gamma_1}(\mathbf{x}), \mathbf{C}_1 \mathbf{y} \rangle_{\mathcal{H}^k} - \max_{\mathbf{y} \in \Theta} \langle \phi_{\gamma_2}(\mathbf{x}), \mathbf{C}_2 \mathbf{y} \rangle_{\mathcal{H}^k} \right)^2 \\
 &\leq \left( \max_{\mathbf{y} \in \Theta} \left( \phi_{\gamma_1}^{\top}(\mathbf{x}) \mathbf{C}_1 \mathbf{y} - \phi_{\gamma_2}^{\top}(\mathbf{x}) \mathbf{C}_2 \mathbf{y} \right) \right)^2 \\
 &= \left( \max_{\mathbf{y} \in \Theta} \left( \phi_{\gamma_1}^{\top}(\mathbf{x}) \mathbf{C}_1 - \phi_{\gamma_2}^{\top}(\mathbf{x}) \mathbf{C}_2 \right) \mathbf{y} \right)^2 \\
 &= \max_{\mathbf{y} \in \Theta} \left( \sum_{v=1}^k y_v \left( \phi_{\gamma_1}^{\top}(\mathbf{x}) \mathbf{C}_1 - \phi_{\gamma_2}^{\top}(\mathbf{x}) \mathbf{C}_2 \right) \mathbf{e}_v \right)^2 \\
 &\leq \sum_{v=1}^k \left( \left( \phi_{\gamma_1}^{\top}(\mathbf{x}) \mathbf{C}_1 - \phi_{\gamma_2}^{\top}(\mathbf{x}) \mathbf{C}_2 \right) \mathbf{e}_v \right)^2,
 \end{aligned} \tag{20}$$

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

Thus, we have

$$\begin{aligned}
 &\mathbb{E}_{\beta} \left[ (G_{\gamma_1, \mathbf{C}_1} - G_{\gamma_2, \mathbf{C}_2})^2 \right] \\
 &= \mathbb{E}_{\beta} \left[ \left( \sum_{i=1}^n \beta_i \left[ \left( 1 - \max_{\mathbf{y}_i \in \Theta} \langle \phi_{\gamma_1}(\mathbf{x}_i), \mathbf{C}_1 \mathbf{y}_i \rangle_{\mathcal{H}^k} \right) - \left( 1 - \max_{\mathbf{y}_i \in \Theta} \langle \phi_{\gamma_2}(\mathbf{x}_i), \mathbf{C}_2 \mathbf{y}_i \rangle_{\mathcal{H}^k} \right) \right] \right)^2 \right] \\
 &= \sum_{i=1}^n \left( \max_{\mathbf{y}_i \in \Theta} \langle \phi_{\gamma_1}(\mathbf{x}_i), \mathbf{C}_1 \mathbf{y}_i \rangle_{\mathcal{H}^k} - \max_{\mathbf{y}_i \in \Theta} \langle \phi_{\gamma_2}(\mathbf{x}_i), \mathbf{C}_2 \mathbf{y}_i \rangle_{\mathcal{H}^k} \right)^2 \\
 &\leq \sum_{i=1}^n \sum_{v=1}^k \left( \left( \phi_{\gamma_1}^{\top}(\mathbf{x}_i) \mathbf{C}_1 - \phi_{\gamma_2}^{\top}(\mathbf{x}_i) \mathbf{C}_2 \right) \mathbf{e}_v \right)^2 \\
 &= \mathbb{E}_{\beta} [(H_{\gamma_1, \mathbf{C}_1} - H_{\gamma_2, \mathbf{C}_2})^2].
 \end{aligned} \tag{21}$$

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

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

Combining Lemmas 1 and 2, Eqs. (13) (21), and (22), we have

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

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

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

This completes the proof.

## REFERENCES

- [1] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 463–482, 2002.
- [2] A. Maurer and M. Pontil, " $k$ -dimensional coding schemes in Hilbert spaces," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5839–5846, 2010.