Research Article

Generalization Bounds for Coregularized Multiple Kernel Learning

Xinxing Wu and Guosheng Hu

Department of Communication and Information Engineering, Shanghai Technical Institute of Electronics & Information, Shanghai 201411, China

Correspondence should be addressed to Xinxing Wu; xinxingwu@yeah.net

Received 10 February 2018; Revised 26 August 2018; Accepted 13 September 2018; Published 1 November 2018

Academic Editor: Amparo Alonso-Betanzos

Copyright © 2018 Xinxing Wu and Guosheng Hu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Multiple kernel learning (MKL) as an approach to automated kernel selection plays an important role in machine learning. Some learning theories have been built to analyze the generalization of multiple kernel learning. However, less work has been studied on multiple kernel learning in the framework of semisupervised learning. In this paper, we analyze the generalization of multiple kernel learning in the framework of semisupervised multiview learning. We apply Rademacher chaos complexity to control the performance of the candidate class of coregularized multiple kernels and obtain the generalization error bound of coregularized multiple kernel learning. Furthermore, we show that the existing results about multiple kernel learning and coregularized kernel learning can be regarded as the special cases of our main results in this paper.

1. Introduction

Kernel-based learning is related to achieve nonlinear machine learning tasks from linear ones. In the real applications, selecting a good or suitable kernel for the kernel-based learning is an important and difficult task. To this end, an approach named multiple kernel learning has been developed, and it allows to automatically choose the best kernel from a predefined kernel class. The earliest work of multiple kernel learning can be traced back to the research in [1], where the authors proposed to automatically pick up a linear combination of candidate kernels for the support vector machines based on a semidefinite programming approach. Theoretical generalization analysis of multiple kernel learning has been widely studied by many researchers [1–7]. In particular, Ying and Campbell in [2] proposed a novel generalization bound (Rademacher chaos complexity) for the study of multiple kernel learning. However, the discussions in [2] were for single view and supervised learning. In this paper, we will employ Rademacher chaos complexity proposed in [2] to study the generalization error of coregularized multiple kernel learning in the semisupervised multiview learning framework.

Semisupervised multiview learning as an area of machine learning is trained with both labeled samples and unlabeled samples, and the unlabeled samples are helpful to reduce the amount of the labeled samples. Semisupervised multiview learning supposes that the train samples can be represented by multiple views. The coregularized least squares algorithm—a semisupervised version of regularized least squares with two views—is a typical multiview learning model that uses the unlabeled samples to estimate the view incompatibility of models [8, 9]. Rosenberg in [10] extended the coregularized least squares algorithm to the case of kernel cotraining. And Brefeld et al. and Rosenberg in [11, 12, 13] discussed the generalization bound of kernel-based learning with multiple (or two) views in the semisupervised learning framework. However, the discussions in [11, 12, 13] supposed that the kernel used to construct the reproducing kernel Hilbert space is predefined. Therefore, their results cannot be used to the analysis of multiple kernel learning.

In this paper, we will discuss the generalization error of coregularized multiple kernel learning in the semisupervised multiview learning framework. And we show that the results in [2] and [11] can be regarded as the special cases of our main

results. The rest of the paper is organized as follows. In Section 2, we introduce some basic notations and definitions for later discussion. In Section 3, we discuss the related research and put forward the question that will be studied in this paper. In Section 4, we present our main results. In Section 5, we give the main proofs for our main results proposed in Section 4. In Section 6, we give a comparative discussion of our results to the existing work and show that the results about multiple kernel learning in [2] and coregularized kernel learning in [11] can be regarded as the special cases of our main results. The last Section 7 concludes this paper.

2. Notations and Definitions

In this section, we introduce notations and definitions for later discussions:

- (i) Let $\mathbb{N}$ be the set of natural numbers and $\mathbb{R}$ be the set of real numbers. Let $\mathbb{N}_n = \{1, 2, \dots, n\}$, $n \in \mathbb{N}$.
- (ii) Let $(\Omega, \mathcal{A}, P)$ be a probability space; that is, Ω alone is called the sample space, $\mathcal{A}$ is a σ -algebra on Ω , and P is a probability measure on $(\Omega, \mathcal{A})$. And Ω has the structure $\Omega = \mathcal{X} \times \mathcal{Y} (\subset \mathbb{R})$, where $\mathcal{X}$ and $\mathcal{Y}$ are the input space and output space, respectively. Denote $P_{\mathcal{X}}$ as the marginal distribution on $\mathcal{X}$. For ignoring the discussion of measure theory, we simply denote $(\Omega, \mathcal{A}, P)$ as (Ω, P) .
- (iii) Let $\mathcal{F}$ be the set of all measurable functions $f : \mathcal{X} \rightarrow \mathcal{Y}$. Assume that $\mathcal{H}$ is a subset of $\mathcal{F}$. That is, $\mathcal{H} \subset \mathcal{F}$, the set $\mathcal{H}$ is called the hypothesis class.
- (iv) Let $S = \{z_i = (x_i, y_i), i \in \mathbb{N}_n\}$ be a finite set of the labeled training samples, and assume these samples are independent and identically distributed (i.i.d.) according to P . Denote the bold letter as a vector; for example, $\mathbf{z}$ presents a vector $(z_1, z_2, \dots, z_n)$.
- (v) For the sign $|\cdot|$, if D is a set, we use $|D|$ to represent the number of elements of a set and if D is a function, we use $|D|$ to represent the absolute value of the function D .
- (vi) If A is a matrix, we use A^T to represent the transpose of the matrix A .
- (vii) Let L be the loss function, $L : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, +\infty)$, and the loss of f on a sample point $z = (x, y)$ is defined by $L(f, z)$ or $L(f(x), y)$.

In learning theory, one of the purposes is to pick up a function f in hypothesis space $\mathcal{H}$ that minimizes the following generalization error:

$$E_P[L(f(X), Y)] \triangleq \int_{\Omega} L(f, z) dP. \quad (1)$$

Generally speaking, the distribution P in the above Equation (1) is unknown. Rather than minimizing $E_P[L(f(X), Y)]$, we usually minimize the empirical or training error below:

$$E_n(f, S) \triangleq \frac{1}{m} \sum_{i \in \mathbb{N}_m} L(f, z_i), \quad (2)$$

where the sign S represents the finite labeled samples and $z_i = (x_i, y_i) \in S, i \in \mathbb{N}_m$.

In this paper, the main quantity we are interested in is the following uniform estimation of the difference between the generalization error and empirical error:

$$\sup_f \{E_P[L(f(X), Y)] - E_n(f, S)\}. \quad (3)$$

For the discussion in the later sections, we introduce the following four definitions and one lemma (Definition 5 is proposed by us).

Definition 1 (Empirical Rademacher Complexity) [2]. Let $\mathcal{H}$ be a class of functions $f : \Omega \rightarrow \mathbb{R}$. The samples $x_i, i \in \mathbb{N}_n$, are independently drawn from the probability space (Ω, P) . The empirical Rademacher complexity can be defined as

$$\mathcal{R}_n(\mathcal{H}) \triangleq E_{\sigma} \left[\sup_{f \in \mathcal{H}} \frac{1}{\sqrt{n}} \sum_{i \in \mathbb{N}_n} \sigma_i f(x_i) \right], \quad (4)$$

where the random variables $\sigma_i, i \in \mathbb{N}_n$, are Rademacher variables, and σ presents a vector $(\sigma_1, \sigma_2, \dots, \sigma_n)$.

Definition 2 (Empirical Rademacher Chaos Complexity) [2]. Let $\mathcal{H}$ be a class of functions $f : \Omega \times \Omega \rightarrow \mathbb{R}$. The samples $x_i, i \in \mathbb{N}_n$, are independently drawn from the probability space (Ω, P) . The empirical Rademacher chaos complexity can be defined as

$$\mathcal{U}_n(\mathcal{H}) \triangleq E_{\sigma} \left[\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i, j \in \mathbb{N}_n, i < j} \sigma_i \sigma_j f(x_i, x_j) \right], \quad (5)$$

where the random variables $\sigma_i, i \in \mathbb{N}_n$, are Rademacher variables, and σ presents a vector $(\sigma_1, \sigma_2, \dots, \sigma_n)$.

Definition 3 (Reproducing Kernel Hilbert Space, RKHS) [14]. The function

$$K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}, \quad (6)$$

is a reproducing kernel of the Hilbert space $\mathcal{H}_K$ if and only if.

- (i) For any $x \in \mathcal{X}, K(\cdot, x) \in \mathcal{H}_K$
- (ii) For any $f \in \mathcal{H}_K$, for any $x \in \mathcal{X}$, $\langle f, K(x) \rangle_{\mathcal{H}_K} = f(x)$

A Hilbert space of functions which possesses a reproducing kernel is called a reproducing kernel Hilbert space.

Remark 1. The second condition in the above Definition 3 is called “the reproducing property”: the value of the function f at the point x is reproduced by the inner product of f with $K(x)$. From the above two conditions, for any, $(x, y) \in \mathcal{X} \times \mathcal{X}$, it is clear that

$$K(x, y) = \langle K(x, \cdot), K(y, \cdot) \rangle_{\mathcal{H}_K}. \quad (7)$$

In real applications, the solution to many reproducing kernel Hilbert space optimization questions is contained in a special subspace of the reproducing kernel Hilbert space,

often known as the “span of the data”. The span of the data for an reproducing kernel Hilbert space $\mathcal{H}_K$ is the linear subspace (Appendix A.3 on page 75 of [13]):

$$\text{span}\{K(x_i, \cdot) : x_i \in \mathcal{X}, i \in \mathbb{N}_m\}. \quad (8)$$

For simplicity, we denote the above linear subspace by $\mathcal{L}_m^K$.

Let $\mathcal{K} \subset \{K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}\}$ be a class of kernels. In this paper, we assume that $\kappa \triangleq \sup_{K \in \mathcal{K}, x \in \mathcal{X}} \sqrt{K(x, x)}$ is finite.

Definition 4 (Subnormalized Functional with Degenerate Dimension n). If a loss function $\ell : (\prod_{i \in \mathbb{N}_n} \mathcal{H}_{K_i} \times \mathcal{Y} \rightarrow [0, +\infty)$ satisfies

$$\ell\left(\underbrace{0, \dots, 0}_n, y\right) \leq 1, \text{ for any } y \in \mathcal{Y}, \quad (9)$$

then we call ℓ a subnormalized functional with degenerate dimension n on $(\prod_{i \in \mathbb{N}_n} \mathcal{H}_{K_i} \times \mathcal{Y})$.

Lemma 1. [13] Let $\mathcal{H}$ be a reproducing kernel Hilbert space with kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, and consider any point $x \in \mathcal{X}$. If $\mathcal{L}^K$ is a closed subspace containing $K(x)$, then the projection of f onto $\mathcal{L}^K$ has the same value at x as f does. That is,

$$f(x) = (\text{Proj}_{\mathcal{L}^K} f)(x). \quad (10)$$

For multiple kernel learning, the main task is to automatically choose a kernel K from a predefined class $\mathcal{K}$ of kernels, and find a function from the reproducing kernel Hilbert space $\mathcal{H}_K$ that is most suitable to model the given samples.

In this paper, our purpose is to minimize

$$\begin{aligned} Q(f^1, f^2) \triangleq & \frac{1}{m} \sum_{i \in \mathbb{N}_m} \ell(f^1(x_i), f^2(x_i), y_i) + \gamma_1 \|f^1\|_{\mathcal{H}_{K_1}}^2 \\ & + \gamma_2 \|f^2\|_{\mathcal{H}_{K_2}}^2 + \lambda \sum_{i=m+1}^{m+u} (f^1(x_i) - f^2(x_i))^2, \end{aligned} \quad (11)$$

over the class $(\cup_{K_1 \in \mathcal{K}} \mathcal{H}_{K_1}) \times (\cup_{K_2 \in \mathcal{K}'} \mathcal{H}_{K_2})$. Here, let $\mathcal{K}$ and $\mathcal{K}'$ be the classes of kernels. $\mathcal{H}_{K_1}$ and $\mathcal{H}_{K_2}$ denote the reproducing kernel Hilbert spaces. m represents the amount of the labeled points $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i \in \mathbb{N}_m$, and u represents the amount of the unlabeled points $x_{m+i} \in \mathcal{X}, i \in \mathbb{N}_u$. The signs γ_1, γ_2 , and λ mean the regularization parameters. The functions f^1 and f^2 , respectively, represent two viewers, and the function $\ell(f^1(\cdot), f^2(\cdot), \cdot)$ is the labeled loss function, which measures the performance of f^1 and f^2 on the labeled points $z_i, i \in \mathbb{N}_m$.

3. Related Work

In [11], Rosenberg and Bartlett used Rademacher complexity to bound the coregularized kernel class in the semisupervised two-view learning framework, and two viewers are two predefined reproducing kernel Hilbert

spaces ($\mathcal{H}_1$ and $\mathcal{H}_2$, respectively). Take labeled points $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i \in \mathbb{N}_m$, and unlabeled points $x_{m+i} \in \mathcal{X}, i \in \mathbb{N}_u$. The coregularized least squares algorithm discussed in [11] can be described as follows:

$$\begin{aligned} (f_z^1, f_z^2) \triangleq & \arg \min_{\substack{f^1 \in \mathcal{H}_1 \\ f^2 \in \mathcal{H}_2}} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \ell(f^1(x_i), f^2(x_i), y_i) + \gamma_1 \|f^1\|_{\mathcal{H}_1}^2 \right. \\ & \left. + \gamma_2 \|f^2\|_{\mathcal{H}_2}^2 + \lambda \sum_{i=m+1}^{m+u} (f^1(x_i) - f^2(x_i))^2 \right\}, \end{aligned} \quad (12)$$

where m represents the amount of the labeled points $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i \in \mathbb{N}_m$, and u represents the amount of the unlabeled points $x_{m+i} \in \mathcal{X}, i \in \mathbb{N}_u$. The signs γ_1, γ_2 , and λ mean the regularization parameters. The functions f^1 and f^2 , respectively, represent two viewers, and the function $\ell(f^1(\cdot), f^2(\cdot), \cdot)$ is the labeled loss function, which measures the performance of f^1 and f^2 on the labeled points $z_i, i \in \mathbb{N}_m$.

In [11], the final output is denoted as $(f_z^1 + f_z^2)/2$.

In [2], Ying and Campbell applied the Rademacher chaos complexity to study the generalization of multiple kernel learning in the supervised learning framework. The multiple kernel learning model they considered is as follows:

$$f_z \triangleq \arg \min_{K \in \mathcal{K}} \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \ell(f(x_i), y_i) + \gamma \|f\|_{\mathcal{H}_K}^2 \right\}, \quad (13)$$

where m represents the amount of the labeled points $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i \in \mathbb{N}_m$. The function $\ell(f(x_i), y_i)$ is the loss function. The sign λ means the regularization parameters. And $\mathcal{H}_K$ denotes the reproducing kernel Hilbert spaces. In Equation (13), $\mathcal{H}_K$ is not predefined and depends on the kernel choose from the class $\mathcal{K}$ of kernel.

In this paper, we are interested in the topic of coregularized multiple kernel learning; that is, the two reproducing kernel Hilbert spaces are not defined in advance. Our discussions are in the framework of semisupervised multiview learning. We give this learning question as the following two-layer minimization formation:

$$\begin{aligned} (f_z^1, f_z^2) \triangleq & \arg \min_{\substack{K_1 \in \mathcal{K} \\ K_2 \in \mathcal{K}'}} \min_{\substack{f^1 \in \mathcal{H}_{K_1} \\ f^2 \in \mathcal{H}_{K_2}}} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \ell(f^1(x_i), f^2(x_i), y_i) \right. \\ & + \gamma_1 \|f^1\|_{\mathcal{H}_{K_1}}^2 + \gamma_2 \|f^2\|_{\mathcal{H}_{K_2}}^2 \\ & \left. + \lambda \sum_{i=m+1}^{m+u} (f^1(x_i) - f^2(x_i))^2 \right\}, \end{aligned} \quad (14)$$

where let $\mathcal{K}$ and $\mathcal{K}'$ be the classes of kernels. $\mathcal{H}_{K_1}$ and $\mathcal{H}_{K_2}$ denote the reproducing kernel Hilbert spaces. m represents the amount of the labeled points $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i \in \mathbb{N}_m$, and u represents the amount of the unlabeled points

$x_{m+i} \in \mathcal{X}, i \in \mathbb{N}_u$. The signs γ_1, γ_2 , and λ mean the regularization parameters. The functions f^1 and f^2 , respectively, represent two viewers, and the function $\ell(f^1(\cdot), f^2(\cdot), \cdot)$ is the labeled loss function, which measures the performance of f^1 and f^2 on the labeled points $z_i, i \in \mathbb{N}_m$.

Remark 2. We will explain the following:

- (1) Equation (14) given in this paper is different from Equation (12):
  - (a) The solution from Equation (14) is to minimize on the class $(\cup_{K_1 \in \mathcal{K}} \mathcal{H}_{K_1}) \times (\cup_{K_2 \in \mathcal{K}'} \mathcal{H}_{K_2})$, while the solution from Equation (12) is to minimize on the class $\mathcal{H}_{K_1} \times \mathcal{H}_{K_2}$.
  - (b) The solution from Equation (14) is through two minimization steps: first, it finds the most suitable reproducing kernel for the given samples. Second, it obtains the best function/model from the found reproducing kernel Hilbert spaces in the first step. While in Equation (12), it only needs to get the best function/model from the reproducing kernel Hilbert spaces which are predefined.
- (2) Equation (14) given in this paper is different from Equation (13):
  - (a) The solution from Equation (14) is to minimize on the product space $(\cup_{K_1 \in \mathcal{K}} \mathcal{H}_{K_1}) \times (\cup_{K_2 \in \mathcal{K}'} \mathcal{H}_{K_2})$, while the solution from Equation (12) is to minimize on the space $(\cup_{K_1 \in \mathcal{K}} \mathcal{H}_{K_1})$.
  - (b) The minimization item in Equation (13) is much simpler because the analysis on Equation (13) is limited to supervised learning and single view and does not deal with unlabeled samples.

From the above discussion, we can see that the generalization analysis about Equation (14) will make more meaningful and bring greater challenge.

In the next section, we will present the main results of this paper.

4. Generalization Bounds

In this section, we assume that the loss function ℓ in Equation (11) is the subnormalized functional with degenerate dimension 2 on $\mathcal{H}_{K_1} \times \mathcal{H}_{K_2} \times \mathcal{Y}, K_1 \in \mathcal{K}, K_2 \in \mathcal{K}'$. In Equation (11), let $f^1 = 0$ and $\tilde{f}^1 = 0$, we have

$$\min_{\substack{K_1 \in \mathcal{K} \\ K_2 \in \mathcal{K}'}} \min_{\substack{f^1 \in \mathcal{H}_{K_1} \\ f^2 \in \mathcal{H}_{K_2}}} \mathcal{Q}(f^1, f^2) \leq \mathcal{Q}(0, 0) = \frac{1}{m} \sum_{i \in \mathbb{N}_m} \ell(0, 0, y_i) \leq 1, \quad (15)$$

Note that $\mathcal{Q}(f^1, f^2) \geq 0$, and then for any samples $(x_i, y_i), i \in \mathbb{N}_m$ and $x_{m+i}, i \in \mathbb{N}_u$, the class of candidate reproducing kernel Hilbert spaces is defined as follows:

$$(f_z^1, f_z^2) \in \mathcal{H}_{\mathcal{K}, \mathcal{K}'} \triangleq \left\{ (f^1, f^2) : \gamma_1 \|f^1\|_{\mathcal{H}_{K_1}}^2 + \gamma_2 \|f^2\|_{\mathcal{H}_{K_2}}^2 + \lambda \sum_{i=m+1}^{m+u} (f^1(x_i) - f^2(x_i))^2 \leq 1, K_1 \in \mathcal{K}, K_2 \in \mathcal{K}' \right\}. \quad (16)$$

That is, the solution (f_z^1, f_z^2) minimizing $\mathcal{Q}(f^1, f^2)$ belongs to the class $\mathcal{H}_{\mathcal{K}} \times \mathcal{H}_{\mathcal{K}'}$. For simplicity, we use $\mathcal{H}_{\mathcal{K}, \mathcal{K}'}$ to denote $\mathcal{H}_{\mathcal{K}} \times \mathcal{H}_{\mathcal{K}'}$ in the next sections.

As the assumption in [11], the final predictor for the coregularized multiple kernel learning is selected from the following class:

$$\hat{\mathcal{H}}_{\mathcal{K}, \mathcal{K}'} \triangleq \left\{ \frac{f+g}{2} : (f, g) \in \mathcal{H}_{\mathcal{K}, \mathcal{K}'} \right\}. \quad (17)$$

Remark 3. In Equations (16) and (17), we can see that $\mathcal{H}_{\mathcal{K}, \mathcal{K}'}$ and $\hat{\mathcal{H}}_{\mathcal{K}, \mathcal{K}'}$ mainly depend on the prescribed set of candidate kernels and the unlabeled data.

For any $f \in \mathcal{H}_{\mathcal{K}, \mathcal{K}'}$, we define the expected loss as Equation (2), and we can use the loss function L to compute the labeled empirical loss in Equation (11). For the given samples $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i \in \mathbb{N}_m$, the loss can be also computed as Equation (3).

If L satisfies the Lipschitz continuous condition on $\hat{\mathcal{H}}_{\mathcal{K}, \mathcal{K}'}$, we introduce the constant defined by

$$L_{\text{sup}} = \sup \left\{ |L(f(\cdot), \cdot)| : \text{for any } |f| \leq \frac{\kappa + \kappa'}{\sqrt{\min\{\gamma_1, \gamma_2\}}} \right\}, \quad (18)$$

and the local Lipschitz constant denoted as

$$L_{\text{sup}}^{\text{loc}} = \sup \left\{ \frac{|L(f(\cdot), \cdot)| - L(f'(\cdot), \cdot)|}{|f - f'|} : \text{for any } |f|, |f'| \leq \frac{\kappa + \kappa'}{\sqrt{\min\{\gamma_1, \gamma_2\}}} \right\}. \quad (19)$$

In the end of this section, we give the main theorems in this paper.

Theorem 1. *Let the function L be a subnormalized functional with degenerate dimension 1 on $\mathcal{H}_{\mathcal{K}, \mathcal{K}'} \times \mathcal{Y}$ and satisfy the Lipschitz continuous condition on $\mathcal{H}_{\mathcal{K}, \mathcal{K}'}$. Let the labeled samples $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i \in \mathbb{N}_m$, be independent random variables drawn from the probability space $(\mathcal{X} \times \mathcal{Y}, P)$, and the unlabeled samples $x_{m+i} \in \mathcal{X}, i \in \mathbb{N}_u$, be independent random variables drawn from the probability space $(\mathcal{X}, P_{\mathcal{X}})$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, for any $f \in \hat{\mathcal{H}}_{\mathcal{K}, \mathcal{K}'}$, the following inequality holds*

$$E_p[L(f(X), Y)] - E_n(f, S) \leq \frac{L_{\sup}^{\text{loc}}}{\sqrt{m}} \left\{ \frac{2}{\gamma_1} \mathcal{U}_m(\mathcal{K}) + \frac{2}{\gamma_2} \mathcal{U}_m(\mathcal{K}') + \frac{\kappa^2}{\gamma_1} + \frac{\kappa'^2}{\gamma_2} + -\frac{1}{\sqrt{m}} \right. \\ \left. \cdot E_{\sigma} \inf_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \sigma^T \cdot \left(\left(\frac{1}{\gamma_1} K_{MU}^1 - \frac{1}{\gamma_2} K_{MU}^2 \right) \cdot O \cdot \Delta_{UU}(\lambda, K_1, K_2) \cdot O^T \right)^2 \cdot \sigma \right\}^{1/2} + 3L_{\sup} \cdot \sqrt{\frac{\ln(2/\delta)}{2m}}, \quad (20)$$

where σ presents a vector $(\sigma_1, \sigma_2, \dots, \sigma_n)$ and $\sigma_i, i \in \mathbb{N}_m$, are Rademacher variables, $(1/\gamma_1) \cdot K_{UU}^1 + (1/\gamma_2) \cdot K_{UU}^2 = O$.

$D_{UU}(K_1, K_2) \cdot O^T$ with diagonal elements $D_{ii}(K_1, K_2) \geq 0, i \in \mathbb{N}_u$, and orthogonal matrix O , and

$$D_{UU}(K_1, K_2) = \begin{pmatrix} d_{11}(K_1, K_2) & & & \\ & d_{22}(K_1, K_2) & & \\ & & \ddots & \\ & & & d_{uu}(K_1, K_2) \end{pmatrix}, \quad (21)$$

$$\Delta_{UU}(\lambda, K_1, K_2) = \begin{pmatrix} \sqrt{\frac{1}{(\lambda^{-1} + d_{11}(K_1, K_2))}} & & & \\ & \sqrt{\frac{1}{(\lambda^{-1} + d_{22}(K_1, K_2))}} & & \\ & & \ddots & \\ & & & \sqrt{\frac{1}{(\lambda^{-1} + d_{uu}(K_1, K_2))}} \end{pmatrix}. \quad (22)$$

Corollary 1. Under the assumption of Theorem 1, we have the following inequality:

$$E_p[L(f(X), Y)] - \frac{1}{m} \sum_{i \in \mathbb{N}_m} L(f(x_i), y_i) \leq \frac{L_{\sup}^{\text{loc}}}{\sqrt{m}} \left\{ \frac{2}{\gamma_1} \mathcal{U}_m(\mathcal{K}) + \frac{2}{\gamma_2} \mathcal{U}_m(\mathcal{K}') + \frac{\kappa^2}{\gamma_1} + \frac{\kappa'^2}{\gamma_2} + -\frac{1}{\sqrt{m}} \cdot E_{\sigma} \inf_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \sigma^T \right. \\ \left. \cdot \left(\left(\frac{1}{\gamma_1} K_{MU}^1 - \frac{1}{\gamma_2} K_{MU}^2 \right) \cdot O \cdot \Delta_{UU}(\lambda, \kappa, \kappa') \cdot O^T \right)^2 \cdot \sigma \right\}^{1/2} + 3L_{\sup} \cdot \sqrt{\frac{\ln(2/\delta)}{2m}}. \quad (23)$$

Corollary 2. Under the assumption of Theorem 1 and assume $\mathcal{K} = \mathcal{K}'$ and $\gamma_1 = \gamma_2$, the following inequality holds

$$E_p[L(f(X), Y)] - \frac{1}{m} \sum_{i \in \mathbb{N}_m} L(f(x_i), y_i) \\ \leq \frac{L_{\sup}^{\text{loc}}}{\sqrt{m}} \sqrt{\frac{4}{\gamma_1} \mathcal{U}_m(\mathcal{K}) + 2 \frac{\kappa^2}{\gamma_1}} + 3L_{\sup} \cdot \sqrt{\frac{\ln(2/\delta)}{2m}}. \quad (24)$$

and Corollary 1 are the extensions of the results in [2] and [11], respectively.

5. Proofs

In this section, we will prove Theorem 1 and Corollaries 1 and 2 in Section 4. As the preparation for the next proof, we give two following lemmas (Lemmas 2 and 3) in advance.

Remark 4. We will proof Theorem 1 and Corollaries 1 and 2 in Section 5. And in Section 6, we will reveal that Theorem 1

Lemma 2. Let the function L be a subnormalized functional with degenerate dimension 1 on $\tilde{\mathcal{H}}_{\mathcal{K}, \mathcal{K}'} \times \mathcal{Y}$, and satisfy the

Lipschitz continuous condition on $\widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}$. Let the labeled samples $z_i, i \in \mathbb{N}_m$, be independently drawn from the $(\mathcal{X} \times \mathcal{Y}, P)$ and the unlabeled samples $x_{m+i} \in \mathcal{X}, i \in \mathbb{N}_u$, be independently from the probability space $(\mathcal{X}, P_{\mathcal{X}})$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have the following inequality:

$$\begin{aligned} & \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}} \{E_P[L(f(X), Y)] - E_n(f, S)\} \\ & \leq 2E_{\sigma} \left\{ \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot L(f(x_i), y_i) \right\} \right\} \\ & + 3L_{\text{sup}} \cdot \sqrt{\frac{\ln(2/\delta)}{2m}}, \end{aligned} \quad (25)$$

where σ presents a vector $(\sigma_1, \sigma_2, \dots, \sigma_n)$ and $\sigma_i, i \in \mathbb{N}_m$, are Rademacher variables.

Proof. For simplicity, let

$$\Delta(\mathbb{N}_m) \triangleq \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}} \left\{ E_P \left[L(f(X), Y) - \frac{1}{m} \sum_{i \in \mathbb{N}_m} L(f(x_i), y_i) \right] \right\}. \quad (26)$$

Replacing the i th sample (x_i, y_i) in the labeled samples with (x'_i, y'_i) , we have

$$\begin{aligned} & \Delta(\mathbb{N}_m) - \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}} \left\{ E_P[L(f(X), Y)] \right. \\ & \left. - \frac{1}{m} \left(\sum_{i \in \mathbb{N}_m/j} L(f(x_i), y_i) + L(f(x'_j), y'_j) \right) \right\} \\ & = \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}} \left[\frac{1}{m} (L(f(x_j), y_j) - L(f(x'_j), y'_j)) \right]. \end{aligned} \quad (27)$$

By McDiarmid's inequality (Theorem D.3 on page 372 of [15]), with probability at least $1 - \delta/2$, we have

$$\begin{aligned} & \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}} \left\{ E_P[L(f(X), Y)] - \frac{1}{m} \sum_{i \in \mathbb{N}_m} L(f(x_i), y_i) \right\} \\ & \leq E_P \left\{ \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}} \left\{ E_P[L(f(X), Y)] - \frac{1}{m} \sum_{i \in \mathbb{N}_m} L(f(x_i), y_i) \right\} \right\} \\ & + L_{\text{sup}} \cdot \sqrt{\frac{\ln(2/\delta)}{2m}}. \end{aligned} \quad (28)$$

By the symmetrization argument (page 36 of [15]), we bound the expectation of the first term on the right-hand side of the above inequality (28) as follows:

$$\begin{aligned} & E_P \left\{ \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}} \left\{ E_P[L(f(X), Y)] - \frac{1}{m} \sum_{i \in \mathbb{N}_m} L(f(x_i), y_i) \right\} \right\} \\ & \leq 2E_P E_{\sigma} \left\{ \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot L(f(x_i), y_i) \right\} \right\}. \end{aligned} \quad (29)$$

Again, applying McDiarmid's inequality to the right-hand side of the above inequality (29), with probability at least $1 - \delta/2$, we have

$$\begin{aligned} & E_P E_{\sigma} \left\{ \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot L(f(x_i), y_i) \right\} \right\} \\ & \leq E_{\sigma} \left\{ \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot L(f(x_i), y_i) \right\} \right\} \\ & + L_{\text{sup}} \cdot \sqrt{\frac{\ln(2/\delta)}{2m}}. \end{aligned} \quad (30)$$

Combining inequalities (28), (29), and (30) yield with probability at least $1 - \delta$:

$$\begin{aligned} & \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}} \left\{ E_P[L(f(X), Y)] - \frac{1}{m} \sum_{i \in \mathbb{N}_m} L(f(x_i), y_i) \right\} \\ & \leq 2E_{\sigma} \left\{ \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot L(f(x_i), y_i) \right\} \right\} \\ & + 3L_{\text{sup}} \cdot \sqrt{\frac{\ln(2/\delta)}{2m}}. \end{aligned} \quad (31)$$

□

Lemma 3. Under the assumption of Lemma 2, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have the following inequality:

$$\begin{aligned} & E_{\sigma} \left\{ \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot L(f(x_i), y_i) \right\} \right\} \\ & \leq L_{\text{sup}}^{\text{loc}} \cdot E_{\sigma} \left\{ \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{X}, \mathcal{X}'}} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot f(x_i) \right\} \right\}, \end{aligned} \quad (32)$$

where, σ presents a vector $(\sigma_1, \sigma_2, \dots, \sigma_n)$ and $\sigma_i, i \in \mathbb{N}_m$, are Rademacher variables.

Proof. As defined in Equation (19), $L_{\text{sup}}^{\text{loc}}$ is the local Lipschitz constant of L . And by the contraction property of Rademacher complexity (Lemma 26.9 on page 331 of [16] and Theorem 7 of [17]), the result is as follows. □

Lemma 4. If Equation (11) has a solution, then, for a fixed $K^1 \in \mathcal{K}$ and a fixed $K^2 \in \mathcal{K}'$, it has a solution (f_z^1, f_z^2) as follows:

$$f_z^1 = \sum_{i \in \mathbb{N}_{m+u}} \alpha_i K^1(x_i, \cdot) \in \left(\bigcup_{K^1 \in \mathcal{K}} \mathcal{L}_{m+u}^{K^1} \right), \quad (33)$$

$$f_z^2 = \sum_{i \in \mathbb{N}_{m+u}} \beta_i K^2(x_i, \cdot) \in \left(\bigcup_{K^2 \in \mathcal{K}'} \mathcal{L}_{m+u}^{K^2} \right), \quad (34)$$

for some $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_{m+u})^T \in \mathbb{R}^{m+u}$ and $\beta = (\beta_1, \beta_2, \dots, \beta_{m+u})^T \in \mathbb{R}^{m+u}$. That is, the solution belongs to $(\bigcup_{K^1 \in \mathcal{K}} \mathcal{L}_{m+u}^{K^1}) \times (\bigcup_{K^2 \in \mathcal{K}'} \mathcal{L}_{m+u}^{K^2})$.

Proof. The result follows in a similar way to Proposition 2.3.1 in [11]. $\square$

Lemma 5. Let the labeled samples $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i \in \mathbb{N}_m$, be independent random variables drawn from the probability space $(\mathcal{X} \times \mathcal{Y}, P)$ and the unlabeled samples $x_{m+i} \in \mathcal{X}$, $i \in \mathbb{N}_u$, be independent random variables drawn

from the probability space $(\mathcal{X}, P_{\mathcal{X}})$. The following inequality holds

$$\begin{aligned} & E_{\sigma} \left\{ \sup_{f \in \mathcal{H}_{\mathcal{X}, \mathcal{K}'}} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot f(x_i) \right\} \right\} \\ & \leq \frac{1}{2\sqrt{m}} \left\{ \frac{2}{\gamma_1} \mathcal{U}_m(\mathcal{K}) + \frac{2}{\gamma_2} \mathcal{U}_m(\mathcal{K}') + \frac{\kappa^2}{\gamma_1} + \frac{\kappa'^2}{\gamma_2} \right. \\ & \quad - \frac{1}{\sqrt{m}} \cdot E_{\sigma} \inf_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \sigma^T \cdot \left(\left(\frac{1}{\gamma_1} K_{MU}^1 - \frac{1}{\gamma_2} K_{MU}^2 \right) \cdot O \right. \\ & \quad \left. \left. \cdot \Delta_{UU}(\lambda, K_1, K_2) \cdot O^T \right)^2 \cdot \sigma \right\}^{1/2}, \end{aligned} \quad (35)$$

where σ presents a vector $(\sigma_1, \sigma_2, \dots, \sigma_n)$ and $\sigma_i, i \in \mathbb{N}_m$, are Rademacher variables, $(1/\gamma_1) \cdot K_{UU}^1 + (1/\gamma_2) \cdot K_{UU}^2 = O \cdot D_{UU}(K_1, K_2) \cdot O^T$ with diagonal elements $D_{ii}(K_1, K_2) \geq 0, i \in \mathbb{N}_u$, and orthogonal matrix O , and

$$D_{UU}(K_1, K_2) = \begin{pmatrix} d_{11}(K_1, K_2) & & & \\ & d_{22}(K_1, K_2) & & \\ & & \ddots & \\ & & & d_{uu}(K_1, K_2) \end{pmatrix}, \quad (36)$$

$$\Delta_{UU}(\lambda, K_1, K_2) = \begin{pmatrix} \sqrt{\frac{1}{(\lambda^{-1} + d_{11}(K_1, K_2))}} & & & \\ & \sqrt{\frac{1}{(\lambda^{-1} + d_{22}(K_1, K_2))}} & & \\ & & \ddots & \\ & & & \sqrt{\frac{1}{(\lambda^{-1} + d_{uu}(K_1, K_2))}} \end{pmatrix}, \quad (37)$$

Proof. As defined in Equation (17), we can rewrite the left-hand side of inequality (35) as

$$\begin{aligned} & E_{\sigma} \left\{ \sup_{f \in \mathcal{H}_{\mathcal{X}, \mathcal{K}'}} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot f(x_i) \right\} \right\} \\ & = E_{\sigma} \left\{ \sup_{(f, g) \in \mathcal{H}_{\mathcal{X}, \mathcal{K}'}} \left\{ \frac{1}{2m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot [f(x_i) + g(x_i)] \right\} \right\}, \end{aligned} \quad (38)$$

where the sign $\mathcal{H}_{\mathcal{X}, \mathcal{K}'}$ is defined in Equation (16).

From the assumptions, we have

$$\mathcal{L}_{m+u}^{K^1} \triangleq \text{span}\{K^1(x_i) : x_i \in \mathcal{X}, i \in \mathbb{N}_{m+u}\} \subseteq \mathcal{H}_{K^1}, K^1 \in \mathcal{K}, \quad (39)$$

$$\mathcal{L}_{m+u}^{K^2} \triangleq \text{span}\{K^2(x_i) : x_i \in \mathcal{X}, i \in \mathbb{N}_{m+u}\} \subseteq \mathcal{H}_{K^2}, K^2 \in \mathcal{K}'. \quad (40)$$

By the reproducing property and Lemma 1, for any $K^1 \in \mathcal{K}$ and $K^2 \in \mathcal{K}'$, for any $i \in \mathbb{N}_{m+u}$, we have

$$f(x_i) = \left(\text{Proj}_{\mathcal{L}_{m+u}^{K_1}} \right) f(x_i), \quad (41)$$

$$f(x_i) = \left(\text{Proj}_{\mathcal{L}_{m+u}^{K_2}} \right) f(x_i). \quad (42)$$

Combining Equations (39), (40), (41), and (42) yields that

$$\begin{aligned} E_{\sigma} & \left\{ \sup_{(f,g) \in \mathcal{H}_{\mathcal{K}, \mathcal{K}'}} \left\{ \frac{1}{2m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot [f(x_i) + g(x_i)] \right\} \right\} \\ & = E_{\sigma} \sup_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \left\{ \frac{1}{2m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot [f(x_i) + g(x_i)] : (f, g) \right. \\ & \quad \left. \in (\mathcal{L}_{m+u}^{K_1} \times \mathcal{L}_{m+u}^{K_2}) \cap \mathcal{H}_{K^1, K^2} \right\}, \end{aligned} \quad (43)$$

where

$$\begin{aligned} \mathcal{H}_{K_1, K_2} & \triangleq \left\{ (f, g) : \gamma_1 \|f\|_{\mathcal{H}_{K_1}}^2 + \gamma_2 \|g\|_{\mathcal{H}_{K_2}}^2 \right. \\ & \quad \left. + \lambda \sum_{i=m+1}^{m+u} (f(x_i) - g(x_i))^2 \leq 1 \right\}. \end{aligned} \quad (44)$$

By Lemma 4, for any $K_1 \in \mathcal{K}$ and $K_2 \in \mathcal{K}'$, we have (this is similar to Section 5.2.1 converting to Euclidean space in [11])

$$\begin{aligned} (f(x_1), f(x_2), \dots, f(x_m))^T &= (K_{MU}^1 K_{MM}^1) \cdot \alpha, \\ (g(x_1), g(x_2), \dots, g(x_m))^T &= (K_{MU}^2 K_{MM}^2) \cdot \beta, \\ \|f\|_{\mathcal{H}_{K_1}}^2 &= \alpha^T \cdot K^1 \cdot \alpha, \\ \|g\|_{\mathcal{H}_{K_2}}^2 &= \beta^T \cdot K^2 \cdot \beta, \\ \sum_{i=m+1}^{m+u} (f(x_i) - g(x_i))^2 &= (\alpha^T \beta^T) \cdot \begin{pmatrix} K_{UU}^1 \\ K_{MU}^1 \\ -K_{UU}^2 \\ -K_{MU}^2 \end{pmatrix} \\ & \quad \cdot (K_{UU}^1 K_{UM}^1 (-K_{UU}^2) (-K_{UM}^2)) \\ & \quad \cdot \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \end{aligned} \quad (45)$$

where

$$\begin{aligned} K^i &= \begin{pmatrix} K_{UU}^i & K_{UM}^i \\ K_{MU}^i & K_{MM}^i \end{pmatrix}_{(m+u) \times (m+u)}, \quad i = 1, 2, \\ K_{UU}^i &= \begin{pmatrix} K^i(x_{m+1}, x_{m+1}) & \dots & K^i(x_{m+1}, x_{m+u}) \\ K^i(x_{m+2}, x_{m+1}) & \dots & K^i(x_{m+2}, x_{m+u}) \\ \vdots & \ddots & \vdots \\ K^i(x_{m+u}, x_{m+1}) & \dots & K^i(x_{m+u}, x_{m+u}) \end{pmatrix}_{u \times u}, \quad i = 1, 2, \\ K_{UM}^i &= \begin{pmatrix} K^i(x_{m+1}, x_1) & \dots & K^i(x_{m+1}, x_m) \\ K^i(x_{m+2}, x_1) & \dots & K^i(x_{m+2}, x_m) \\ \vdots & \ddots & \vdots \\ K^i(x_{m+u}, x_1) & \dots & K^i(x_{m+u}, x_m) \end{pmatrix}_{u \times m}, \quad i = 1, 2, \\ K_{MU}^i &= \begin{pmatrix} K^i(x_1, x_{m+1}) & \dots & K^i(x_1, x_{m+u}) \\ K^i(x_2, x_{m+1}) & \dots & K^i(x_2, x_{m+u}) \\ \vdots & \ddots & \vdots \\ K^i(x_m, x_{m+1}) & \dots & K^i(x_m, x_{m+u}) \end{pmatrix}_{m \times u}, \quad i = 1, 2, \\ K_{MM}^i &= \begin{pmatrix} K^i(x_1, x_1) & \dots & K^i(x_1, x_m) \\ K^i(x_2, x_1) & \dots & K^i(x_2, x_m) \\ \vdots & \ddots & \vdots \\ K^i(x_m, x_1) & \dots & K^i(x_m, x_m) \end{pmatrix}_{m \times m}, \quad i = 1, 2. \end{aligned} \quad (46)$$

Hence, we have

$$\begin{aligned} E_{\sigma} & \left\{ \sup_{(f,g) \in \mathcal{H}_{\mathcal{K}, \mathcal{K}'}} \left\{ \frac{1}{2m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot [f(x_i) + g(x_i)] \right\} \right\} \\ & = E_{\sigma} \sup_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \left\{ \frac{1}{2m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot [f(x_i) + g(x_i)] : (f, g) \right. \\ & \quad \left. \in (\mathcal{L}_{m+u}^{K_1} \times \mathcal{L}_{m+u}^{K_2}) \cap \mathcal{H}_{K^1, K^2} \right\} \\ & = E_{\sigma} \sup_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \sup_{\substack{\alpha \in \mathbb{R}^{m+u} \\ \beta \in \mathbb{R}^{m+u}}} \\ & \quad \cdot \left\{ \frac{\sigma^T \cdot (K_{MU}^1 K_{MM}^1) \cdot \alpha + \sigma^T \cdot (K_{MU}^2 K_{MM}^2) \cdot \beta}{2m} : (\alpha^T \beta^T) \right. \\ & \quad \cdot \Lambda \cdot \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \leq 1 \left. \right\}, \end{aligned} \quad (47)$$

where

$$\Lambda = \begin{pmatrix} \gamma_1 K^1 & 0 \\ 0 & \gamma_2 K^2 \end{pmatrix} + \lambda \cdot \begin{pmatrix} K_{UU}^1 \\ K_{MU}^1 \\ -K_{UU}^2 \\ -K_{MU}^2 \end{pmatrix} \cdot (K_{UU}^1 K_{UM}^1 (-K_{UU}^2) (-K_{UM}^2)). \quad (48)$$

Note that the matrix Λ is not full rank, and by using the similar steps in [11], we can rewrite Equation (47) as

$$E_\sigma \left\{ \sup_{(f,g) \in \mathcal{H}_{\mathcal{K}, \mathcal{K}'}} \left\{ \frac{1}{2m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot [f(x_i) + g(x_i)] \right\} \right\} \\ = \frac{1}{2m} E_\sigma \sup_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \left\| \Lambda(\widehat{K^1}, K^2)^{-1/2} \cdot K(\widehat{K^1}, K^2)^T \cdot \sigma \right\|, \quad (49)$$

where

$$\Lambda(\widehat{K^1}, K^2) = (K_{MU}^1 \ K_{MM}^1 \ K_{MU}^2 \ K_{MM}^2) \cdot \begin{pmatrix} V & 0 \\ 0 & W \end{pmatrix}, \\ K(\widehat{K^1}, K^2) = \begin{pmatrix} \gamma_1 E_{K_1} & 0 \\ 0 & \gamma_2 E_{K_2} \end{pmatrix} + \lambda \cdot R \cdot R^T, \\ R = \begin{pmatrix} V^T & 0 \\ 0 & W^T \end{pmatrix} \cdot \begin{pmatrix} K_{UU}^1 \\ K_{MU}^1 \\ -K_{UU}^2 \\ -K_{MU}^2 \end{pmatrix}, \quad (50)$$

$$V^T K^1 V = E_{K_1}, \\ W^T K^2 W = E_{K_2},$$

In the above equations, the matrices E_{K_1} and E_{K_2} are, respectively, diagonal matrices containing nonzero eigenvalues. And we write the projections of $\mathbf{a}$ and $\mathbf{b}$ onto the column spaces of K^1 and K^2 as $V \cdot \mathbf{a}$ and $W \cdot \mathbf{b}$.

Next, we try to relate Equation (47) to Rademacher Chaos complexity. Note that

$$K(\widehat{K^1}, K^2) \cdot \Lambda(\widehat{K^1}, K^2)^{-1} \cdot K(\widehat{K^1}, K^2)^T \\ = \frac{1}{\gamma_1} K_{MM}^1 + \frac{1}{\gamma_2} K_{MM}^2 - \lambda \left\{ \left(\frac{1}{\gamma_1} K_{MU}^1 - \frac{1}{\gamma_2} K_{MU}^2 \right) \right. \\ \cdot \left(I + \lambda \left(\frac{1}{\gamma_1} K_{UU}^1 + \frac{1}{\gamma_2} K_{UU}^2 \right) \right)^{-1} \cdot \left(\frac{1}{\gamma_1} K_{UM}^1 - \frac{1}{\gamma_2} K_{UM}^2 \right) \Big\} \\ = \frac{1}{\gamma_1} K_{MM}^1 + \frac{1}{\gamma_2} K_{MM}^2 - \left(\left(\frac{1}{\gamma_1} K_{MU}^1 - \frac{1}{\gamma_2} K_{MU}^2 \right) \cdot O \right. \\ \cdot \Delta_{UU}(\lambda, K_1, K_2) \cdot O^T \Big)^2, \quad (51)$$

The first equation can be easily obtained by the discussion of Section 5.2.4 in [11]. The second equation from

$$\lambda \left(I + \lambda \left(\frac{1}{\gamma_1} K_{UU}^1 + \frac{1}{\gamma_2} K_{UU}^2 \right) \right)^{-1} = O \cdot (\lambda^{-1} \cdot I + D_{UU}(K_1, K_2))^{-1} \cdot O^T \\ = O \cdot \begin{pmatrix} \lambda^{-1} + d_{11}(K_1, K_2) & & & \\ & \lambda^{-1} + d_{22}(K_1, K_2) & & \\ & & \ddots & \\ & & & \lambda^{-1} + d_{UU}(K_1, K_2) \end{pmatrix}^{-1} \cdot O^T \\ = \left(O \cdot \begin{pmatrix} \sqrt{\frac{1}{(\lambda^{-1} + d_{11}(K_1, K_2))}} & & & \\ & \sqrt{\frac{1}{(\lambda^{-1} + d_{22}(K_1, K_2))}} & & \\ & & \ddots & \\ & & & \sqrt{\frac{1}{(\lambda^{-1} + d_{UU}(K_1, K_2))}} \end{pmatrix} \right)^2 \cdot O^T \quad (52)$$

Then, we have

$$\begin{aligned}
& E_{\sigma} \sup_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \left\| \Lambda(\widehat{K^1}, K^2)^{-1/2} \cdot K(\widehat{K^1}, K^2)^T \cdot \sigma \right\| \\
&= E_{\sigma} \sup_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \sqrt{\sigma^T \cdot K(\widehat{K^1}, K^2) \cdot \Lambda(\widehat{K^1}, K^2)^{-1} \cdot K(\widehat{K^1}, K^2)^T \cdot \sigma} \\
&= E_{\sigma} \sup_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \left\{ \sigma^T \cdot \frac{1}{\gamma_1} K_{MM}^1 \cdot \sigma + \sigma^T \cdot \frac{1}{\gamma_2} K_{MM}^2 \cdot \sigma - \sigma^T \cdot \left(\left(\frac{1}{\gamma_1} K_{MU}^1 - \frac{1}{\gamma_2} K_{MU}^2 \right) \cdot O \cdot \Delta_{UU}(\lambda, K_1, K_2) \cdot O^T \right)^2 \cdot \sigma \right\}^{1/2} \\
&= E_{\sigma} \sup_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \left\{ \frac{1}{\gamma_1} \sum_{i,j \in \mathbb{N}_m} \sigma_i \cdot \sigma_j \cdot K^1(x_i, x_j) + \frac{1}{\gamma_2} \sum_{i,j \in \mathbb{N}_m} \sigma_i \cdot \sigma_j \cdot K^2(x_i, x_j) - \sigma^T \cdot \left(\left(\frac{1}{\gamma_1} K_{MU}^1 - \frac{1}{\gamma_2} K_{MU}^2 \right) \cdot O \cdot \Delta_{UU}(\lambda, K_1, K_2) \cdot O^T \right)^2 \cdot \sigma \right\}^{1/2} \\
&\leq E_{\sigma} \sup_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \left\{ \frac{2}{\gamma_1} \sum_{i,j \in \mathbb{N}_m, i < j} \sigma_i \cdot \sigma_j \cdot K^1(x_i, x_j) + \frac{2}{\gamma_2} \sum_{i,j \in \mathbb{N}_m, i < j} \sigma_i \cdot \sigma_j \cdot K^2(x_i, x_j) + \frac{1}{\gamma_1} \sum_{i \in \mathbb{N}_m} K^1(x_i, x_i) + \frac{1}{\gamma_2} \sum_{i \in \mathbb{N}_m} K^2(x_i, x_i) \right. \\
&\quad \left. - \sigma^T \cdot \left(\left(\frac{1}{\gamma_1} K_{MU}^1 - \frac{1}{\gamma_2} K_{MU}^2 \right) \cdot O \cdot \Delta_{UU}(\lambda, K_1, K_2) \cdot O^T \right)^2 \cdot \sigma \right\}^{1/2} \\
&\leq \left\{ \frac{2m}{\gamma_1} U_m(\mathcal{K}) + \frac{2m}{\gamma_2} U_m(\mathcal{K}') + \frac{\kappa^2 m}{\gamma_1} + \frac{\kappa'^2 m}{\gamma_2} - E_{\sigma} \inf_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \sigma^T \cdot \left(\left(\frac{1}{\gamma_1} K_{MU}^1 - \frac{1}{\gamma_2} K_{MU}^2 \right) \cdot O \cdot \Delta_{UU}(\lambda, K_1, K_2) \cdot O^T \right)^2 \cdot \sigma \right\}^{1/2}.
\end{aligned} \tag{53}$$

The second equation uses Equation (51), the forth inequality is obtained by using Jensen's inequality twice, and the last inequality uses the definition of Rademacher chaos complexity and the finite of kernels. $\square$

Proof (Proof of Theorem 1). For any $f \in \widehat{\mathcal{H}}_{\mathcal{K}, \mathcal{K}'}$, it is easy to show that

$$\begin{aligned}
& E_P[L(f(X), Y)] - \sum_{i \in \mathbb{N}_m} L(f(x_i), y_i) \\
&\leq \sup_{f \in \widehat{\mathcal{H}}_{\mathcal{K}, \mathcal{K}'}} \left\{ E_P[L(f(X), Y)] - \sum_{i \in \mathbb{N}_m} L(f(x_i), y_i) \right\},
\end{aligned} \tag{54}$$

Combining Lemmas 2, 3, 4, and 5, the result is as follows. $\square$

Proof (Proof of Corollary 1). By Gershgorin Circle Theorem (Theorem 7.2.1 on page 381 of [18]), the $D_{UU}(K_1, K_2)$ in Equation (21) can be estimated as follows:

$$d_{ii}(K_1, K_2) \leq u \cdot \left(\frac{\kappa^2}{\gamma_1} + \frac{\kappa'^2}{\gamma_2} \right), i \in \mathbb{N}_u, \tag{55}$$

Then, we can write Equation (22) as

$$\begin{aligned}
& \Delta_{UU}(\lambda, K_1, K_2) \geq \Delta_{UU}(\lambda, \kappa, \kappa') \\
&\triangleq \begin{pmatrix} \sqrt{\frac{1}{\left(\lambda^{-1} + u \cdot \left(\kappa^2/\gamma_1 + \kappa'^2/\gamma_2 \right) \right)}} & & & \\ & \sqrt{\frac{1}{\left(\lambda^{-1} + u \cdot \left(\kappa^2/\gamma_1 + \kappa'^2/\gamma_2 \right) \right)}} & & \\ & & \ddots & \\ & & & \sqrt{\frac{1}{\left(\lambda^{-1} + u \cdot \left(\kappa^2/\gamma_1 + \kappa'^2/\gamma_2 \right) \right)}} \end{pmatrix},
\end{aligned} \tag{56}$$

So, the result is as follows. $\square$

FIGURE 1: Semisupervised learning as supervised learning when $u = 0$. And if $\mathcal{K}$ has a single kernel, we think that it is the special case of multiple kernel learning. The scope of the discussion in [2] is the intersection of the green and blue ellipses, the scope of the discussion in [11] is the yellow ellipse, and the cope of the discussion in this paper is the blue ellipse.

Proof (Proof of Corollary 2). By the assumption in Corollary 2, we have

$$\begin{aligned} & \frac{2}{\gamma_1} \mathcal{U}_m(\mathcal{K}) + \frac{2}{\gamma_2} \mathcal{U}_m(\mathcal{K}') + \frac{\kappa^2}{\gamma_1} + \frac{\kappa'^2}{\gamma_2} = \frac{4}{\gamma_1} \mathcal{U}_m(\mathcal{K}) + 2 \frac{\kappa^2}{\gamma_1} \\ & = \frac{4}{\gamma_2} \mathcal{U}_m(\mathcal{K}') + 2 \frac{\kappa'^2}{\gamma_2}, \end{aligned} \quad (57)$$

$$\begin{aligned} & \frac{1}{\sqrt{m}} \cdot E_{\sigma} \inf_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \sigma^T \\ & \cdot \left(\left(\frac{1}{\gamma_1} K_{MU}^1 - \frac{1}{\gamma_2} K_{MU}^2 \right) \cdot O \cdot \Delta_{UU}(\lambda, \kappa, \kappa') \cdot O^T \right)^2 \cdot \sigma = 0. \end{aligned} \quad (58)$$

By substituting Equations (57) and (58) into Equation (24), the Corollary 2 follows. $\square$

6. Discussion

In the above two sections, we give our main results and proof these results. In this section, we will give a comparative discussion of our results to the existing work (in [2] and [11]).

First, we can see that the term

$$\begin{aligned} & \frac{1}{\sqrt{m}} \cdot E_{\sigma} \inf_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \sigma^T \\ & \cdot \left(\left(\frac{1}{\gamma_1} K_{MU}^1 - \frac{1}{\gamma_2} K_{MU}^2 \right) \cdot O \cdot \Delta_{UU}(\lambda, K_1, K_2) \cdot O^T \right)^2 \cdot \sigma \end{aligned} \quad (59)$$

in Theorem 1 reflects the compatibility of two viewers on the unlabeled samples.

- (1) For multiple kernel learning in supervised learning.

If we let $u = 0$ and $\gamma_1 = \gamma_2 = \lambda$, then we have

$$\frac{2}{\gamma_1} \mathcal{U}_m(\mathcal{K}) + \frac{2}{\gamma_2} \mathcal{U}_m(\mathcal{K}') + \frac{\kappa^2}{\gamma_1} + \frac{\kappa'^2}{\gamma_2} = \frac{4}{\lambda} \mathcal{U}_m(\mathcal{K}) + 2 \frac{\kappa^2}{\lambda}, \quad (60)$$

$$\begin{aligned} & \frac{1}{\sqrt{m}} \cdot E_{\sigma} \inf_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \sigma^T \\ & \cdot \left(\left(\frac{1}{\gamma_1} K_{MU}^1 - \frac{1}{\gamma_2} K_{MU}^2 \right) \cdot O \cdot \Delta_{UU}(\lambda, \kappa, \kappa') \cdot O^T \right)^2 \cdot \sigma = 0. \end{aligned} \quad (61)$$

By Corollary (1) and Equations (60) and (61), we can get that

$$\begin{aligned} & E_P[L(f(X), Y)] - \frac{1}{m} \sum_{i \in \mathbb{N}_m} L(f(x_i), y_i) \\ & \leq \frac{L_{\sup}^{\text{loc}}}{\sqrt{m}} \left\{ \frac{4}{\lambda} \mathcal{U}_m(\mathcal{K}) + 2 \frac{\kappa^2}{\lambda} \right\}^{1/2} + 3L_{\sup} \cdot \sqrt{\frac{\ln(2/\delta)}{2m}}. \end{aligned} \quad (62)$$

Then, the main result in [2] recovers from Corollary (1). Therefore, the main result in [2] becomes the special case of Corollary (1).

Remark 5. For the discussion in Section 3, if we set $u = 0$ and $\gamma_1 = \gamma_2 = \lambda$ and by combining Equation (17), then we have that Equation (14) reduces to Equation (13). Furthermore, let $|\mathcal{K}| = |\mathcal{K}'| = 1$ and $\mathcal{K} = \mathcal{K}'$, and we have

$$\mathcal{U}_m(\mathcal{K}) = \mathcal{U}_m(K) = \frac{2}{m} \sum_{i \in \mathbb{N}_m} K(x_i, x_i). \quad (63)$$

Substituting inequality (63) into inequality (62), we can obtain the generalization bound for the single kernel learning in the framework of supervised learning as follows:

$$\begin{aligned} E_P[L(f(X), Y)] - \frac{1}{m} \sum_{i \in \mathbb{N}_m} L(f(x_i), y_i) \\ \leq \frac{L_{\sup}^{\text{loc}}}{\sqrt{m}} \left\{ \frac{8}{m\lambda} \sum_{i \in \mathbb{N}_m} K(x_i, x_i) + 2 \frac{\kappa^2}{\lambda} \right\}^{1/2} + 3L_{\sup} \cdot \sqrt{\frac{\ln(2/\delta)}{2m}}. \end{aligned} \quad (64)$$

(2) For coregularized kernel learning in semisupervised learning.

If we let $K_1 \in \mathcal{K}$, $|\mathcal{K}| = 1$, $K_2 \in \mathcal{K}'$, and $|\mathcal{K}'| = 1$, by equation we have

$$\begin{aligned} E_{\sigma} \left\{ \sup_{(f,g) \in \mathcal{H}_{\mathcal{K}, \mathcal{K}'}} \left\{ \frac{1}{2m} \sum_{i \in \mathbb{N}_m} \sigma_i \cdot [f(x_i) + g(x_i)] \right\} \right\} \\ = \frac{1}{2m} E_{\sigma} \sup_{\substack{K^1 \in \mathcal{K} \\ K^2 \in \mathcal{K}'}} \left\| \Lambda(\widehat{K^1}, K^2)^{-1/2} \cdot K(\widehat{K^1}, K^2)^T \cdot \sigma \right\| \quad (65) \\ = \frac{1}{2m} E_{\sigma} \left\| \Lambda(\widehat{K^1}, K^2)^{-1/2} \cdot K(\widehat{K^1}, K^2)^T \cdot \sigma \right\|. \end{aligned}$$

And note that Equation (65) is the same as the supremum evaluation in Section 5.2.2 in [11]. So, the main result in [11] recovers from Theorem 1. Then, we have that the main result in [11] can be regarded as the special case of Theorem 1.

Remark 6. For the discussion in Section 3, if we set $K_1 \in \mathcal{K}$, $|\mathcal{K}| = 1$, $K_2 \in \mathcal{K}'$, and $|\mathcal{K}'| = 1$, then we have that Equation (14) reduces to Equation (12).

In Figure 1, we show the relations between the main results in this paper and the results in [2] and [11].

7. Conclusion

In this paper, based on semisupervised two-viewers learning, we study the generalization bound of coregularized multiple kernel learning. The main research tool is Rademacher chaos complexity which we use to control the performance of the candidate class of coregularized multiple kernels. In this paper, we mainly blend the work in [2] and [11] to discuss the generalization error of coregularized multiple kernel learning in the semisupervised multiview learning framework. First, we discuss the differences between the learning question proposed by us and the learning questions in [2] and [11]. Then, we analyze the generalization bound of coregularized multiple kernel learning in the semisupervised multiview learning framework. And we show that the existing results in [2] and [11] can be regarded as the special cases of our main results.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Shanghai Talents Development Funds (Grant no. 201629).

References

- [1] G. R. G. Lanckriet, N. Christianini, P. L. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semi-definite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.
- [2] Y. Ying and C. Campbell, "Rademacher chaos complexities for learning the kernel problem," *Neural Computation*, vol. 22, no. 11, pp. 2858–2886, 2010.
- [3] C. A. Micchelli and M. Pontil, "Learning the kernel function via regularization," *Journal of Machine Learning Research*, vol. 6, pp. 1099–1125, 2005.
- [4] N. Srebro and S. Ben-David, "Learning bounds for support vector machines with learned kernels," in *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 169–183, Springer, Berlin, Heidelberg, Germany, December 2006.
- [5] Y. Lei and L. Ding, "Refined rademacher chaos complexity bounds with applications to the multikernel learning problem," *Neural Computation*, vol. 26, no. 4, pp. 739–760, 2014.
- [6] M. Kloft and G. Blanchard, "The local rademacher complexity of ℓ_p -norm multiple kernel learning," in *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 2438–2446, Granada Congress and Exhibition Centre, Granada, Spain, December 2011.
- [7] C. Cortes, M. Kloft, and M. Mohri, "Learning kernels using local rademacher complexity," in *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 2760–2768, Curran Associates Inc., Lake Tahoe, NV, USA, December 2013.
- [8] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the International Conference on Computational Learning Theory*, pp. 92–100, ACM, Madison, WI, USA, July 1998.
- [9] V. Sindhwani, P. Niyogi, and M. Belkin, "A co-regularized approach to semi-supervised learning with multiple views," in *Proceedings of the International Conference on Machine Learning Workshop on Learning with Multiple Views*, pp. 824–831, ACM, Bonn, Germany, August 2005.
- [10] D. S. Rosenberg, *Kernel Co-Training: An Algorithm for Semi-Supervised Learning*, Technical report, Statistics Department Student SeminarTalk, University of California at Berkeley, 2006.
- [11] D. S. Rosenberg and P. L. Bartlett, "The rademacher complexity of co-regularized kernel classes," in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, vol. 2, pp. 396–403, San Juan, Puerto Rico, March 2007.
- [12] U. Brefeld, T. Scheffer, and S. Wrobel, "Efficient co-regularised least squares regression," in *Proceedings of the*

- International Conference on Machine Learning*, pp. 137–144, ACM, Pittsburgh, PA, USA, 2006.
- [13] D. S. Rosenberg, *Semi-Supervised Learning with Multiple Views*, Ph.D. thesis, University of California, Berkeley, 2008.
  - [14] A. Berline and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Springer Science+Business Media, New York, NY, USA, 2004.
  - [15] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, The Massachusetts Institute of Technology Press, Cambridge, MA, USA, 2012.
  - [16] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, Cambridge, UK, 2014.
  - [17] R. Meir and T. Zhang, “Generalization error bounds for bayesian mixture algorithms,” *Journal of Machine Learning Research*, vol. 4, no. 5, pp. 839–860, 2003.
  - [18] G. Golub and C. Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MA, USA, 4th edition, 2013.