*Appl. Math. J. Chinese Univ.*
2024, 39(3): 533-544

# Convergence analysis for complementary-label learning with kernel ridge regression

NIE Wei-lin[1] WANG Cheng[1] XIE Zhong-hua[2]

**Abstract**. Complementary-label learning (CLL) aims at finding a classifier via samples with complementary labels. Such data is considered to contain less information than ordinary-label samples. The transition matrix between the true label and the complementary label, and some loss functions have been developed to handle this problem. In this paper, we show that CLL can be transformed into ordinary classification under some mild conditions, which indicates that the complementary labels can supply enough information in most cases. As an example, an extensive misclassification error analysis was performed for the Kernel Ridge Regression (KRR) method applied to multiple complementary-label learning (MCLL), which demonstrates its superior performance compared to existing approaches.

## §1 Introduction

Weakly supervised learning problems have attracted great interest in machine learning society since high-quality supervised data sometimes is difficult or expensive to obtain. One notable instance of such a problem is CLL, which was introduced by Ishida et al. (2017) [8]. In CLL, the training sample is labeled with a class it does not belong to, and the goal is to learn a proper classifier as in the traditional multi-class classification problem.

We follow the settings in Cucker and Smale, (2002) [3]. Let the input space be $X \subset \mathbb{R}^d$ and the output space $Y = \{e_1, \cdots, e_n\}$ where $e_i$ is the vector with the $i$-th element 1 and others 0, standing for the $i$-th class. For any classifier $f : X \to Y$, the misclassification error is $\Pr_{X\times Y}\{f(x) \neq y\} = \mathbb{E}I_{f(x)\neq y}$ where $I$ is the indicator function. In traditional problems, samples are drawn according to a joint distribution $\rho$ defined on the sample space $Z := X \times Y$.

Received: 2024-03-04. Revised: 2024-04-03.
MR Subject Classification: 68T05, 68Q32, 62J02, 41A46.
Keywords: multiple complementary-label learning, partial label learning, error analysis, reproducing kernel Hilbert spaces.
Digital Object Identifier(DOI): https://doi.org/10.1007/s11766-024-5173-6.
Supported by the Indigenous Innovation's Capability Development Program of Huizhou University (HZU202003, HZU202020), Natural Science Foundation of Guangdong Province(2022A1515011463), the Project of Educational Commission of Guangdong Province(2023ZDZX1025), National Natural Science Foundation of China(12271473), and Guangdong Province's 2023 Education Science Planning Project (Higher Education Special Project)(2023GXJK505).

In the CLL problem, the true labels of the training samples are always unknown. Different from most literature, we denote the complementary label $\tilde{y}$ as a vector in $\{0,1\}^n$. 0 means a complementary label, and 1 indicates a possible true label. We should be aware that the complementary output space is $\tilde{Y} = \{\vec{1} - e_A, A \subset \{0,1,\cdots,n\}\}$. Here $\vec{1} = (1,\cdots,1) \in \mathbb{R}^n$, $A$ is an index set, and we denote $e_A = (a_1,\cdots,a_n)$ where $a_i = 1$ if $i \in A$ and $a_i = 0$ otherwise. Then the complementary-label samples are drawn from another distribution $\tilde{\rho}$ on the space $\tilde{Z} : X \times \tilde{Y}$. Also, we can consider the sample as $(x, y, \tilde{y})$ where the true label $y$ is a latent variable.

The generation method of complementary labels is important in CLL problems. Ishida et al., (2017) [8] considered a single complementary label for each sample. They assumed the complementary label is uniformly distributed. That is, all the labels other than the true label has the same probability to be the complementary label. Then one versus all (OVA) and pairwise comparison (PC) loss were applied in an empirical risk minimization (ERM) scheme. Later Yu et al. (2018) [17] studied a transition matrix $Q$ with different elements $Q_{ij} = \Pr\{\bar{y} = j | y = i\}$, which extended the above setting. Here $\bar{y}$ denotes the complementary label. A new loss function was introduced to deal with this condition, and convergence analysis was conducted. Feng et al., (2020) [6] introduced the multiple CLL (MCLL) problem, in which the data contain multi-complementary labels. But for a fixed size of complementary-label set, labels in this set were still uniformly distributed. Recently, Wang et al. (2023) [13] proposed a mild condition, stating that the sampling of complementary labels is independent of the input variable $x$. Based on this condition, the authors reformulated the expected risk and addressed the problem by minimizing the empirical risk.

After generating the complementary-label samples, the above literature and Ishida et al., (2019) [9] designed several complementary losses, to approximate the conditional distribution $\rho(y|x)$ given $x$ via the relationship $Q$. Different from these analyses, Gao et al., (2021) [7] directly modelled $\rho(\bar{y}|x)$ by minimizing the gap between $\bar{y}$ and $1 - f(x)$ where $f$ is the prediction function. Xu et al., (2020) [16] studied the two conditional distributions simultaneously, by a ganerative adversarial net (GAN).

Theoretical analysis of classifier consistency was considered in more recent literature. In Wang et al., (2023) [12], the authors considered least-square-based loss to single uniformly distributed CLL. They proved the ERM estimator with respect to their loss is consistent with a traditional ERM problem with the least square loss and ordinary labels. Both Wang et al., (2023) [13] and Liu et al., (2023) [10] claimed that their estimators are consistent with the Bayes classifier. The latter paper further conducted a regret transfer bound, which leads to an estimation for the misclassification error.

It is also of theoretical interest to derive a sharp convergence rate for CLL problems. $O_p(n^2 m^{-1/2})$ rate was obtained in Ishida et al., (2017) [8] for the excess generalization error. When each size $m_i$ of samples with complementary-label $\bar{y} = i$ is $m/n$, Yu et al., (2018) [17] stated a $O_p(n^{3/2} m^{-1/2})$ rate. For MCLL, Feng et al., (2020) [6] presented an excess generalization error $O_p(n \sum_{j=1}^{n} m_j^{-1/2})$. There $m_j$ is the number of examples whose complementary-label

size is $j$. Wang et al., (2023) [13] proved a sharper bound $O_p(\sum_{j=1}^{n} m_j^{-1/2} + m_j'^{-1/2})$ where $m_j + m_j' = m$.

In summary, most of the existing literature considers the (multi) complementary label contain less information than the ordinary-label, as the lack of true label in the sample data. Hence restrictive conditions on the sample distribution are required. And complicated loss functions are needed to derive consistent estimators. In this paper, we propose a weak condition for the distribution of the complementary label. Under this condition, we show that MCLL can be equivalent to learning with ordinary labels. Furthermore, if the conditional distribution $\tilde{\rho}(\tilde{y}|x)$, given $x$ (to be defined in the next section), satisfies an additional mild condition, a comparison theorem can be established. As an example, we focus on applying the kernel ridge regression (KRR) method to the MCLL, and we provide optimal bounds for both the excess generalization error and the excess misclassification error associated with the KRR method.

## §2 Bayes classification consistency

In this section, we prove that CLL can be transformed to ordinary-label learning under weak condition on $\rho(\tilde{y}|x)$. Firstly we recall the basic settings as follows. $X \subset \mathbb{R}^d$ is the input space, the output space is $Y = \{e_1, \cdots, e_n\}$ where $e_k$ stands for the $k$-th class. For any input $x \in X$, we assume the true label is $y \in Y$, and the complentary-label is $\tilde{y} \in \tilde{Y}$, where $\tilde{Y} = \{\vec{1} - e_A, A \subset \{1, \cdots, n\}\}$. A complementary label $\tilde{y} = \vec{1} - e_A$ means the input does not belong to the class with indexes in $A$. In the sequel, we denote $y^{(k)}$ as the $k$-th element of vector $y$, and $y_i$ as the (latent) output of the $i$-th sample.

We assume the true sample data $\mathbf{z} = (x_i, y_i)_{i=1}^m \in Z^m := (X \times Y)^m$ is drawn according to a joint distribution $\rho$ on $Z$. The marginal distribution of $\rho$ is $\rho_X$ and the conditional distribution is $\rho(y|x)$ on $Y$. And the complementary-label sample we obtained is $\tilde{\mathbf{z}} = (x_i, \tilde{y}_i)_{i=1}^m \in \tilde{Z}^m := (X \times \tilde{Y})^m$, which is drawn according to another joint distribution $\tilde{\rho}$ on $X \times \tilde{Y}$. $\tilde{\rho}$ can be decomposed to $\rho_X$ and $\tilde{\rho}(\tilde{y}|x)$ as well. A classifier can be considered as a function $g : X \to Y$. However, in practice, we often find some scoring function $f : X \to \mathbb{R}^n$ from data, then use a splitter function to derive a classifier. The splitter function is $S(\alpha) = e_{j_\alpha}$ where $j_\alpha = \arg\max_{k=1,\cdots,n} \alpha^{(k)}$ for a score vector $\alpha \in \mathbb{R}^n$.

The Bayes classifier $f_c(x) = e_{j_\rho(x)}$, with $j_\rho(x) = \arg\max_{j=1,\cdots,n} \Pr\{y = e_j|x\}$, is the minimizer of the misclassification error $R(f) = \Pr\{f(x) \neq y\}$, which is our goal function. It is also known that if we denote the regression function by $f_\rho(x) = \mathbb{E}(y|x) = (\Pr\{y^{(k)} = 1|x\})_{k=1}^n$, then $f_c(x) = S(f_\rho(x))$. Hence we would like to estimate the performance of a classifier $g$ by the excess misclassification error $R(g) - R(f_c) = R(g) - R(S(f_\rho))$.

To this end, we need some conditions on the generation for the complementary-label set.

**Assumption 1.** *Assume the largest element with index $j_\rho(x)$ of $f_\rho(x)$ is unique. And*

$$\Pr\{\tilde{y}^{(k)} = 0|x\} > \Pr\{\tilde{y}^{(j_\rho(x))} = 0|x\}, \quad k = 1, \cdots, n. \tag{1}$$

The uniqueness assumption is often considered necessary and practical because, in many real-world scenarios, the true label for a given instance is not ambiguous. The formula (1)

means the true label has the least probability of appearing in the complementary-label set, which is also realistic. Compared with the existing assumptions, we can see (1) is much weaker.

**Example 1.** *In Ishida et al., (2017) [8],* $\Pr\{\bar{y}^{(k)} = 1|x, y^{(k)} = 1\} = 0$, *and* $\Pr\{\bar{y}^{(k')} = 1|x, y^{(k)} = 1\} = \frac{1}{n-1}$ *for any* $k' \neq k$ *were assumed. Then from the notation* $\tilde{y} = \vec{1} - \bar{y}$, *i.e.,* $\tilde{y}^{(k)} = 1 - \bar{y}^{(k)}$, *we deduce that*

$$\begin{aligned}
&\Pr\{\tilde{y}^{(k)} = 0|x\} = \sum_{j=1}^{n} \Pr\{\tilde{y}^{(k)} = 0, y^{(j)} = 1|x\} \\
&= \sum_{j=1}^{n} \Pr\{\tilde{y}^{(k)} = 0|x, y^{(j)} = 1\} \cdot \Pr\{y^{(j)} = 1|x\} \\
&= \sum_{j=1, j\neq k}^{n} \frac{1}{n-1} \cdot f_\rho^{(j)}(x) = \frac{1}{n-1}\Big[\sum_{j=1}^{n} f_\rho^{(j)}(x) - f_\rho^{(k)}(x)\Big] \\
&= \frac{1}{n-1}(1 - f_\rho^{(k)}(x)).
\end{aligned}$$

*If* $k = j_\rho(x)$, *the above formula is the smallest one as* $j_\rho(x) = \arg\max_k f_\rho^{(k)}(x)$.

It is worth noting that our assumption covers cases that have not been previously considered. For example, the distribution of complementary-label can depend on the input $x$, $\Pr\{\bar{y}^(k) = 1|y^(k) = 1\}$ for some $k$ can be positive, which means people may make a mistake when annotating a label for an input. $\Pr\{\bar{y} = \vec{0}|x\}$ or $\Pr\{\bar{y} = \vec{1}|x\}$ can be positive too, which allows a complementary-label to be empty and universal set. Now we prove the main result in this section that MCLL under such conditions is consistent with the Bayes classifier. This implies that any classification method designed for ordinary-label samples can also be applied to MCLL problems. Furthermore, our assumption may be extended to other partial label learning problems, Cour et al., (2011) [2] and etc..

**Theorem 1.** *With the notations above and Assumption 1, we have* $S(\tilde{f}_\rho(x)) = S(f_\rho(x))$, *where* $\tilde{f}_\rho(x) = \mathbb{E}(\tilde{y}|x)$.

*Proof.* From the definition

$$\begin{aligned}
\tilde{f}_\rho(x) = \mathbb{E}(\tilde{y}|x) &= \sum_{A\subset[n]} (\vec{1} - e_A) \Pr\{\tilde{y} = \vec{1} - e_A|x\} \\
&= \sum_{A\subset[n]} \vec{1} \cdot \Pr\{\tilde{y} = \vec{1} - e_A|x\} - \sum_{A\subset[n]} e_A \cdot \Pr\{\tilde{y} = \vec{1} - e_A|x\}.
\end{aligned}$$

Here $[n] = \{1, 2, \cdots, n\}$. Then the $k$-th element of $\tilde{f}_\rho(x)$ is

$$\sum_{A\subset[n]} \Pr\{\tilde{y} = \vec{1} - e_A|x\} - \sum_{A\subset[n]} I_{k\in A} \Pr\{\tilde{y} = \vec{1} - e_A|x\}.$$

Note that the second term is

$$\sum_{\substack{A\subset[n],\\ k\in A}} \Pr\{\tilde{y} = \vec{1} - e_A|x\} = \Pr\{\tilde{y}^{(k)} = 0|x\},$$

we have $\tilde{f}_\rho^{(j_\rho(x))}(x) > \tilde{f}_\rho^{(k)}(x)$ for any $k \neq j_\rho(x)$ from Assumption 1 where $j_\rho(x) = \arg\max_j f_\rho^{(j)}(x)$. Hence the result is proved. □

## §3 Comparison Theorem

Assumption 1 and Theorem 1 indicate that the index of the largest element in $\tilde{f}_\rho$ should be the same as the one in $f_\rho$, i.e., $j_\rho(x) = \tilde{j}_\rho(x)$. However, it is not enough to get a satisfactory convergence result with only this condition. Consider a single CLL problem, for a given input $x$, if the complementary-label distribution satisfies $\tilde{f}_\rho^{(k)}(x) = \frac{1}{n} - \epsilon$ for all $k \neq \tilde{j}_\rho(x)$ and $\tilde{f}_\rho^{(\tilde{j}_\rho(x))}(x) = \frac{1}{n} + (n-1)\epsilon$ for a very small $\epsilon > 0$. Then $\tilde{f}_\rho^{(j_\rho(x))}(x) - \tilde{f}_\rho^{(k)}(x) = n\epsilon$. In this case, we can hardly find the true label correctly since the largest element is not prominent enough in the conditional probabilities $\tilde{f}_\rho^{(k)}(x)$. This simple example suggests the distribution for the complementary label should not be close to a uniform one. Motivated by this observation, we introduce another condition below.

**Assumption 2.** *For any $1 \leq k \leq n$ and $x \in X$, we assume*

$$f_\rho^{(j_\rho(x))}(x) - f_\rho^{(k)}(x) \leq \alpha(\tilde{f}_\rho^{(\tilde{j}_\rho(x))}(x) - \tilde{f}_\rho^{(k)}(x))$$

*for some constant $\alpha > 0$ where $j_\rho(x) = \arg\max_j f_\rho^{(j)}(x)$ and $\tilde{j}_\rho(x) = \arg\max_j \tilde{f}_\rho^{(j)}(x)$.*

This is a weak condition for the distribution of complementary-label as well. Indeed, a trivial bound holds that $f_\rho^{(j_\rho(x))}(x) - f_\rho^{(k)}(x) \leq 1$ for any $k$ and $x$. If $\tilde{f}_\rho^{(\tilde{j}_\rho(x))}(x) - \tilde{f}_\rho^{(k)}(x) \geq \epsilon$ for any $k$ and $x$, then Assumption 2 holds with $\alpha = \frac{1}{\epsilon}$. This means the first and second largest element in $\tilde{f}_\rho(x)$ should not be too close, which is a mild condition.

A comparison theorem can be proved under this condition for least squares loss. We denote the generalization error $\mathcal{E}(f) := \int_Z \|f(x) - y\|_2^2 d\rho = \int_Z \sum_{j=1}^n (f^{(j)}(x) - y^{(j)})^2 d\rho$ and $\tilde{\mathcal{E}}(f) = \int_{\tilde{Z}} \|f(x) - \tilde{y}\|_2^2 d\tilde{\rho}$. We can verify that $\tilde{\mathcal{E}}(f) - \tilde{\mathcal{E}}(\tilde{f}_\rho) = \int_X \|f(x) - \tilde{f}_\rho(x)\|_2^2 d\rho_X := \|f - \tilde{f}_\rho\|_{2,\rho}^2$ and $\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_{2,\rho}^2$.

**Theorem 2.** *(Comparison) With the notations above and Assumption 1,2, we have for any $f : X \to \mathbb{R}^n$,*

$$R(S(f)) - R(f_c) \leq \alpha\sqrt{2(\tilde{\mathcal{E}}(f) - \tilde{\mathcal{E}}(\tilde{f}_\rho))}.$$

*Proof.* Note that from the previous section, $S(\tilde{f}_\rho) = S(f_\rho) = f_c$, hence

$$\begin{aligned}
&R(S(f)) - R(S(\tilde{f}_\rho)) = \Pr\{y \neq S(f(x))\} - \Pr\{y \neq S(\tilde{f}_\rho(x))\} \\
&= \Pr\{y = S(\tilde{f}_\rho(x))\} - \Pr\{y = S(f(x))\} \\
&= \int_X \Big[\Pr\{y = S(\tilde{f}_\rho(x))|x\} - \Pr\{y = S(f(x))|x\}\Big] d\rho_X \\
&= \int_X \Big[\Pr\{y = e_{tildej_\rho(x)}|x\} - \Pr\{y = e_{j_f(x)}|x\}\Big] d\rho_X
\end{aligned}$$

where $j_f(x) = \arg\max_j f^{(j)}(x)$. On the other hand,

$$f_\rho(x) = \mathbb{E}(y|x) = \sum_{k=1}^n e_k \Pr\{y = e_k|x\} = \Big(\Pr\{y = e_k|x\}\Big)_{k=1}^n.$$

Hence

$$R(S(f)) - R(S(\tilde{f}_\rho)) = \int_X \Big[f_\rho^{(\tilde{j}_\rho(x))}(x) - f_\rho^{(j_f(x))}(x)\Big] d\rho_X.$$

By Theorem 1 $\tilde{j}_\rho(x) = j_\rho(x)$ for any $x \in X$ and Assumption 2 we have

$$R(S(f)) - R(S(\tilde{f}_\rho)) \le \alpha \int_X \left[\tilde{f}_\rho^{(\tilde{j}_\rho(x))}(x) - \tilde{f}_\rho^{(j_f(x))}(x)\right] d\rho_X.$$

From Jensen's inequality,

$$\begin{aligned}
&\tilde{f}_\rho^{(\tilde{j}_\rho(x))}(x) - \tilde{f}_\rho^{(j_f(x))}(x) \\
&= [\tilde{f}_\rho^{(\tilde{j}_\rho(x))}(x) - f^{(\tilde{j}_\rho(x))}(x)] + [f^{(\tilde{j}_\rho(x))}(x) - f^{(j_f(x))}(x)] + [f^{(j_f(x))}(x) - \tilde{f}_\rho^{(j_f(x))}(x)] \\
&\le [\tilde{f}_\rho^{(j_{\tilde{\rho}}(x))}(x) - f^{(j_{\tilde{\rho}}(x))}(x)] + [f^{(j_f(x))}(x) - \tilde{f}_\rho^{(j_f(x))}(x)] \\
&\le \sqrt{2\sum_{k=1}^{n}(f^{(k)}(x) - \tilde{f}_\rho^{(k)}(x))^2},
\end{aligned}$$

we deduce that

$$\begin{aligned}
R(S(f)) - R(S(\tilde{f}_\rho)) &\le \alpha \sqrt{2\int_X \sum_{k=1}^{n}(f^{(k)}(x) - \tilde{f}_\rho^{(k)}(x))^2 d\rho_X} \\
&= \alpha \sqrt{2\int_X \|f(x) - \tilde{f}_\rho(x)\|_2^2 d\rho_X}
\end{aligned}$$

which proves the result from the fact that $\tilde{\mathcal{E}}(f) - \tilde{\mathcal{E}}(\tilde{f}_\rho) = \|f - \tilde{f}_\rho\|_{2,\rho}^2$. □

## §4 Error analysis

In this section, we focus on analyzing the error bound for the classical Kernel Ridge Regression (KRR) method applied to the MCLL problem. Let $K$ be a Mercer kernel defined on $X \times X$, which is continuous, symmetric and positive semi-definite. $\mathcal{H}_K := \overline{\{span(K_x) : x \in X\}}$ is the reproducing kernel Hilbert space (RKHS) where $K_x(t) = K(x,t),\ x,t \in X$. Denote $\kappa := \sup_{x\in X}\sqrt{K(x,x)}$ and $\mathcal{H}_K^n = \{(f^{(1)}, \cdots, f^{(n)}) : f^{(k)} \in \mathcal{H}_K, k = 1, \cdots, n\}$. The KRR scheme is defined as

$$\tilde{f}_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}_K^n} \frac{1}{m}\sum_{i=1}^{m} \|f(x_i) - \tilde{y}_i\|_2^2 + \lambda\|f\|_{\vec{K}}^2,$$

where $\|f\|_{\vec{K}}^2 = \sum_{k=1}^{n}\|f^{(k)}\|_K^2$. Then a classifier can be derived by splitter function $S(\tilde{f}_{\mathbf{z}})$.

We would like to estimate the excess misclassification error for $S(\tilde{f}_{\mathbf{z}})$ in this section. To this end, we have to introduce some notations as in the classical learning theory as Wang and Guo (2012), [11].

**Definition 1.** *For a metric space $(\mathcal{H}, D)$, and $\mathcal{F} \subset \mathcal{H}$, the covering number $\mathcal{N}(\mathcal{F}, \eta, D)$ is defined to be the minimal integer $N$, such that there exists a function set $f_1, \cdots, f_N$, for any $f \in \mathcal{F}$, we can find some $j \in \{1, 2, \cdots, N\}$ satisfying $D(f, f_j) \le \eta$.*

**Assumption 3.** *Denote $B_1 = \{f \in \mathcal{H}_K : \|f\|_K \le 1\}$, we assume*

$$\log\mathcal{N}(B_1, \eta, \|\cdot\|_\infty) \le c_s\eta^{-s}.$$

Denote integral operator $L_K g(t) := \int_Z g(x)K(x,t)d\rho_X$. And $L_K^r\sum_{i\ge1} c_i\phi_i = \sum_{i\ge1} c_i\mu_i\phi_i$ where $\{\mu_i, \phi_i\}_{i\ge1}$ are the eigen-pairs of $L_K$ and $\{\phi_i\}_{i\ge1}$ form an orthogonal basis of $L_{\rho_X}^2$. Then

the regularity condition is as follows.

**Assumption 4.** *Assume* $\tilde{f}_\rho^{(k)} \in L_K^r(L^2_{\rho_X})$ *with some constant* $r > 0$. *And*

$$M_\rho := \max_{1\le k\le n} \|L_K^{-r}\tilde{f}_\rho^{(k)}\|_\rho.$$

These two assumptions are general in the learning theory literature as in Wu and Zhou, (2006) [14]. The first one describes the capacity for the hypothesis space, and the second one characterize the regularity of the goal function. We also need the Bernstein inequality from Bennett, (1962) [1] for the error analysis.

**Lemma 1.** *Let* $\xi$ *be a random variable on a probability space* $Z$ *with variance* $\sigma^2(\xi) = \sigma^2$, *and satisfy* $|\xi(z) - \mathbb{E}\xi| \le M$ *for almost all* $z \in Z$. *Then for all* $\varepsilon > 0$,

$$\Pr\Big\{\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbb{E}\xi \ge \varepsilon\Big\} \le \exp\Big\{-\frac{m\varepsilon^2}{2(\sigma^2 + \frac{1}{3}M\varepsilon)}\Big\}.$$

Set the right-hand side to $\delta$, we can prove that

$$\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbb{E}\xi \le \Big(\frac{2M}{3m} + \sqrt{\frac{2\sigma^2}{m}}\Big)\log\frac{1}{\delta} \tag{2}$$

holds with confidence $1-\delta$. Now we state the convergence result as follows.

**Theorem 3.** *Under assumptions 1, 2, 3, 4, let* $\lambda = m^{-\frac{1}{(1+s)\min\{2r+1,2\}}}$, *then with confidence* $1-\delta$, *we have*

$$R(S(\tilde{f}_{\mathbf{z}})) - R(f_c) \le \tilde{C}\sqrt{n}\log^{\frac{1}{2}}\frac{4}{\delta}\cdot m^{-\min\{\frac{1}{4(1+s)},\frac{r}{(1+s)(2r+1)}\}}$$

*for some constant* $\tilde{C}$ *independent of* $m, n$ *and* $\delta$.

*Proof.* Firstly we note that from the previous section there holds

$$R(S(\tilde{f}_{\mathbf{z}})) - R(f_c) \le \alpha\sqrt{2(\tilde{\mathcal{E}}(\tilde{f}_{\mathbf{z}}) - \tilde{\mathcal{E}}(\tilde{f}_\rho))}$$

for the least squares loss. So what is left is to estimate the excess generalization error.

We denote a stepping-stone function $\tilde{f}_\lambda = (\tilde{f}_\lambda^{(1)}, \cdots, \tilde{f}_\lambda^{(n)})$ where

$$\tilde{f}_\lambda^{(k)} = \arg\min_{f\in\mathcal{H}_K}\int_{\tilde{Z}}(f(x)-\tilde{y})^2 d\tilde{\rho} + \lambda\|f\|_K^2, \quad k = 1, \cdots, n.$$

Then the error decomposition can be stated as follows.

$$\begin{aligned}
&\tilde{\mathcal{E}}(\tilde{f}_{\mathbf{z}}) - \tilde{\mathcal{E}}(\tilde{f}_\rho) \le \tilde{\mathcal{E}}(\tilde{f}_{\mathbf{z}}) - \tilde{\mathcal{E}}(\tilde{f}_\rho) + \lambda\|\tilde{f}_{\mathbf{z}}\|^2_{\vec{K}}\\
&\le \Big[\Big(\tilde{\mathcal{E}}(\tilde{f}_{\mathbf{z}}) - \mathcal{E}_{\tilde{\mathbf{z}}}(\tilde{f}_\rho)\Big) - \Big(\mathcal{E}_{\tilde{\mathbf{z}}}(\tilde{f}_{\mathbf{z}}) - \mathcal{E}_{\tilde{\mathbf{z}}}(\tilde{f}_\rho)\Big)\Big]\\
&+\Big[\Big(\mathcal{E}_{\tilde{\mathbf{z}}}(\tilde{f}_\lambda) - \mathcal{E}_{\tilde{\mathbf{z}}}(\tilde{f}_\rho)\Big) - \Big(\tilde{\mathcal{E}}(\tilde{f}_\lambda) - \tilde{\mathcal{E}}(\tilde{f}_\rho)\Big)\Big]\\
&+\Big[\tilde{\mathcal{E}}(\tilde{f}_\lambda) - \tilde{\mathcal{E}}(\tilde{f}_\rho) + \lambda\|\tilde{f}_\lambda\|^2_{\vec{K}}\Big] := S_1 + S_2 + \tilde{D}(\lambda)
\end{aligned}$$

where $\mathcal{E}_{\tilde{\mathbf{z}}}(f) = \frac{1}{m}\sum_{i=1}^m \|f(x_i) - \tilde{y}_i\|_2^2$. The first two terms are sample errors and the third one is regularization error.

It is known that $\tilde{f}_\lambda^{(k)} = (L_K + \lambda I)^{-1} L_K \tilde{f}_\rho^{(k)}$ from classical analysis, Cucker and Zhou, (2007) [4]. Since

$$\tilde{\mathcal{E}}(f_\lambda) - \tilde{\mathcal{E}}(\tilde{f}_\rho) = \sum_{k=1}^{n} \|\tilde{f}_\lambda^{(k)} - \tilde{f}_\rho^{(k)}\|_\rho^2 \le \lambda^{\min\{2r,2\}}(\kappa^2+1)M_\rho n$$

and

$$\lambda \|\tilde{f}_\lambda\|_{\vec{K}}^2 = \lambda \sum_{k=1}^{n} \|\tilde{f}_{\lambda,k}\|_K^2 \le \lambda^{\min\{2r,1\}}(\kappa^{4r-2}+1)M_\rho^2 n,$$

we can estimate the regularization error bound

$$\tilde{D}(\lambda) \le C_1 \lambda^{\min\{2r,1\}} n,$$

where $C_1 = (\kappa^2 + \kappa^{4r-2} + 1)M_\rho^2$.

We next estimate $S_2$ by the Bernstein inequality. Let random variable $\xi(\tilde{z}) = \|\tilde{f}_\lambda(x) - \tilde{y}\|_2^2 - \|\tilde{f}_\rho(x) - \tilde{y}\|_2^2$, then $S_2 = \frac{1}{m}\sum_{i=1}^m \xi(\tilde{z}_i) - \mathbb{E}\xi(\tilde{z})$. Since $|\tilde{f}_\lambda^{(k)}(x)| \le |\tilde{f}_\rho^{(k)}(x)| \le 1$, and $\tilde{y}^{(k)} \in \{0,1\}$, we deduce that $|\xi(\tilde{z})| \le 4n$ and $\sigma^2(\xi) \le 16n(\tilde{\mathcal{E}}(\tilde{f}_\lambda) - \tilde{\mathcal{E}}(\tilde{f}_\rho)) \le 16n\tilde{D}(\lambda)$. From Bernstein inequality we have with confidence $1-\delta$

$$S_2 \le \Big(\frac{11n}{m} + \tilde{D}(\lambda)\Big)\log\frac{2}{\delta}.$$

Now we present a covering number-based upper bound for $S_1$. Denote

$$B_R^n = \{f \in \mathcal{H}_K^n : \|f^{(k)}\|_K^2 \le R^2, k = 1, \cdots, n\},$$

we notice that $\log \mathcal{N}(B_R^n, \eta, \|\cdot\|_\infty) \le c_s R^s \eta^{-s} n$ from Assumption 3. Here the metric is $\|f\|_\infty = \max_{k=1,\cdots,n} \|f^{(k)}\|_\infty$. For any $f \in B_R^n$, denote $\zeta(\tilde{z}) = \|f(x) - \tilde{y}\|_2^2 - \|\tilde{f}_\rho(x) - \tilde{y}\|_2^2$, then

$$\Big(\tilde{\mathcal{E}}(f) - \tilde{\mathcal{E}}(\tilde{f}_\rho)\Big) - \Big(\mathcal{E}_{\tilde{\mathbf{z}}}(f) - \mathcal{E}_{\tilde{\mathbf{z}}}(\tilde{f}_\rho)\Big) = \frac{1}{m}\sum_{i=1}^m \zeta(\tilde{z}_i) - \mathbb{E}(\zeta).$$

We can verify that $|\zeta(\tilde{z})| \le n(\kappa+1)^2R^2$, and $\sigma^2(\zeta) \le (\kappa+3)^2R^2 \cdot (\tilde{\mathcal{E}}(f) - \tilde{\mathcal{E}}(\tilde{f}_\rho))$. Therefore with confidence $1 - N\delta$

$$\Big(\tilde{\mathcal{E}}(f_j) - \tilde{\mathcal{E}}(\tilde{f}_\rho)\Big) - \Big(\mathcal{E}_{\tilde{\mathbf{z}}}(f_j) - \mathcal{E}_{\tilde{\mathbf{z}}}(\tilde{f}_\rho)\Big) \le \frac{3n(\kappa+3)^2R^2}{m}\log\frac{2}{\delta} + \frac{1}{2}(\tilde{\mathcal{E}}(f_j) - \tilde{\mathcal{E}}(\tilde{f}_\rho))$$

where $\{f_j\}_{j=1}^N$ forms a $\eta$-net of $B_R^n$ with $N = \mathcal{N}(B_R^n, \eta, \|\cdot\|_\infty)$. On the other hand,

$$\begin{aligned}
&\|\tilde{f}_{\mathbf{z}}(x) - \tilde{y}\|_2^2 - \|f_j(x) - \tilde{y}\|_2^2 \\
&= \sum_{k=1}^{n} [(\tilde{f}_{\mathbf{z}}^{(k)}(x) - f_j^{(k)}(x))(\tilde{f}_{\mathbf{z}}^{(k)}(x) + f_j^{(k)}(x) - 2\tilde{y}^{(k)})] \le 2\eta(\kappa+1)Rn.
\end{aligned}$$

Therefore with confidence $1 - N\delta$, there holds

$$\begin{aligned}
S_1 &= \Big(\tilde{\mathcal{E}}(\tilde{f}_{\mathbf{z}}) - \mathcal{E}_{\tilde{\mathbf{z}}}(\tilde{f}_\rho)\Big) - \Big(\mathcal{E}_{\tilde{\mathbf{z}}}(\tilde{f}_{\mathbf{z}}) - \mathcal{E}_{\tilde{\mathbf{z}}}(\tilde{f}_\rho)\Big) \\
&= \Big[\Big(\tilde{\mathcal{E}}(\tilde{f}_{\mathbf{z}}) - \mathcal{E}_{\tilde{\mathbf{z}}}(\tilde{f}_\rho)\Big) - \Big(\mathcal{E}_{\tilde{\mathbf{z}}}(\tilde{f}_{\mathbf{z}}) - \mathcal{E}_{\tilde{\mathbf{z}}}(\tilde{f}_\rho)\Big)\Big] \\
&\quad + \Big(\tilde{\mathcal{E}}(\tilde{f}_{\mathbf{z}}) - \tilde{\mathcal{E}}(f_j)\Big) + \Big(\mathcal{E}_{\tilde{\mathbf{z}}}(f_j) - \mathcal{E}_{\tilde{\mathbf{z}}}(\tilde{f}_{\mathbf{z}})\Big) \\
&\le \frac{3n(\kappa+3)^2R^2}{m}\log\frac{2}{\delta} + 5\eta n(\kappa+1)R + \frac{1}{2}(\tilde{\mathcal{E}}(\tilde{f}_{\mathbf{z}}) - \tilde{\mathcal{E}}(\tilde{f}_\rho)).
\end{aligned}$$

Scaling $N\delta$ to $\delta$ and choosing $\eta = Rm^{-\frac{1}{1+s}}$, we have with confidence $1-\delta$

$$S_1 \le C_2 \frac{nR^2}{m^{\frac{1}{1+s}}}\log\frac{2}{\delta} + \frac{1}{2}(\tilde{\mathcal{E}}(\tilde{f}_{\mathbf{z}}) - \tilde{\mathcal{E}}(\tilde{f}_\rho)),$$

where $C_2 = 6c_s(\kappa+3)^2 + 5(\kappa+1)$.

From the definition of $\tilde{f}_{\mathbf{z}}$ we have

$$\tilde{f}_{\mathbf{z}}^{(k)} = \arg\min_{f^{(k)}\in\mathcal{H}_K} \frac{1}{m}\sum_{i=1}^{m} |f^{(k)}(x_i) - \tilde{y}_i^{(k)}|^2 + \lambda\|f^{(k)}\|_K^2,$$

then $\lambda\|\tilde{f}_{\mathbf{z}}^{(k)}\|_K^2 \le \frac{1}{m}\sum_{i=1}^{m}\|\tilde{y}_i\|_2^2 \le n$. Hence $R^2 = \frac{1}{\lambda}$. Then with confidence $1-2\delta$, the whole excess generalization error can be bounded by

$$\tilde{\mathcal{E}}(\tilde{f}_{\mathbf{z}}) - \tilde{\mathcal{E}}(\tilde{f}_\rho) \le C_3 n \log\frac{2}{\delta}\Big(\frac{1}{m^{\frac{1}{1+s}}\lambda} + \lambda^{\min\{2r,1\}}\Big)$$

where $C_3 = 2(2C_1 + C_2 + 11)$. By taking $\lambda = m^{-\frac{1}{(1+s)\min\{2r+1,2\}}}$, we have with confidence $1-\delta$,

$$\tilde{\mathcal{E}}(\tilde{f}_{\mathbf{z}}) - \tilde{\mathcal{E}}(\tilde{f}_\rho) \le 2C_3 n \log\frac{4}{\delta}\cdot m^{-\min\{\frac{1}{2(1+s)}, \frac{2r}{(1+s)(2r+1)}\}}. \tag{3}$$

Hence we can deduce the misclassification error bound

$$R(S(\tilde{f}_{\mathbf{z}})) - R(f_c) \le 2\alpha\sqrt{C_3}\sqrt{n}\log^{\frac{1}{2}}\frac{4}{\delta}\cdot m^{-\min\{\frac{1}{4(1+s)}, \frac{r}{(1+s)(2r+1)}\}}$$

with confidence $1-\delta$. This proves the result. □

In this theorem, we establish that under certain mild conditions, solving the Kernel Ridge Regression (KRR) problem yields a classifier that converges to the Bayes classifier at the rate $\mathcal{O}_p(\sqrt{n}m^{-\frac{1}{4(1+s)}})$ if $r \ge \frac{1}{2}$ (i.e., $\tilde{f}_\rho^{(k)} \in \mathcal{H}_K$, we refer to Cucker and Smale, (2002) [3] for more details). Moreover, if kernel $K \in C^\infty$, then the capacity parameter $s$ tends to 0 from Zhou, (2002) [19], this rate can be close to $\mathcal{O}_p(\sqrt{n}m^{-\frac{1}{4}})$, which is optimal compared with the existing literature and matches the findings of Wang et al., (2023) [13].

## §5 Comparisons and discussions

In this section, we will show some comparisons with existing results and discussions on different loss functions.

Most of the existing results focus on the excess generalization error bound, so we will compare their results with (3). In the first paper on the CLL problem, Ishida et al., (2017) [8] proposed two complementary losses for the single CLL problem: one-versus-all (OVA) and pairwise-comparison (PC), based on the ordinary ones. Excess generalization error bound was given in the rate $O_p(n^2/sqrtm)$. Gao and Zhang, (2021) [7] developed another series of losses to directly model $\rho(\bar{y}|x)$, which has a learning rate of $O_p(n^2/\sqrt{m})$. In Liu et al., (2023) [10], an order-preserving loss was proposed. A comparison theorem was stated for one of such loss and the final excess generalization error bound was $O_p(n/\sqrt{m})$. Notice that all the above results are for the single CLL problem and on assumption is needed that the corresponding Rademacher complexities are bounded by $O_p(1/\sqrt{m})$.

Feng et al., (2020) [6] considered the MCLL problem. When the loss function is Lipschitz with respect to the first parameter and bounded, their excess generalization error bound was $O_p(n^2\sum_{j=1}^{n-1}\Pr\{s=j\}\frac{1}{\sqrt{m_j}})$. Here $s$ is the size of the multi-complementary label set, and $m_j$ is the size of the corresponding sample set. When $\Pr\{s=1\}=1$ the problem reduces to the single CLL problem and the best rate $O_p(n^2/\sqrt{m})$ can be achieved. But for true MCLL problem,

this rate may degenerate greatly. More recently, Wang et al., (2023) [13] constructed a novel unbiased risk estimator to deal with both single and multiple CLL problems. For each class $k$, the data set was separated into two parts: negative sample set $D_k^N$ and unlabeled sample set $D_k^U$, where $|D_k^N|+|D_k^U| = m$. Error bound can be written as $O_p(\sum_{k=1}^n(1/\sqrt{|D_k^N|}+1/\sqrt{|D_k^U|}))$. When the sizes of the two parts for each $k$ are all in the rate $O(1/m)$, the rate becomes the optimal one $O_p(n/\sqrt{m})$.

In all, for single CLL problem, the excess generalization error has been estimated well in different literature, $O_p(n/\sqrt{m})$ can be derived when Rademacher complexity bounds hold. Error estimated for MCLL are less considered. Optimal rate can also be obtained when the sample distributions satisfy some special conditions. However, in our result (3), $O_p(n/\sqrt{m})$ bound can be achieved if the kernel is carefully chosen.

In the previous sections, we conduct a comparison theorem and error analysis for the least squares loss. Results for other loss functions may be deduced by similar proofs. Zhang, (2004) [18] proved comparison theorems for some basic losses such as hinge loss, exponential loss, logistic loss. In Fan and Xiang, (2020) [5], a large series of losses called large-margin unified machines (LUM) losses were studied and excess misclassification error was estimated. Comparison theorem for LUM losses is also provided. Though the above results are on the binary class classification problems, similar ideas may be introduced to derive the corresponding comparison theorem for multi-class classification. In the error analysis, the regularization error bound may be more involved for alternative losses which depends on the kernel. We refer to Xiang and Zhou, (2009) [15] for a detail analysis of general convex loss and varying Gaussian kernels.

## §6 Conclusions

In this paper, we propose a straightforward and intuitive condition for the distribution of complementary labels. Under this condition, we show that the MCLL problem is equivalent to a traditional multi-class classification problem. In other words, the MCLL problem aims to find the largest element of the conditional expectation function $f_\rho(x)$, which is also the objective of multi-class classification. Moreover, we introduce a comparison condition between $\tilde{\rho}(\tilde{y}|x)$, the distribution of complementary labels, and $\rho(y|x)$, the distribution of true labels. This comparison condition allows us to establish an excess misclassification error bound based on the excess generalization error bound. To illustrate these concepts, we consider the Kernel Ridge Regression (KRR) method as an example. KRR is a popular machine learning algorithm that can be applied to both the MCLL problem and traditional multi-class classification. We demonstrate that under the proposed conditions, the KRR method achieves an optimal learning rate in terms of both the sample size $m$ and the number of classes $n$.

## Declarations

**Conflict of interest** The authors declare no conflict of interest.

## References

[1] G Bennett. *Probability inequalities for the sum of independent random variables*, Journal of the American Statistical Association, 1962, 57: 33-45.

[2] T Cour, B Sapp, B Taskar. *Learning from partial labels*, Journal of Machine Learning Research, 2011, 12: 1501-1536.

[3] F Cucker, S Smale. *On the mathematical foundations of learning*, Bulletin of the American Mathematical Society, 2002, 39: 1-49.

[4] F Cucker, D X Zhou. *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University Press, 2007.

[5] J Fan, D H Xiang. *Quantitative convergence analysis of kernel based large-margin unified machines*, Communications on Pure & Applied Analysis, 2020, 19(8): 4069-4083.

[6] L Feng, J Q Lv, B Han, G Niu, B An, M Sugiyama. *Learning with multiple complementary labels*, In Proceedings of the 37th International Conference on Machine Learning, 2020, 3072-3081.

[7] Y Gao, M L Zhang. *Discriminative complementary-label learning with weighted loss*, In Proceedings of the 38th International Conference on Machine Learning, 2021, 139: 3587-3597.

[8] T Ishida, G Niu, W Hu, M Sugiyama. *Learning from complementary labels*, Neural Information Processing Systems, 2017.

[9] T Ishida, G Niu, A K Menon, M Sugiyama. *Complementary -label learning for arbitrary losses and models*, In Proceedings of the 36th International Conference on Machine Learning, 2019, 97: 2971-2980.

[10] S Liu, Y Cao, Q Zhang, L Feng, B An. *Consistent complementary-label learning via order-preserving losses*, In Proceedings of the 26th International Conference on Artificial Intelligence and Statistics, 2023, 206: 8734-8748.

[11] C Wang, Z C Guo. *ERM learning algorithm for multi-class classification*, Applicable Analysis, 2012, 91(7): 1339-1349.

[12] C Wang, X Xu, D Liu, X Niu, S Han. *Simple and effective complementary label learning based on mean square error loss*, Journal of Physics: Conference Series, 2023, https://doi.org/10.1088/1742-6596/2504/1/012016.

[13] W Wang, T Ishida, Y J Zhang, G Niu, M Sugiyama. *Learning with complementary labels revisited: A consistent approach via negative-unlabeled learning*, Arxiv, 2023.

[14] Q Wu, Y Ying, D X Zhou. *Learning rates of least-square regularized regression*, Foundations of Computational Mathematics, 2006, 6: 171-192.

[15] D H Xiang, D X Zhou. *Classification with gaussians and convex loss*, Journal of Machine Learning Research, 2009, 10: 1447-1468.

[16] Y Xu, M Gong, J Chen, T Liu, K Zhang, K Batmanghelich. *Generative-discriminative complementary learning*, In Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 6526-6533.

[17] X Yu, T Liu, M Gong, D Tao. *Learning with biased complementary labels*, In Computer Vision-ECCV 2018, Lecture Notes in Computer Science, Springer, Cham, 2018, 11205: 68-85.

[18] T Zhang. *Statistical behavior and consistency of classification methods based on convex risk minimization*, Annals of Statistics, 2004, 32(1): 56-134.

[19] D X Zhou. *The covering number in learning theory*, Journal of Complexity, 2002, 18: 739-767.

[1]School of Mathematics and Statistics, Huizhou University, Huizhou 516007, China.
Emails: niewl@hzu.edu.cn, wangch@hzu.edu.cn

[2]School of Computer Science and Engineering, Huizhou University, Huizhou 516007, China.
Email: eezhxie@gmail.com