Classification with non-i.i.d. sampling
Zheng-Chu Guo\textsuperscript{a,b}, Lei Shi\textsuperscript{b,*}

\textsuperscript{a} School of Mathematics and Computational Science, Sun Yat-sen University, Guangzhou 510275, PR China
\textsuperscript{b} Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China

\begin{abstract}
We study learning algorithms for classification generated by regularization schemes in reproducing kernel Hilbert spaces associated with a general convex loss function in a non-i.i.d. process. Error analysis is studied and our main purpose is to provide an elaborate capacity dependent error bounds by applying concentration techniques involving the $\ell^2$-empirical covering numbers.
\end{abstract}

\section{Introduction}

In this paper, we consider learning algorithms for classification with non-i.i.d. sampling processes.

In a binary classification problem, the input space is a compact subset $X \subset \mathbb{R}^d$ and the outputs space $Y = \{-1, 1\}$ represents two classes. Classification algorithms produce binary classifiers $C : X \to Y$. Let $\rho$ be a probability measure defined on $Z := X \times Y$. The prediction ability of a classifier $C$ is measured by the misclassification error which is defined as

$$R(C) = \text{Prob}_{(x,y) \in (Z,\rho)}[C(x) \neq y] = \int_X \rho_x(y \neq C(x)) d \rho_X. \quad (1.1)$$

Here $\rho_X$ is the marginal distribution of $\rho$ on $X$ and $\rho_x$ is the conditional distribution at $x \in X$. The best classifier that minimizes the misclassification error is the \textit{Bayes rule} given by

$$f_c(x) = \begin{cases} 1, & \text{if } \rho_x(y = 1) \geq \rho_x(y = -1), \\ -1, & \text{if } \rho_x(y = 1) < \rho_x(y = -1). \end{cases} \quad (1.2)$$

Since $\rho_x$ is unknown, $f_c$ cannot be computed directly. The goal of classification algorithms is to find classifiers which approximate $f_c$ from a finite sample $z = \{z_i = (x_i, y_i)\}_{i=1}^m \in Z^m$. The classifiers considered here are induced by real-valued functions $f : X \to \mathbb{R}$ as $C = \text{sgn}(f)$ which is defined by $\text{sgn}(f)(x) = 1$ if $f(x) \geq 0$ and $\text{sgn}(f)(x) = -1$ otherwise. We define a loss function $\phi : \mathbb{R} \to \mathbb{R}_+$ and use the error $\phi(y f(x))$ to measure the difference between the output $y$ and the prediction $\text{sgn}(f(x))$.

\textit{Definition 1.} A function $\phi : \mathbb{R} \to \mathbb{R}_+$ is called a classifying loss function if it is convex, differentiable at 0 with $\phi'(0) < 0$, and the smallest zero of $\phi$ is 1.