

# Minimax optimal rates of convergence for multicategory classifications <sup>\*</sup>

Di-Rong Chen    Xu You  
Department of Applied Mathematics  
Beijing Univ. of Aeronautics and Astronautics  
Beijing 100083, P. R. China

January 17, 2005

## Abstract

In the problem of classification (or pattern recognition), given a set of  $n$  samples, we attempt to construct a classifier  $g_n$  with small misclassification error. It is important to study the convergence rates of the misclassification error as  $n$  to infinity. It is known that such a rate can't exist for the set of *all* distributions. In this paper we obtain the optimal convergence rates for a class of distributions  $\mathcal{D}^{(\lambda, \omega)}$  in multicategory classification and nonstandard binary classification.

**Keywords :** rate of convergence, error probability, modulus of continuity, multicategory classification.

**2000 MR Subject Classification:** 62C20, 62H30, 62G20, 41A46

## 1 Introduction

Pattern recognition (or classification) is about guessing or predicting the unknown class of an observation. An *observation* is often a collection of numerical measurements represented by a  $d$ -dimensional vector  $x$ . The unknown nature of the observation is called a *class*. It is denoted by  $y$  and takes values in the set  $\{0, 1, \dots, m-1\}$ , where  $m \geq 2$  is an integer. To model the learning problem, we introduce a probability setting, and let  $Z = (X, Y)$  be an  $\mathbb{R}^d \times \{0, 1, \dots, m-1\}$ -valued random pair. Any

---

<sup>\*</sup>Research supported in part by NSF of China under grant 10171007. The work was partially done while the first author was visiting the Institute for Mathematical Sciences, National University of Singapore in 2003. The visit was supported by the Institute

measurable function  $g : \mathbb{R}^d \rightarrow \{0, 1, \dots, m-1\}$  defines a classifier on a classification rule. The probability of error for a classifier  $g$  is

$$L(g) = P\{g(X) \neq Y\}.$$

Denote by  $\mathcal{G}$  the set of all classifiers and  $L^* = \inf_{g \in \mathcal{G}} L(g)$ .

In the model of learning, we are given  $n$  independent and identically samples  $D_n = \{Z_i\}_{i=1}^n$ ,  $Z_i = (X_i, Y_i)$ ,  $1 \leq i \leq n$ , from the distribution on  $Z$ . An estimate  $M_n$  is a mapping with input  $D_n$  and output  $g_n \in \mathcal{G}$ . A sequence  $\{M_n\}$  of estimates is said to be *universally consistent* if, for any distribution on  $Z$ ,

$$\lim_{n \rightarrow \infty} EL(g_n) - L^* = 0,$$

where  $EL(g_n)$  is the expectation of  $L(g_n)$  taken with respect to  $D_n$ . The goal of classification is to construct consistent estimates. There are vast papers to deal with the case  $m = 2$  (see [1], [2] and references therein). Support vector machines, Boosting have been proved much efficient in application. For  $m > 2$ , some difficulties arise. There are also many papers focusing on the multicategory classification. For example, multicategory Support vector machines with hing loss is constructed in [3]. Recently, the universal consistency of the multicategory Support vector machines are established in [4] by developing the methods in [5] and [6].

It is more desired in both theory and application, for universally consistent estimates, to know how fast the corresponding error  $L(g_n)$  converges to  $L^*$  in certain sense. In this paper we are interested in the the convergence rate of  $EL(g_n) - L^*$  as  $n \rightarrow \infty$ . Disappointingly, no estimate exists to guarantee a specified rate for *all* distributions. Therefore we have to restrict ourself with a class of distributions.

Before proceeding further, we introduce the notions concerning with the rate of convergence.

For a class  $\mathcal{D}$  of distributions of  $(X, Y)$  and a sequence  $\{a_n\}$  of positive numbers, let

$$r_{mm}(\mathcal{D}, \{a_n\}) = \limsup_{n \rightarrow \infty} \inf M_n \sup_{(X, Y) \in \mathcal{D}} \frac{EL(g_n) - L^*}{a_n},$$

where the supremum is taken over all distributions in  $\mathcal{D}$  and infimum is taken over all sequences  $\{M_n\}$  of estimates.

**Definition 1.1** Suppose that  $\mathcal{D}$  is a class of distributions of  $Z$  and  $\{a_n\}$  is a sequence of positive numbers. Let  $r_{mm}(\mathcal{D}, \{a_n\})$  be defined as above. We call  $\{a_n\}$  a minimax lower rate of convergence if  $r_{mm}(\mathcal{D}, \{a_n\}) > 0$ .

We call  $\{a_n\}$  a minimax upper rate of convergence for the class  $\mathcal{D}$ , if for some  $\{M_n\}$

$$\limsup_{n \rightarrow \infty} \sup_{(X, Y) \in \mathcal{D}} \frac{EL(g_n) - L^*}{a_n} < \infty.$$

A sequence  $\{a_n\}$  is a minimax optimal rate of convergence for the class  $\mathcal{D}$  if it is both a minimax upper and a lower rate of convergence for the class  $\mathcal{D}$ .

For binary classification, i.e.,  $m = 2$ , and regression problem, the optimal convergence rates are determined. See, for example, [7], [8], [9] and [10]. In these papers, the distributions of  $(X, Y)$  are required to be smooth in some sense. More precisely, in regression problem,  $Y$  is a smooth function of  $X$ , up to a standard normal noise; in binary classification, the conditional probability  $P\{Y = 1|X\}$  is itself a smooth function of  $X$ . The functions have all  $\alpha$ -derivatives,  $\alpha = (\alpha_1, \dots, \alpha_d)$ ,  $\sum_{i=1}^d \alpha_i = \lambda$ , and the derivatives have a Hölder exponent  $\beta > 0$ , where  $\lambda$  and  $\beta$  are constants. Under the above condition and an assumption on the distribution of  $X$ , the optimal rate of convergence is  $\{n^{-\frac{\beta}{2\beta+d}}\}$  for binary classification.

The purpose of this paper is to generalize the above mentioned results. We work in a general setting. The Hölder condition for the derivatives is replaced with a concave modulus of continuity. More important, we consider the multicategory case. The optimal rates for the multicategory classification are determined. Moreover, we show the results also hold for binary classification with nonstandard loss. It worth to note that the function classes defined by a concave modulus of continuity play an important role in the theory of functions, in particular in approximation theory of functions.

For our purpose, we generalize a well known inequality, relating classification error with approximation error, to multicategory and the nonstandard binary cases (see Theorem 2.2 and Theorem 3.2). It is of independent interest.

## 2 Multicategory classification

In this section we establish the optimal convergence rate for multicategory classification.

Let  $\eta_i(x) = P\{Y = i|X = x\}$ ,  $i = 0, 1, \dots, m - 1$ , be the conditional probability given  $X = x$ . The Bayes classifier  $g^*$  is given by

$$g^*(x) = \arg \max_{i=0,1,\dots,m-1} \eta_i(x), \quad x \in \mathbb{R}^d.$$

It is the minimizer of  $L(g)$ , i.e.,  $L(g^*) = L^*$ . In fact, it is not difficult to prove the following conclusion.

**Lemma 2.1** For any classifier  $g_n(x)$ , it holds

$$L(g_n) - L(g^*) = \int_{\mathbb{R}^d} (\eta_{g^*(x)}(x) - \eta_{g_n(x)}(x)) d\rho_X,$$

where  $\rho_X$  is the marginal distribution of  $Z$  on  $\mathbb{R}^d$ . ■

We first derive the upper rate of convergence. Let  $\mathcal{F}$  be a set of functions. Suppose that, based on  $D_n$ , a function  $\eta^{(n)}$  is constructed to estimate a function  $\eta \in \mathcal{F}$  in  $L_2(X, d\rho_X)$ . Such a construction is said to be a consistent approximation on  $\mathcal{F}$  if

$$\lim_{n \rightarrow \infty} E \|\eta - \eta^{(n)}\|_2 = 0, \quad \forall \eta \in \mathcal{F}.$$

It is called a strongly consistent approximation if, with probability one

$$\lim_{n \rightarrow \infty} \|\eta - \eta^{(n)}\|_2 = 0, \quad \forall \eta \in \mathcal{F}.$$

In learning theory, the goal of regression to approximate  $\eta_i, i = 0, 1, \dots, m - 1$ . Recent advances in this subject are referred to [11]. Once the approximants  $\eta_i^{(n)}, 0 \leq i \leq m - 1$ , are obtained, we define the plug-in classifier as following

$$g_n(x) = \arg \max_{i=0,1,\dots,m-1} \eta_i^{(n)}(x).$$

The following result establishes an inequality between the error for approximation of functions and the error for the plug-in classifier. It generalizes a well known result for  $m = 1$  (see [1]). However, the existing proof for such results does not work.

**Theorem 2.2** Let  $g_n(x)$  be the plug-in classifier determined by the approximants  $\eta_i^{(n)}$  of  $\eta_i, 0 \leq i \leq m - 1$ . Then

$$L(g_n) - L(g^*) \leq \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_2,$$

where  $\|f\|_p = (\int_{\mathbb{R}^d} |f|^p d\rho_X)^{\frac{1}{p}}$

**Proof.** By definition of  $\eta_{g_n}^{(n)}$  we have  $\eta_{g_n}^{(n)}(x) \geq \eta_{g^*}^{(n)}(x)$ . Therefore

$$\begin{aligned} 0 &\leq \eta_{g^*}(x) - \eta_{g_n}(x) \\ &= \eta_{g^*}(x) - \eta_{g^*}^{(n)}(x) + \eta_{g^*}^{(n)}(x) - \eta_{g_n}^{(n)}(x) + \eta_{g_n}^{(n)}(x) - \eta_{g_n}(x) \\ &\leq \eta_{g^*}(x) - \eta_{g^*}^{(n)}(x) + \eta_{g_n}^{(n)}(x) - \eta_{g_n}(x) \\ &\leq |\eta_{g^*}(x) - \eta_{g^*}^{(n)}(x)| + |\eta_{g_n}^{(n)}(x) - \eta_{g_n}(x)|. \end{aligned}$$

Consequently,

$$0 \leq \eta_{g^*}(x) - \eta_{g_n}(x) \leq \sum_{i=0}^{m-1} |\eta_i(x) - \eta_i^{(n)}(x)|. \quad (2.1)$$

It follows from Lemma 2.1 that

$$L(g_n) - L(g^*) \leq \sum_{i=0}^{m-1} \int_{\mathbb{R}^d} |\eta_i(x) - \eta_i^{(n)}(x)| d\rho_X.$$

The proof is complete by Cauchy-Schwartz inequality. ■

As a consequence of Theorem 2.2, we know that any consistent approximation on a set  $\mathcal{F}$  of functions induces a consistent classification, provided that the conditional probabilities  $\eta_i \in \mathcal{F}, i = 0, 1, \dots, m - 1$ . Hence, there indeed exist universally consistent classifiers in multicategory classification.

With Theorem 2.2 we can reveal a fact that  $L(g_n) - L(g^*)$  convergence to zero faster than the  $L_2$ -error of the corresponding approximation method.

**Corollary 2.3** Suppose that  $g_n$  is the plug-in classifier constructed from a consistent approximation for  $\eta_i, i = 0, 1, \dots, m-1$ . Then

$$\lim_{n \rightarrow \infty} \frac{EL(g_n) - L(g^*)}{\sqrt{E \left\{ \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_2 \right\}^2}} = 0,$$

where the expectation  $E$  is taken with respect to the training data  $D_n$ .

**Proof.** From Lemma 2.1, we have

$$L(g_n) - L(g^*) = \int_{\eta_{g^*} \neq \eta_{g_n}} (\eta_{g^*(x)}(x) - \eta_{g_n(x)}(x)) d\rho_X.$$

Let us bound the right hand side as following. For arbitrary  $\epsilon > 0$ , there exists  $\delta > 0$ , subject to

$$\sum_{i \neq j} P\{0 < \eta_i - \eta_j < \delta\} < \epsilon^2.$$

It together with (2.1) yields

$$\begin{aligned} \int_{0 < \eta_{g^*} - \eta_{g_n} < \delta} (\eta_{g^*} - \eta_{g_n}) d\rho_X &\leq \sum_{i=0}^{m-1} \int_{0 < \eta_{g^*} - \eta_{g_n} < \delta} |\eta_i - \eta_i^{(n)}| d\rho_X \\ &\leq \epsilon \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_2. \end{aligned}$$

Similarly it holds

$$\int_{\eta_{g^*} - \eta_{g_n} \geq \delta} (\eta_{g^*} - \eta_{g_n}) d\rho_X \leq \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_2 \sqrt{P\{\eta_{g^*} - \eta_{g_n} \geq \delta\}}.$$

But by (2.1) we have

$$P\{\eta_{g^*} - \eta_{g_n} \geq \delta\} \leq P\left\{ \sum_{i=0}^{m-1} |\eta_i - \eta_i^{(n)}| \geq \delta \right\} \leq \frac{1}{\delta} \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_1.$$

Therefore,

$$L(g_n) - L(g^*) \leq \epsilon \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_2 + \frac{1}{\delta} \left( \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_2 \right) \left( \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_1 \right). \quad (2.2)$$

By Cauchy-Schwartz inequality we have

$$E \left\{ \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_2 \right\} \leq \sqrt{E \left\{ \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_2 \right\}^2}$$

and

$$\begin{aligned} & E \left\{ \left( \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_2 \right) \left( \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_1 \right) \right\} \\ & \leq \sqrt{E \left\{ \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_2 \right\}^2} \sqrt{E \left\{ \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_1 \right\}^2}. \end{aligned}$$

Consequently,

$$\frac{EL(g_n) - L(g^*)}{\sqrt{E \left\{ \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_2 \right\}^2}} \leq \epsilon + \frac{1}{\delta} \sqrt{E \left\{ \sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_1 \right\}^2},$$

which completes the proof by  $\|f\|_1 \leq \|f\|_2$ . ■

For the strongly consistent approximation we have the following result. For its proof, it only needs to make use of inequality (2.2). The details are omitted.

**Corollary 2.4** Suppose that  $g_n$  is the plug-in classifier constructed from a strongly consistent approximation for  $\eta_i, i = 0, 1, \dots, m-1$ . Then with probability one

$$\lim_{n \rightarrow \infty} \frac{L(g_n) - L(g^*)}{\sum_{i=0}^{m-1} \|\eta_i - \eta_i^{(n)}\|_2} = 0.$$

To introduce the class of distributions discussed in this paper, we first define a set of smooth functions. ■

**Definition 2.1** Let  $\omega(x)$  be a concave modulus of continuity. For given  $\lambda \in \mathbb{N}$ , let  $\mathcal{F}^{(\lambda, \omega)}$  be the set of functions  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  such that for every  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d)$  with  $\alpha_i \in \mathbb{N}, \sum_{i=1}^d \alpha_i = \lambda$ ,

$$|D^\alpha f(x) - D^\alpha f(z)| \leq \omega(\|x - z\|),$$

where  $D^\alpha$  denotes the partial derivation.

**Definition 2.2** Let  $\mathcal{F}^{(\lambda, \omega)}$  be defined as above. We denote by  $\mathcal{D}^{(\lambda, \omega)}$  the class of distributions  $\rho$  of  $(X, Y)$  satisfying the following conditions.

- i)  $X$  is uniformly distributed on  $[0, 1]^d$ , that is to say  $d\rho_X = dx$ ;
- ii)  $Y \in \{0, \dots, m-1\}$  a.s. ;
- iii)  $\eta_i(x) \in \mathcal{F}^{(\lambda, \omega)}$ ,  $i = 0, 1, \dots, m-1$ , where  $\eta_i(x) = P\{Y = i | X = x\}$ .

By assumption, the function  $h(t) = t^{2\lambda+d}\omega^2(t)$  is monotone increasing. Therefore, there exists, for any  $n$ , a unique solution  $q_n$  to equation

$$t^{2\lambda+d}\omega^2(t) = \frac{1}{n}.$$

Let  $\epsilon_n = q_n^\lambda \omega(q_n)$ . Then  $q_n^{-d} = n\epsilon_n^2$ . It is known from [12] that the metric entropy of  $\mathcal{F}^{(\lambda, \omega)}$  in  $L_2(X, \rho_X)$  is of order  $q_n^{-d}$ .

It is also known from [9,10] that, for any  $D_n$ , we can construct a function  $\eta_i^{(n)}(x)$  subject to

$$\limsup_{n \rightarrow \infty} \sup_{\eta \in \mathcal{F}^{(\lambda, \omega)}} \frac{E\|\eta - \eta^{(n)}\|_2}{q_n^\lambda \omega(q_n)} < +\infty. \quad (2.3)$$

Applying (2.3) to  $\eta = \eta_i$ ,  $i = 0, 1, \dots, m-1$ , respectively, and appealing to Theorem 2.2, we have the upper rate of convergence for multicategory classification.

**Theorem 2.5** For multicategory classification, the sequence  $\{q_n^\lambda \omega(q_n)\}$  is a min-max upper rate of convergence for the class  $\mathcal{D}^{(\lambda, \omega)}$ . ■

Now let us consider the lower rate of convergence. To this end, we associate an arbitrary function  $\eta(x) \in \mathcal{F}^{(\lambda, \omega)}$ ,  $0 \leq \eta(x) \leq 1$ , with a distribution  $\rho \in \mathcal{D}^{(\lambda, \omega)}$  on  $Z$  by setting

$$\eta_0(x) = \eta(x), \quad \eta_i(x) = \frac{1 - \eta(x)}{m-1}, \quad i = 1, \dots, m-1. \quad (2.4)$$

**Theorem 2.6** Let  $\eta(x) \in \mathcal{F}^{(\lambda, \omega)}$  with  $0 \leq \eta(x) \leq 1$  and  $\rho$  be the distribution on  $Z$  defined as above. Then, for any classifier  $g(x)$

$$L(g) - L(g^*) \geq \frac{m}{m-1} \int_X |\eta(x) - \frac{1}{m}| (I_{\{g=0, g^*=1\}}(x) + I_{\{g=1, g^*=0\}}(x)) dx.$$

**Proof.** The Bayes rule is not unique. We choose, without loss of any generality, the Bayes rule  $g^*$  by

$$g^*(x) = \begin{cases} 0, & \text{if } \frac{1}{m} \leq \eta(x); \\ 1, & \text{otherwise.} \end{cases}$$

Suppose that  $x$  satisfies  $I_{\{g=0, g^*=1\}}(x) + I_{\{g=1, g^*=0\}}(x) = 1$ . We claim that

$$\eta_{g^*(x)}(x) - \eta_{g(x)}(x) = \frac{m}{m-1} \left| \eta(x) - \frac{1}{m} \right|.$$

In fact, if  $g^*(x) = 1$ , then  $g(x) = 0$  and  $\eta(x) < 1/m$ . Hence

$$\eta_{g^*(x)}(x) - \eta_{g(x)}(x) = \frac{m}{m-1} \left( \frac{1}{m} - \eta(x) \right).$$

Similarly, for  $g(x) = 0$  and  $g^*(x) = 1$  we have

$$\eta_{g^*(x)}(x) - \eta_{g(x)}(x) = \frac{m}{m-1} \left( \eta(x) - \frac{1}{m} \right).$$

This verifies (2.4). On the other hand, it follows from Lemma 2.1 that

$$L(g) - L(g^*) = \int_{g \neq g^*} (\eta_{g^*}(x) - \eta_g(x)) dx$$

The proof is complete by (2.4). ■

**Theorem 2.7** For multicategory classification, the sequence  $\{q_n^\lambda \omega(q_n)\}$  is a min-max lower rate of convergence for the class  $\mathcal{D}^{(\lambda, \omega)}$ .

**Proof.** We prove the result by the method of [7]. First we define a subclass of distributions of  $(X, Y)$  contained in  $\mathcal{D}^{(\lambda, \omega)}$ . We pack infinitely many disjoint cubes into  $[0, 1]^d$  in the following way. For a given probability distribution  $\{p_j\}$ . Let  $\{B_j\}$  be a partition of  $[0, 1]$  such that  $B_j$  is an interval of length  $p_j$ , we pack disjoint cubes of volume  $p_j^d$  into the rectangle

$$B_j \times [0, 1]^{d-1}.$$

Denote these cubes by

$$A_{j,1}, \dots, A_{j,S_j},$$

where

$$S_j = \lfloor \frac{1}{p_j} \rfloor^{d-1}.$$

Moreover, denote by  $\mathcal{C}$  the set of all vectors

$$c = (c_{1,1}, \dots, c_{1,S_1}; c_{2,1}, \dots, c_{2,S_2}; \dots), \quad c_{j,k} \in \{-1, 1\}.$$

Let  $a_{j,k}$  be the center of  $A_{j,k}$ , choose a function  $f : \mathbb{R}^d \rightarrow [0, \frac{1}{4}]$  such that

- i) the support of  $f$  is a subset of  $[-\frac{1}{2}, \frac{1}{2}]^d$ ;
- ii)  $\int f(x) dx > 0$ ;
- iii)  $f(x) \in \mathcal{F}^{(\lambda, \omega)}$ .

For  $c \in \mathcal{C}$  we define the function

$$\eta^{(c)}(x) = \frac{1}{m} + \sum_{j=1}^{\infty} \sum_{k=1}^{S_j} c_{j,k} f_{j,k}(x),$$

where

$$f_{j,k}(x) = p_j^k f\left(\frac{x - a_{j,k}}{p_j}\right) \omega(p_j).$$

It is easily seen

$$\eta^{(c)}(x) \in \mathcal{F}^{(\lambda, \omega)}.$$

Let  $\rho \in \mathcal{D}^{(\lambda, \omega)}$  by setting  $\eta_i, i = 0, 1, \dots, m-1$ , as in (2.4) with  $\eta = \eta^{(c)}$ . Furthermore, let  $\mu$  be the measure defined by  $\mu(A) = \int_A |\eta^{(c)}(x) - \frac{1}{m}| dx$ . Clearly,

$$\mu(A) = \int_A \sum_{j,k} f_{j,k} dx.$$

For any  $x$ , there is a pair  $\{j, k\}$  such that  $x \in A_{j,k}$ . If  $c_{j,k} = 1$ , we have  $\eta_0(x) = \eta^{(c)}(x) \geq 1/m$  and therefore  $g^*(x) = 0$ . For  $c_{j,k} = -1$ ,  $\eta_0(x) = \eta^{(c)}(x) < 1/m$ . We let  $g^*(x) = 1$ . This means  $g^*(x) = \frac{1}{2} + \sum_{j,k} \frac{c_{j,k} I_{A_{j,k}}}{2}$ . Thus for any classifier  $g_n$ ,

$$I_{\{g_n \neq g^*\}} = I_{\{g_n=0, g^*=1\}} + I_{\{g_n=1, g^*=0\}}.$$

It follows from Theorem 2.5 and the last equality that

$$\begin{aligned} L(g_n) - L(g^*) &\geq \frac{1}{(k-1)^2 m(m-1)} \int_X (g_n - g^*)^2 (I_{\{g_n=0, g^*=1\}} + I_{\{g_n=1, g^*=0\}}) d\mu \\ &= \frac{1}{(k-1)^2 m(m-1)} \int_X ((g_n - \frac{1}{2}) - (g^* - \frac{1}{2}))^2 I_{\{g_n \neq g^*\}} d\mu \\ &\geq \frac{1}{(k-1)^2 m(m-1)} \int_X ((\hat{g}_n - \frac{1}{2}) - (g^* - \frac{1}{2}))^2 I_{\{g_n \neq g^*\}} d\mu, \end{aligned}$$

where  $\hat{g}$  denotes the projection of  $g$  on the orthogonal system  $\{I_{A_{j,k}}/2\}$  in  $L_2(X, d\mu)$ , and the last inequality holds due to  $\hat{g}^* = g^*$ . Consequently,

$$\begin{aligned} L(g_n) - L(g^*) &\geq \frac{1}{(k-1)^2 m(m-1)} \int_X \sum_{j,k} f_{j,k} (\hat{g}_n - g^*)^2 dx \\ &= \frac{1}{(k-1)^2 m(m-1)} \sum_{j,k} \frac{(\hat{c}_{n,j,k} - c_{j,k})^2}{4} \int_{A_{j,k}} f_{j,k} dx, \end{aligned}$$

where the constants  $\hat{c}_{n,j,k}$  are determined by the expression of the projection of  $g_n - 1/2$

$$\hat{g}_n - \frac{1}{2} = \sum_{j,k} \frac{\hat{c}_{n,j,k} I_{A_{j,k}}}{2}.$$

Therefore we obtain by the definition of  $f_{j,k}$

$$L(g_n) - L(g^*) \geq \frac{1}{4(k-1)^2 m(m-1)} \|f\|_1 \sum_{j,k} (\hat{c}_{n,j,k} - c_{j,k})^2 p_j^{\lambda+d} \omega(p_j).$$

Let

$$\tilde{c}_{n,j,k} = \begin{cases} 1 & \text{if } \hat{c}_{n,j,k} \geq 0 \\ -1 & \text{otherwise.} \end{cases}$$

Because of

$$|\hat{c}_{n,j,k} - c_{j,k}|^2 \geq \frac{|\tilde{c}_{n,j,k} - c_{j,k}|}{2} = I_{\{\tilde{c}_{n,j,k} \neq c_{j,k}\}},$$

we get

$$L(g_n) - L(g^*) \geq \frac{1}{4(k-1)^2 m(m-1)} \|f\|_1 \sum_{j,k} I_{\{\tilde{c}_{n,j,k} \neq c_{j,k}\}} p_j^{\lambda+d} \omega(p_j).$$

This proves

$$EL(g_n) - L(g^*) \geq \frac{1}{4(k-1)^2 m(m-1)} \|f\|_1 R_n(c), \quad (2.5)$$

where

$$R_n(c) = \sum_{j: np_j^{2k+d}\omega(p_j) \leq 1} \sum_{k=1}^{S_j} p_j^{\lambda+d} \omega(p_j) P\{\tilde{c}_{n,j,k} \neq c_{j,k}\}.$$

By the same arguments as in [7] and a selection for  $\{p_j\} : p_j = q_n, j \leq \frac{1}{q_n}$ , we can deduce for some constant  $K$

$$\sup_{c \in \mathcal{C}} R_n(c) \geq K \left( \frac{1}{qn} - 1 \right) q_n^{\lambda+1} \omega(q_n) = K q_n^\lambda \omega(q_n) (1 - o(1)), \quad n \rightarrow \infty.$$

The details are omitted and referred to [7]. On the other hand, it follows from (2.5) for any sequence  $\{a_n\}$  of positive numbers

$$\limsup_{n \rightarrow \infty} \inf_{g_n} \sup_{(X,Y) \in \mathcal{D}(\lambda,\omega)} \frac{EL(g_n) - L(g^*)}{a_n} \geq K' \limsup_{n \rightarrow \infty} \inf_{g_n} \sup_{c \in \mathcal{C}} \frac{R_n(c)}{a_n},$$

where  $K' = \frac{\|f\|_1}{4(k-1)^2 m(m-1)}$ . Hence

$$\limsup_{n \rightarrow \infty} \inf_{g_n} \sup_{(X,Y) \in \mathcal{D}(\lambda,\omega)} \frac{EL(g_n) - L(g^*)}{q_n^\lambda \omega(q_n)} > 0.$$

The proof is complete. ■

The optimal rate has been established by Theorem 2.5 and Theorem 2.7.

**Theorem 2.8** For multicategory classification, the sequence  $\{q_n^\lambda \omega(q_n)\}$  is the minimax optimal rate of convergence for the class  $\mathcal{D}(\lambda,\omega)$ .

### 3 Classification with nonstandard cost

In this section we briefly discuss the classification with nonstandard cost. It turns out the minimax optimal rate is the same as in Section 2.

For simplicity, we only consider the two class case, i.e.,  $m = 2$ . Direct extensions to cases with multiple classes are straightforward. Let  $Z = (X, Y)$  be a pair of random variables taking their values from  $\mathbb{R}^d \times \{0, 1\}$ .

For  $0 < \alpha < 2$ , define the loss function as

$$l_\alpha(y, g(x)) = \alpha I_{\{y=0, g(x)=1\}} + (2 - \alpha) I_{\{y=1, g(x)=0\}}. \quad (3.1)$$

In particular, when  $\alpha = 1$ , it is the usual case we consider in pattern recognition.

For arbitrary decision function  $g$ , the error of the decision function  $g$  is

$$L^{(\alpha)}(g) = \int_{\mathbb{R}^d \times \{0,1\}} l_\alpha(y, g(x)) d\rho = \alpha P\{Y = 0, g = 1\} + (2 - \alpha) P\{Y = 1, g = 0\}.$$

Then the classifier  $g^*$ , also known as Bayes decision, given by

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) \geq \frac{\alpha}{2} \\ 0 & \text{otherwise,} \end{cases}$$

is a minimizer of  $L^{(\alpha)}(g)$ . In fact we have

**Theorem 3.1** For any decision function  $g$  it holds the equality

$$L^{(\alpha)}(g) - L^{(\alpha)}(g^*) = 2 \int_{\mathbb{R}^d} \left| \eta - \frac{\alpha}{2} \right| I_{\{g^*(x) \neq g(x)\}} dx. \quad (3.2)$$

**Proof.** For any  $x$ , it follows from (3.1) that

$$l^{(\alpha)}(y, g^*(x)) = \min\{\alpha(1 - \eta), (2 - \alpha)\eta\}.$$

Moreover, for any decision  $g$  with  $g(x) \neq g^*(x)$ , it holds

$$l^{(\alpha)}(y, g(x)) = \max\{\alpha(1 - \eta), (2 - \alpha)\eta\}.$$

The above two equalities give (3.2). The proof is complete. ■

First we derive the minimax upper rate of convergence. As in Section 2, we construct a plug-in classifier  $g_n$  from an approximant  $\eta^{(n)}$  of  $\eta$  as following.

$$g_n(x) = \begin{cases} 1 & \text{if } \eta^{(n)}(x) \geq \frac{\alpha}{2} \\ 0 & \text{otherwise} \end{cases}$$

**Theorem 3.2** For any plug-in classifier  $g_n$  constructed from an approximant  $\eta^{(n)}$  of  $\eta$  we have

$$L^{(\alpha)}(g_n) - L^{(\alpha)}(g^*) \leq 2 \sqrt{\int_{\mathbb{R}^d} |\eta(x) - \eta^{(n)}(x)|^2 dx}.$$

**Proof.** By Theorem 3.1, we have

$$L^{(\alpha)}(g_n) - L^{(\alpha)}(g^*) = 2 \int_{\mathbb{R}^d} |\eta(x) - \frac{\alpha}{2}| I_{\{g^*(x) \neq g_n(x)\}} dx.$$

Then by the construction of  $g_n$

$$\begin{aligned} & L^{(\alpha)}(g_n) - L^{(\alpha)}(g^*) \\ = & 2 \int_{\eta(x) \geq \frac{\alpha}{2}, \eta^{(n)}(x) < \frac{\alpha}{2}} |\eta(x) - \frac{\alpha}{2}| dx + 2 \int_{\eta(x) < \frac{\alpha}{2}, \eta^{(n)}(x) \geq \frac{\alpha}{2}} |\eta(x) - \frac{\alpha}{2}| dx \\ \leq & 2 \int_{\mathbb{R}^d} |\eta(x) - \eta^{(n)}(x)| dx. \end{aligned}$$

It together with Cauchy-Schwartz inequality yields the conclusion. The proof of the theorem is complete.  $\blacksquare$

With Theorem 3.2, a minimax upper rate of convergence for estimating regression function  $\eta(x)$  immediately gives an upper rate of convergence for classification. Recall that  $q_n$  is defined in Section 2. Appealing to (2.3) again we conclude the following.

**Theorem 3.3** For the classification with nonstandard cost, the sequence  $\{q_n^\lambda \omega(q_n)\}$  is a minimax upper rate of convergence for the class  $\mathcal{D}^{(\lambda, \omega)}$ .  $\blacksquare$

As a corollary of Theorem 3.1, we generalize a well known result (see [1]) to binary classification with nonstandard cost. The proof is similar to that of Corollary 2.3, and therefore is omitted.

**Corollary 3.4** Let  $\eta^{(n)}(x)$  be a weakly consistent regression estimation of  $\eta(x)$ , that is

$$\lim_{n \rightarrow \infty} E\{\|\eta - \eta^{(n)}\|_1\} = 0.$$

Then for the plug-in classifier  $g_n$ , it holds

$$\lim_{n \rightarrow \infty} \frac{EL^{(\alpha)}(g_n) - L^{(\alpha)}(g^*)}{\sqrt{E\{\|\eta - \eta^{(n)}\|_2^2\}}} = 0.$$

For the strongly consistent approximation, as Corollary 2.4, we also have

**Corollary 3.5** Suppose that  $g_n$  is the plug-in classifier constructed from a strongly consistent approximation for  $\eta$ . Then with probability one

$$\lim_{n \rightarrow \infty} \frac{L^{(\alpha)}(g_n) - L^{(\alpha)}(g^*)}{\|\eta - \eta^{(n)}\|_2} = 0.$$

We turn to the minimax lower rate of convergence. With Theorem 3.1 and the method of establishing Theorem 2.6 we have

**Theorem 3.6** For the classification with nonstandard cost, the sequence  $\{q_n^\lambda \omega(q_n)\}$  is a minimax lower rate of convergence for the class  $\mathcal{D}^{(\lambda, \omega)}$ . ■

Finally, a combination of Theorem 3.3 and Theorem 3.5 gives the optimal rate of convergence as follow.

**Theorem 3.7** For the classification with nonstandard cost, the sequence  $\{q_n^\lambda \omega(q_n)\}$  is the optimal rate of convergence for the class  $\mathcal{D}^{(\lambda, \omega)}$ . ■

## References

- [1] L. Devroye, L. Györfi, G. Lugosi, A probability theory of pattern recognition. Springer Verlag, 1996.
- [2] V. Vapnik, Statistical learning theory. Wiley, 1998.
- [3] Y. Lee, Y. Lin, G. Wahba, Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. Journal of American Statistical Association, 2002.
- [4] D. R. Chen, D. H. Xiang, The consistency of multicategory support vector machines, Adv. in Comput. Math., to appear.
- [5] Q. Wu, D. X. Zhou, Analysis of support vector machine classification, submitted, 2003.
- [6] D. R. Chen, Q. Wu, Y. Ying, D. X. Zhou, Support vector machine soft margin classifier: error analysis, J.Machine Learning Research, 5(2004), 1143-1175.
- [7] A. Antos, Performance limits of nonparametric estimators. Ph D thesis, University of Budapest, 1999.
- [8] A. Antos, B. Kégl, T. Linder, G. Lugosi, Data-dependent margin-based generalization bounds for classification. Journal of Machine Learning Research, 3(2002), 73-98.
- [9] Y. H. Yang, Minimax nonparametric classification-part I: rates of convergence. IEEE Transaction on Information Theory, 7(1999), 2271-2284.
- [10] Y. H. Yang, Minimax nonparametric classification-part II: model selection for adaptation. IEEE Transaction on Information Theory, 7(1999), 2285-2292.
- [11] F. Cucker, S. Smale, On the Mathematical foundations of learning. Bulletin of the American Mathematical Society, 39(2001), 1-49.
- [12] G. G. Lorentz, Metric entropy and approximation, Bull. Amer. Math. Soc. 72(1966), 903-937.