Journal of Machine Learning Research 7 (2006) 2435-2447 Submitted 11/05; Revised 6/06; Published 10/06

# Consistency of Multiclass Empirical Risk Minimization Methods Based on Convex Loss

**Di-Rong Chen** DRCHEN@BUAA.EDU.CN
*Department of Mathematics, and LMIB*
*Beijing University of Aeronautics and Astronautics*
*Beijing 100083, P. R. CHINA*

**Tao Sun** ST.126@163.COM.CN
*Department of Mathematics*
*Beijing University of Aeronautics and Astronautics*
*Beijing 100083, P. R. CHINA*

**Editor:** Peter Bartlett

## Abstract

The consistency of classification algorithm plays a central role in statistical learning theory. A consistent algorithm guarantees us that taking more samples essentially suffices to roughly reconstruct the unknown distribution. We consider the consistency of ERM scheme over classes of combinations of very simple rules (base classifiers) in multiclass classification. Our approach is, under some mild conditions, to establish a quantitative relationship between classification errors and convex risks. In comparison with the related previous work, the feature of our result is that the conditions are mainly expressed in terms of the differences between some values of the convex function.

**Keywords:** multiclass classification, classifier, consistency, empirical risk minimization, constrained comparison method, Tsybakov noise condition

## 1. Introduction

We consider the consistency of empirical risk minimization (ERM) algorithm in multiclass classification.

Given an input vector $x \in \mathcal{X} \subseteq \mathbb{R}^d$, we would like to predict its corresponding label $y \in \{1,2,\ldots,K\}$. A classifier $f$ is a function defined on $\mathcal{X}$ with values in $\{1,2,\ldots,K\}$. The quality of this classifier can be measured by the classification error

$$\mathcal{R}(f) = \mathbb{E}_{X,Y} I_{\{f(X)\neq Y\}},$$

where $I_A$ is the characteristic function of set $A$, and $X,Y$ are drawn from an unknown underlying distribution $D$. It is clear that $\mathcal{R}(f) = \mathbb{P}\{Y \neq f(X)\}$. If we know the conditional density $\mathbb{P}\{Y = c|X = x\}$, then the classifier $\phi_B$ given by

$$\phi_B(x) := \arg \max_{c\in\{1,2,\ldots,K\}} \mathbb{P}\{Y = c|X = x\},$$

referred to as Bayes rule, minimizes $\mathcal{R}(f)$ over all classifiers: $\mathcal{R}(\phi_B) = \inf \mathcal{R}(f)$. Henceforth, let $\mathcal{R}^*$ stand for the number $\inf \mathcal{R}(f)$. However, the conditional density is unknown in practice.

Instead, we are given $n$ samples $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ of independent random variable drawn from the underlying distribution $D$. The goal of statistical learning is to find a classifier based on the samples and a pre-chosen set $\mathcal{F}$ of vector functions with $K$-components. For this purpose, a very successful method used in binary classification is to solve a minimization problem of a risk based on a convex loss $\phi$. Main examples of $\phi$ include the exponential loss $\phi(x)=e^{-x}$ used in AdaBoost, the logit loss $\phi(x)=\ln(1+e^{-x})$ and the hinge loss $\phi(x)=(1-x)_+$ used in support vector machine, where $(u)_+=\max\{0,u\}$ for a number $u\in\mathbb{R}$.

Probably since one can solve a multiclass classification problem $(K>2)$ by solving several binary classification problems, there are much fewer studies on multiclass classification algorithms based directly on minimizing empirical risk with convex loss. Recently, Zhang (2004b) proposes a natural version of EMR scheme in solving a multiclass problem:

$$\hat{\mathbf{f}}=\arg\min_{\mathbf{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\Psi_{Y_i}(\mathbf{f}(X_i)), \tag{1}$$

where $\Psi_c$ is a mapping from $\mathbb{R}^K$ to $\mathbb{R}$, which is usually constructed by some convex loss function $\phi$. In the following, we use bold symbols such as $\mathbf{f}$ and $\mathbf{q}$ to denote vectors, and $f_c$ and $q_c$ to denote their $c$-th component. We also use $\mathbf{f}(\cdot)$ to denote a vector function. Once obtaining $\hat{\mathbf{f}}$, we have a classifier $C(\hat{\mathbf{f}})$, where $C(\mathbf{f})$ is defined by

$$C(\mathbf{f})(x)=\arg\max_c f_c(x),\qquad \forall\mathbf{f}=(f_1(x),\ldots,f_K(x)).$$

A natural question is how close the optimal Bayes error $\mathcal{R}^*$ can be approximately reached by $\mathcal{R}(C(\hat{\mathbf{f}}))$. A very desirable property is the consistency of algorithm: the excess error $\mathcal{R}(C(\hat{\mathbf{f}}))-\mathcal{R}^*\to 0$ in some sense, as the size $n$ of samples increases to $\infty$. A consistent algorithm guarantees us that taking more samples essentially suffices to roughly reconstruct the unknown distribution. A good learning algorithm should be consistent.

In recent years, a large part of research has been focused on classifiers which base their decision on a certain combination of (base) classifiers. Suppose that $\mathcal{H}$ is a set of classifiers and $\lambda$ is a positive number. Let $\mathcal{F}=\mathcal{F}_\lambda$ be the following set of vector functions

$$\mathcal{F}_\lambda=\Big\{\mathbf{f}=\Big(\sum_{j=1}^{J}\beta_j T_c(h_j(\cdot))\Big)_{c=1}^{K}:\beta_j>0,h_j\in\mathcal{H},J=1,2,\ldots,\sum_{j=1}^{J}\beta_j=\lambda\Big\},$$

where $T_c, c=1,\ldots,K$, are functions defined on $\{1,2,\ldots,K\}$ by

$$T_c(h)=\begin{cases}K-1, & \text{if } h=c,\\ -1 & \text{if } h\neq c.\end{cases}$$

A classifier $C(\mathbf{f})$ with $\mathbf{f}\in\mathcal{F}_\lambda$ may be thought as one that, upon observing $x$, takes a weighted vote of classifiers $h_1,\ldots,h_J$, using weights $\beta_1,\ldots,\beta_J$.

For $K=2$, the vector function $\mathbf{f}=(f_1,f_2)\in\mathcal{F}_\lambda$ satisfies $f_1+f_2=0$. Therefore $\mathcal{F}_\lambda$ is usually regarded as the set of functions $f=\sum_{j=1}^{J}\beta_j T_1(h_j(\cdot)),h_j\in\mathcal{H},\sum_{j=1}^{J}\beta_j=\lambda$. In different versions of boosting, bagging and arcing algorithms, the output classifiers are constructed by weighted voting schemes. Their consistency is established in Lugosi and Vayatis (2004) under the assumption that the Bayes classifier can be approximated by $\mathcal{F}_\lambda$ and $\mathcal{H}$ has a finite VC dimension.

The computational feasibility of schemes (1) has been recognized all along. Moreover, in binary classification, as revealed recently in binary classification problem, a striking feature of ERM (1) using a convex loss is that one can upper bound the excess error by the excess $\{\Psi_c\}_c$-risk $\mathcal{E}(\hat{\mathbf{f}}) - \mathcal{E}^*$, where $\mathcal{E}(\mathbf{f}) = \mathbb{E}_{X,Y}\Psi_Y(\mathbf{f}(X))$ is the expectation of $\Psi_Y(\mathbf{f}(X))$, referred to as the $\{\Psi_c\}$-risk, and $\mathcal{E}^*$ is the infimum $\inf_{\mathbf{f}} \mathcal{E}(\mathbf{f})$ of $\mathcal{E}(\mathbf{f})$ over an appropriate set (not restricted to $\mathcal{F}$). Consequently, we have a very important implication relation (e.g., Bartlett et al., 2005; Lugosi and Vayatis, 2004; Chen et al., 2004; Zhang, 2004a)

$$\mathcal{E}(\hat{\mathbf{f}}) \to \mathcal{E}^* \Rightarrow \mathcal{R}(C(\hat{\mathbf{f}})) \to \mathcal{R}^*.$$

The notion of classification calibrated in Bartlett et al. (2005) is extended to multiclass classification problem and is used to characterize above implication in Tewari and Bartlett (2005). Such an implication is also established under the so called infinite-sample-consistency (ISC) condition on $\{\Psi_c\}_c$ (see Zhang, 2004b). Moreover, an quantitative relation between the excess error and the excess $\{\Psi_c\}$-risk is obtained for One-versus-All method in Zhang (2004b).

In this paper we consider the constrained comparison method in multiclass classification problem. One of our goals is to generalize the results of consistency for weighted voting schemes in Lugosi and Vayatis (2004) to multiclass case. We first establish an inequality concerning with the excess error and the excess $\{\Psi_c\}$-risk. The inequality is interesting in its own right.

The paper is organized as following. In Section 2, we upper bound the excess error by the excess $\{\Psi_c\}$-risk under some mild conditions. In comparison with the previous work, our conditions are mainly expressed in terms of the differences between some values of function $\phi$. On the other hand, the sufficient conditions ensuring the quantitative relationships, even in case $K = 2$, are expressed previously in terms of the infimum $\inf_{\mathbf{f}} \mathbb{E}(\Psi_Y(\mathbf{f}(X))|X = x)$. In Section 3, we apply the results in Section 2 to establish a consistency result in multiclass case, similar to that of Lugosi and Vayatis (2004).

## 2. Bounding Classification Error by Convexity

In this section, we upper bound, under some conditions on convex loss $\phi$, the excess classification error $\mathcal{R}(C(\mathbf{f})) - \mathcal{R}^*$ by the excess $\{\Psi_c\}$-risk $\mathcal{E}(\mathbf{f}) - \mathcal{E}^*$ for the constrained comparison method. Similar result is established for the One-versus-All method (see Zhang, 2004b). The two methods are different: in the One-versus-All method, one can deal with each component of the vector function separately. The conditions and proofs here are different from those in Zhang (2004b). Moreover, a tighter upper bound is given under Tsybakov noise condition.

Recall that $\mathbb{P}\{Y = c|X = x\}$ is the conditional probability. Let

$$\mathbf{q}(x) = (q_c(x))_{c=1}^K, \qquad q_c(x) = \mathbb{P}\{Y = c|X = x\}.$$

Suppose that $\phi$ is a convex function on $\mathbb{R}$. The constrained comparison method proposed in Zhang (2004b) uses $\Psi_c$ below.

$$\Psi_c(\mathbf{f}) = \sum_{k=1,k\neq c}^{K} \phi(-f_k), \qquad \mathbf{f} \in \Omega := \Big\{\mathbf{f} \in \mathbb{R}^K : \sum_c f_c = 0\Big\}.$$

Then the risk $\mathcal{E}(\mathbf{f})$ may by expressed as

$$\mathcal{E}(\mathbf{f}) = \mathbb{E}_X W(\mathbf{q}(X), \mathbf{f}(X)), \tag{2}$$

with $W(\mathbf{q},\mathbf{f}) = \sum_{c=1}^{K}(1-q_c)\phi(-f_c)$.

Note that we use $q_c$ to denote the $c$-th component of a $K$-dimensional vector $\mathbf{q} \in \Lambda_K$, where $\Lambda_K$ is the set of possible conditional probability vectors:

$$\Lambda_K = \Big\{ q \in \mathbb{R}^K : \sum_{c=1}^{K} q_c = 1, q_c \geq 0 \Big\}.$$

Denote by $\mathcal{B}$ the set of all $K$-dimensional vectors of Borel measurable functions on $\mathcal{X}$ and $\mathcal{B}_\Omega = \{\mathbf{f} \in \mathcal{B} : \forall x \in \mathcal{X}, \mathbf{f}(x) \in \Omega\}$. Let $\mathcal{E}^* = \inf_{\mathbf{f}\in\mathcal{B}_\Omega} \mathcal{E}(\mathbf{f})$.

For any $\mathbf{q} \in \Lambda_K$, let $W^*(\mathbf{q}) := \inf_{\mathbf{f}\in\Omega} W(\mathbf{q},\mathbf{f})$. It is easily seen that

$$\mathcal{E}^* = \mathbb{E}W^*(\mathbf{q}(X)).$$

**Lemma 2.1** *Assume that $\phi$ is a decreasing and convex function on $\mathbb{R}$. Let $W(\mathbf{q},\mathbf{f})$ be given as above. Suppose that $\mathbf{q} \in \Lambda_K$ and $\mathbf{f} \in \mathcal{B}_\Omega$ satisfy that there are $i, j$ such that $q_i < q_j$ and $f_j < f_i$. Then $W(\mathbf{q},\mathbf{f}') \leq W(\mathbf{q},\mathbf{f})$, where $\mathbf{f}' = (f_1', \cdots, f_K')$ is given by $f_i' = f_j' = \frac{f_i+f_j}{2}$, and $f_c' = f_c$, $c \neq i, j$.*

**Proof.** Without loss of generally, we can assume that $q_1 < q_2$ and $f_2 < f_1$. Then

$$\frac{f_1+f_2}{2} \leq \frac{(1-q_1)f_1 + (1-q_2)f_2}{2-q_1-q_2}.$$

By assumption, we have

$$\begin{aligned}(2-q_1-q_2)\phi\Big(-\tfrac{f_1+f_2}{2}\Big) &\leq (2-q_1-q_2)\phi\Big(-\tfrac{(1-q_1)f_1+(1-q_2)f_2}{2-q_1-q_2}\Big)\\ &\leq (1-q_1)\phi(-f_1) + (1-q_2)\phi(-f_2).\end{aligned}$$

Therefore the proof is complete by

$$\begin{aligned} & W(\mathbf{q},\mathbf{f}) - W(\mathbf{q},\mathbf{f}')\\ = \;& (1-q_1)\phi(-f_1) + (1-q_2)\phi(-f_2) - (2-q_1-q_2)\phi(-\tfrac{f_1+f_2}{2}) \geq 0.\end{aligned}$$

■

**Lemma 2.2** *Assume that $\phi$ is a decreasing and convex function on $\mathbb{R}$. Suppose that there exist positive constants $k > 0$ and $\alpha \geq 1$ such that for any $\mathbf{q} \in \Lambda_K$,*

$$k(q_j - q_i)^\alpha \leq W^*(\mathbf{q}') - W^*(\mathbf{q}), \tag{3}$$

*where $j = \arg\max_c q_c$ and $q_i < q_j$, and $\mathbf{q}'$ is given by $\mathbf{q}' = (q_1', \cdots, q_K')$, where $q_i' = q_j' = \frac{q_i+q_j}{2}$, and $q_c' = q_c, c \neq i, j$. Then for any $\mathbf{f} \in \mathcal{B}_\Omega$,*

$$k(\mathcal{R}(C(\mathbf{f})) - \mathcal{R}^*) \leq \mathcal{E}(\mathbf{f}) - \mathcal{E}^*)^{\frac{1}{\alpha}}.$$

**Proof.** Recall that $q_c(x)$ is the conditional probability $\mathbb{P}\{Y = c|X = x\}$. For any $\mathbf{f}$ we have by definition of $\mathcal{R}(C(\mathbf{f}))$

$$\mathcal{R}(C(\mathbf{f})) - \mathcal{R}^* = \int_{\mathcal{X}} \Big( q_{\phi_B(x)}(x) - q_{C(\mathbf{f})(x)}(x) \Big) d\rho_X. \tag{4}$$

Let $\mathbf{q}(x) = (q_c(x))_{c=1}^K$. Also by (2)

$$\mathcal{E}(\mathbf{f}) - \mathcal{E}^* = \int_{\mathcal{X}} \Big( W(\mathbf{q}(\mathbf{x}), \mathbf{f}(\mathbf{x})) - W^*(\mathbf{q}(x)) \Big) d\rho_X. \tag{5}$$

Let $x \in \mathcal{X}$ be given such that $q_{C(\mathbf{f})(x)}(x) \neq q_{\phi_B(x)}(x)$. Denote $j = \phi_B(x)$ and $i = C(\mathbf{f})(x)$. We regard $\mathbf{q}(x)$, $\mathbf{f}(x)$ and $\mathbf{f}'(x)$ as $\mathbf{q}$ , $\mathbf{f}$ and $\mathbf{f}'$ in Lemma 2.1 respectively.

By assumption, we have $W(\mathbf{q}(x), \mathbf{f}'(x)) = W(\mathbf{q}'(x), \mathbf{f}'(x)) \geq W^*(\mathbf{q}'(x))$. It follows from Lemma 2.1 that $W^*(\mathbf{q}'(x)) \leq W(\mathbf{q}(x), \mathbf{f}(x))$. Therefore by (3)

$$k(q_j(x) - q_i(x))^\alpha \leq W(\mathbf{q}(x), \mathbf{f}(x)) - W^*(\mathbf{q}(x)).$$

Integrating the above inequality over the set $\mathcal{X}' = \{x \in \mathcal{X} : C(\mathbf{f})(x) \neq \phi_B(x)\}$, we have

$$k \int_{\mathcal{X}'} \Big( q_{\phi_B(x)}(x) - q_{C(\mathbf{f})(x)}(x) \Big)^\alpha d\rho_X \leq \int_{\mathcal{X}'} \Big( W(\mathbf{q}(\mathbf{x}), \mathbf{f}(\mathbf{x})) - W^*(\mathbf{q}(x)) \Big) d\rho_X.$$

By Hölder inequality, for $\alpha \geq 1$

$$\Big( \int_{\mathcal{X}'} \Big( q_{\phi_B(x)}(x) - q_{C(\mathbf{f})(x)}(x) \Big) d\rho_X \Big)^\alpha \leq \int_{\mathcal{X}'} \Big( q_{\phi_B(x)}(x) - q_{C(\mathbf{f})(x)}(x) \Big)^\alpha d\rho_X.$$

Then we have the desired inequality by the definition of $\mathcal{X}'$, (4) and (5). The proof is complete. ■

In the following we impose some conditions on $\phi$.

**Assumption 2.3** *1. $\phi$ is a differentiable, convex and decreasing function on $\mathbb{R}$ such that $\lim_{x \to +\infty} \phi(x) = 0$ and $\lim_{x \to -\infty} \phi(x) = +\infty$.*

*2. For any $\mathbf{q} = (q_c)_{c=1}^K \in \Lambda_K$ with all $q_c < 1, c \in \{1, \ldots, K\}$, there is a minimizer $\mathbf{f}^* = (f_c^*)_{c=1}^K$ of $W(\mathbf{q}, \mathbf{f})$. Moreover, $\phi$ is twice differentiable at points $-f_c^*, c = 1, \ldots, K$, and $\phi''(-f_c^*) > 0, c \in \{1, \ldots, K\}$.*

For any $\mathbf{q} = (q_c)_{c=1}^K$, let $j = \arg\max_c q_c$ and $i \in \{1, \ldots, K\}$ with $q_i < q_j$. We introduce $\mathbf{q}^t = (q_c^t)_{c=1}^K \in \Lambda_K$ for $0 \leq t \leq \frac{q_j - q_i}{2}$ as following.

$$q_i^t = q_i + t,\ q_j^t = q_j - t, \text{ and } q_c^t = q_c,\ c \neq i, j.$$

Clearly, $q_c^t < 1$ for $0 < t < \frac{q_j - q_i}{2}$ and any $1 \leq c \leq K$. Therefore, for any $t$, there is a $\mathbf{f}^{t,*} = (f_c^{t,*})_{c=1}^K$ minimizing $W(\mathbf{q}^t, \mathbf{f})$, that is, $W^*(\mathbf{q}^t) = W(\mathbf{q}^t, \mathbf{f}^{t,*})$.

Under a condition weaker than Assumption 2.3, Zhang (2004b) proves that the excess error $\mathcal{R}(C(\mathbf{f})) - \mathcal{R}^*$ is small whenever the excess $\{\Psi\}$-risk $\mathcal{E}(\mathbf{f}) - \mathcal{E}^*$ is small. Our goal however is, under Assumption 2.3, to establish an inequality between the above two quantities. We give a sufficient condition for (3) in terms of the differences between any pair of $\phi(-f_c^{t,*}), c \in \{1, \ldots, K\}$.

**Theorem 2.4** *Assume that $\phi$ satisfies Assumption 2.3. Suppose that there exist positive constants $k_1 > 0$ and $\beta \geq 0$ such that for any $\mathbf{q} \in \Lambda_K$,*

$$k_1(q_j - q_i - 2t)^\beta \leq \phi(-f_j^{t,*}) - \phi(-f_i^{t,*}), \qquad 0 < t < \frac{q_j - q_i}{2}, \tag{6}$$

*whenever* $j = \arg\max_c q_c$, $q_i < q_j$ *and* $\mathbf{f}^{t,*} = (f_c^{t,*})_{c=1}^K$ *is a minimizer of* $W(\mathbf{q}^t, \mathbf{f})$. *Then for any vector* $\mathbf{f} \in \mathcal{B}_\Omega$,

$$\mathcal{R}(C(\mathbf{f})) - \mathcal{R}^* \leq \frac{2(\beta+1)}{k_1} (\mathcal{E}(\mathbf{f}) - \mathcal{E}^*)^{\frac{1}{\beta+1}}.$$

**Proof.** We establish condition (3) with $\alpha = \beta + 1$ and $k = \frac{k_1}{2(\beta+1)}$. As above, let $\mathbf{f}^{t,*} = (f_c^{t,*})_{c=1}^K$ be the minimizer of $W(\mathbf{q}^t, \mathbf{f})$. The first-order optimality condition is the set of equations

$$(1 - q_c^t)\phi'(-f_c^{t,*}) = \mu, \qquad c = 1, \ldots, K,$$

where $\mu$, independent of $c$, is the Lagrangian multiplier. Assumption 2.3 implies that $f_{c,t}$ is differentiable with respect to $t, c = 1, \ldots, K$. Moreover, the constraint $\sum_{c=1}^K f_c^{t,*} = 0 \, (\forall t \in (0, \frac{q_2 - q_1}{2}))$ yields $\sum_{c=1}^K \frac{d f_c^{t,*}}{dt} = 0$. Consequently,

$$\begin{aligned} & \frac{dW^*(\mathbf{q}^t)}{dt} \\ = \; & \phi(-f_j^{t,*}) - \phi(-f_i^{t,*}) - \sum_{c=1}^K (1 - q_c^t)\phi'(-f_c^{t,*}) \frac{d f_c^{t,*}}{dt} \\ = \; & \phi(-f_j^{t,*}) - \phi(-f_i^{t,*}). \end{aligned}$$

Therefore, we have by (6)

$$\frac{dW^*(\mathbf{q}^t)}{dt} \geq k_1 (q_j - q_i - 2t)^\beta, \qquad 0 < t < \frac{q_j - q_i}{2}.$$

Integrating the above inequality over $[0, \frac{q_j - q_i}{2}]$ gives (3) with $\alpha = \beta + 1$ and $k = \frac{k_1}{2(\beta+1)}$. Our conclusion follows from Lemma 2.2. The proof is complete. ■

We consider the exponential loss as the first example.

**Example 2.5** *Let* $\phi(x) = e^{-x}$. *Then for any vector* $\mathbf{f} \in \mathcal{B}_\Omega$, *we have*

$$\mathcal{R}(C(\mathbf{f})) - \mathcal{R}^* \leq \frac{4\sqrt[K]{K-1}}{\sqrt[K]{(\frac{2K}{2K-1})^{2K-2}}} \sqrt{\mathcal{E}(\mathbf{f}) - \mathcal{E}^*}.$$

**Proof.** For $\mathbf{q} = (q_c)_{c=1}^K \in \Lambda_K$ with all $q_c < 1$, the unique minimizer $\mathbf{f}^* = (f_c^*)_{c=1}^K$ is determined by $(1 - q_c)\exp(f_c^*) = \mu, c = 1, \cdots, K$, with $\mu$ the Lagrangian multiplier. Assumption 2.3 holds for $\phi$. By $\sum_c f_c^* = 0$ we have $\mu = \sqrt[K]{\prod_{c=1}^K (1 - q_c)}$. Therefore

$$\phi(-f_k^*) = \frac{\sqrt[K]{\prod_{c=1}^K (1 - q_c)}}{1 - q_k}, \qquad k = 1, \ldots, K.$$

Let $j = \arg\max_c q_c$ and $i \in \{1, \ldots, K\}$ such that $q_i < q_j$. Recall that $\mathbf{q}^t$ and $\mathbf{f}^{t,*}$ be defined as before. We apply the above equality and obtain

$$\phi(-f_j^{t,*}) - \phi(-f_i^{t,*}) = \frac{\sqrt[K]{\prod_{c \neq i,j} (1 - q_c)}}{((1 - q_j + t)(1 - q_i - t))^{\frac{K-1}{K}}} (q_j - q_i - 2t).$$

If $K=2$, $\prod_{c\neq i,j}(1-q_c)$ is understood as 1. If $K>2$, the $K-2$ nonnegative numbers $q_c, c\neq i,j$, may be arranged in the decreasing order, so that it is easily seen that they are not larger than $\frac{1}{2},\ldots,\frac{1}{K-1}$ respectively. Therefore

$$\prod_{c\neq i,j}(1-q_c)\geq\prod_{c=2}^{K-1}(1-\frac{1}{c})=\frac{1}{K-1}.$$

On the other hand, $(1-q_j+t)(1-q_i-t)\leq(1-\frac{q_i+q_j}{2})^2$ for $0\leq t\leq\frac{q_i-q_j}{2}$. Note $\frac{q_i+q_j}{2}\geq\frac{q_j}{2}\geq\frac{1}{2K}$. Consequently,

$$\phi(-f_j^{t,*})-\phi(-f_i^{t,*})\geq\frac{\sqrt[K]{(\frac{2K}{2K-1})^{2K-2}}}{\sqrt[K]{K-1}}(q_j-q_i-2t).$$

This is (6) with $\beta=1$ and $k_1=\frac{\sqrt[K]{(\frac{2K}{2K-1})^{2K-2}}}{\sqrt[K]{K-1}}$. The conclusion follows from Theorem 2.4. ■

Let $p\geq 1$ and $\phi(x)=(\frac{1}{K-1}-x)_+^p$, where $(x)_+=\max\{x,0\}$. The resulting risk is just the one used in $p$-norm Support vector machine (SVM). Chen and Xiang (2004) have established the inequality for $p=1$

$$\frac{\mathcal{R}(C(\mathbf{f}))-\mathcal{R}^*}{K-1}\leq\mathcal{E}(\mathbf{f})-\mathcal{E}^*.$$

**Example 2.6** *Let $\phi(x)=(\frac{1}{K-1}-x)_+^2$. Then for any vector $\mathbf{f}\in\mathcal{B}_\Omega$, we have*

$$\mathcal{R}(C(\mathbf{f}))-\mathcal{R}^*\leq\frac{4(\frac{K-1}{K})^2}{k_2}\sqrt{\mathcal{E}(\mathbf{f})-\mathcal{E}^*},$$

*where $k_2=2(\frac{2K-1}{2K})^2+(\frac{2K-1}{2K})^4\left(\left(\frac{1}{2}\right)^2+\cdots+\left(\frac{K-2}{K-1}\right)^2\right)$ for $K>2$ and $k_2=\frac{1}{8}$ for $K=2$.*

**Proof.** For $\mathbf{q}=(q_c)_{c=1}^k$ with all $q_c<1$, by the method of Lagrange multiplier we conclude that the minimizer $\mathbf{f}^*=(f_c^*)_{c=1}^K$ satisfies $-f_c^*<\frac{1}{K-1}, c\in\{1,\ldots,K\}$. Thus, Assumption 2.3 is satisfied by $\phi$. Moreover, we have

$$\phi(-f_k^*)=\left(\frac{K}{K-1}\right)^2\frac{1}{(1-q_k)^2\sum_{c=1}^K\frac{1}{(1-q_c)^2}},\qquad k=1,\ldots,K.$$

Let $j=\arg\max_c q_c$ and $i\in\{1,\ldots,K\}$ such that $q_i<q_j$. Moreover, $\mathbf{q}^t$ and $\mathbf{f}^{t,*}$ are defined as before. An application of the above equality to $\mathbf{q}^t$ and $\mathbf{f}^{t,*}$ yields

$$\begin{aligned}&\phi(-f_j^{t,*})-\phi(-f_i^{t,*})\\ =\;&\left(\frac{K}{K-1}\right)^2\frac{(q_j-q_i-2t)(2-q_i-q_j)}{(1-q_i-t)^2(1-q_j+t)^2\left(\frac{1}{(1-q_i+t)^2}+\frac{1}{(1-q_j+t)^2}+\sum_{c\neq i,j}\frac{1}{(1-q_c)^2}\right)},\end{aligned}$$

where $\sum_{c\neq i,j}\frac{1}{(1-q_c)^2}$ is understood as 0 for $K=2$. It is easily seen that

$$\begin{aligned}&(1-q_i-t)^2(1-q_j+t)^2\left(\frac{1}{(1-q_i+t)^2}+\frac{1}{(1-q_j+t)^2}\right)\\ \leq\;&2(1-\frac{q_j+q_i}{2})^2\leq 2(\frac{2K-1}{2K})^2,\qquad\qquad\forall t\in[0,\frac{q_j-q_i}{2}],\end{aligned}$$

where the second inequality holds by $1/K \le q_j$.

As in Example 2.5, again we arrange $q_c, c \neq i, j$, in decreasing order so that they are not larger than $\frac{1}{2}, \ldots, \frac{1}{K-1}$ respectively. It follows that, for $0 \le t \le \frac{q_j - q_i}{2}$,

$$(1-q_i-t)^2(1-q_j+t)^2 \sum_{c\neq i,j} \frac{1}{(1-q_c)^2} \le \Big(\frac{2K-1}{2K}\Big)^4 \Big(\Big(\frac{1}{2}\Big)^2 + \cdots + \Big(\frac{K-2}{K-1}\Big)^2\Big).$$

Therefore, the condition (6) holds with $\beta = 1$ and $k_1 = \Big(\frac{K}{K-1}\Big)^2 k_2$. The conclusion follows from Theorem 2.4. ■

**Remark 2.7** *For $\phi(x) = (\frac{1}{K-1} - x)_+^p$ with $p > 1$, we can also apply Theorem 2.4 and get an inequality $\mathcal{R}(C(\mathbf{f})) - \mathcal{R}^* \le k'\sqrt{\mathcal{E}(\mathbf{f}) - \mathcal{E}^*}$, where $k'$ is a constant. The argument is similar to that of Example 2.8. We point out that $-f_c^* < \frac{1}{K-1}$ for any $\mathbf{q} = (q_c)_{c=1}^K$ with all $q_c < 1, c = 1, \ldots, K$, which ensures that $\phi$ satisfies Assumption 2.3. Moreover, by simple computation,*

$$\phi(-f_k^*) = \Big(\frac{K}{K-1}\Big)^{\frac{p}{p-1}} \frac{1}{\sum_{c=1}^K \Big(\frac{1-q_k}{1-q_c}\Big)^{\frac{p}{p-1}}}, \qquad k = 1, \ldots, K.$$

The bounds in Lemma 2.2 and Theorem 2.4 may be improved under the so-called Tsybakov noise condition. For any $x \in \mathcal{X}$, let

$$m(x) = q_{\phi_B(x)}(x) - \max\{q_i(x) : q_i(x) < q_{\phi_B(x)}(x), i = 1 \ldots, K\}$$

if the set $\{q_i(x) : q_i(x) < q_{\phi_B(x)}(x), i = 1 \ldots, K\}$ is not empty, and $m(x) = 0$ otherwise.

**Definition 2.8** *Let $s \in [0,1]$. We say that $\mathbb{P}$ satisfies Tsybakov noise condition with exponent $s$, if there is a constant $c$ such that*

$$\mathbb{P}\{X \in \mathcal{X} : 0 < m(X) < t\} \le c t^{\frac{s}{1-s}}, \qquad 0 < t \le 1.$$

As in binary classification (see Bartlett and Mendelson, 2002), Tsybakov noise condition with exponent $s$ implies that there is a constant $c$ such that, for any $\mathbf{f} \in \mathcal{B}_\Omega$,

$$\mathbb{P}\{x : x \in \mathcal{X}, q_{\phi_B(x)}(x) \neq q_{C(\mathbf{f})(x)}(x)\} \le c(\mathcal{R}(C(\mathbf{f})) - \mathcal{R}^*)^s. \tag{7}$$

In fact, Tsybakov noise condition and (4) tell us

$$\begin{aligned} & \mathcal{R}(C(\mathbf{f})) - \mathcal{R}^* \\ \ge & \int_{\mathcal{X}} \Big(q_{\phi_B(x)}(x) - q_{C(\mathbf{f})(x)}(x)\Big) I_{\{t \le m(x)\}} d\rho_X \\ \ge & \; t\Big(\mathbb{P}\{x : x \in \mathcal{X}, q_{\phi_B(x)}(x) \neq q_{C(\mathbf{f})(x)}(x)\} - c t^{\frac{s}{1-s}}\Big). \end{aligned}$$

Minimizing the last term over $t$ establishes (7).

**Theorem 2.9** *Suppose that $\mathbb{P}$ satisfies Tsybakov noise condition with exponent s. If the conditions of Lemma 2.2 are satisfied, then for any vector $\mathbf{f} \in \mathcal{B}_\Omega$ we have*

$$\mathcal{R}(C(\mathbf{f})) - \mathcal{R}^* \le k_\phi (\mathcal{E}(\mathbf{f}) - \mathcal{E}^*)^{\frac{1}{\alpha-(\alpha-1)s}}, \tag{8}$$

*where $k_\phi$ is a constant.*

*Consequently, under Tsybakov noise condition with exponent s and conditions of Theorem 2.4, we have for any vector $\mathbf{f} \in \mathcal{B}_\Omega$*

$$\mathcal{R}(C(\mathbf{f})) - \mathcal{R}^* \le k_\phi (\mathcal{E}(\mathbf{f}) - \mathcal{E}^*)^{\frac{1}{\beta+1-\beta s}}.$$

**Proof.** For $\mathbf{f} \in \mathcal{B}_\Omega$ and $t \in (0,1]$ set $\mathcal{X}_1 = \{x : x \in \mathcal{X}, 0 < q_{\phi_B(x)}(x) - q_{C(\mathbf{f})(x)}(x) < t\}$ and $\mathcal{X}_2 = \{x : x \in \mathcal{X}, t \le q_{\phi_B(x)}(x) - q_{C(\mathbf{f})(x)}(x)\}$. Clearly, $\mathcal{X}_1 \subseteq \{x : x \in \mathcal{X}, q_{\phi_B(x)}(x) \ne q_{C(\mathbf{f})(x)}(x)\}$, which implies $\mathbb{P}(\mathcal{X}_1) \le c(\mathcal{R}(C(\mathbf{f})(x)) - \mathcal{R}^*)^s$ by (7). On the other hand,

$$\begin{aligned} & \int_{\mathcal{X}_2} \Big( q_{\phi_B(x)}(x) - q_{C(\mathbf{f})(x)}(x) \Big) d\rho_X \\ \le \;& t^{-\alpha+1} \int_{\mathcal{X}} \Big( q_{\phi_B(x)}(x) - q_{C(\mathbf{f})(x)}(x) \Big)^\alpha d\rho_X \\ \le \;& \frac{1}{kt^{\alpha-1}} (\mathcal{E}(\mathbf{f}) - \mathcal{E}^*), \end{aligned}$$

where the last inequality follows from the proof of Lemma 2.2. Therefore we have by (4) that

$$\mathcal{R}(C(\mathbf{f})) - \mathcal{R}^* \le tc(\mathcal{R}(C(\mathbf{f})) - \mathcal{R}^*)^s + \frac{1}{kt^{\alpha-1}} (\mathcal{E}(\mathbf{f}) - \mathcal{E}^*).$$

Minimizing the right hand side of above inequality over $t \in (0,1]$ yields the inequality (8) for some constant $c_\phi$.

As a consequence, the second conclusion follows from Theorem 2.4 and (8) with $\alpha = \beta + 1$. The proof is complete. ■

## 3. Consistency of Weighted Voting Schemes

In this section, we consider the consistency of weight voting schemes by the results of section 2.

Recall that $\mathcal{B}_\Omega$ is given in Section 2. It is easily seen that, for any set $\mathcal{H}$ of classifiers, $\mathcal{F}_\lambda \subset \mathcal{B}_\Omega$.

**Assumption 3.1** *Recall that $\mathcal{E}^*$ is defined in Section 2. Suppose that the set $\mathcal{H}$ of classifiers satisfies*

$$\lim_{\lambda \to \infty} \inf_{\mathbf{f} \in \mathcal{F}_\lambda} \mathcal{E}(\mathbf{f}) = \mathcal{E}^*.$$

The notion of VC dimension plays an important role in classification (see Devroye et al., 1996; Vapnik, 1998). Recall that for a collection $\mathcal{A}$ of some sets $A$, the VC dimension $V_\mathcal{A}$ of $\mathcal{A}$ is defined to be the largest number $d$, when exists, such that $\mathcal{A}$ shatters a set of some $d$ points (see Devroye et al., 1996). If there exists no such an integer $d$ we define $V_\mathcal{A} = \infty$.

With $n$ samples $\{(X_i, Y_i)\}_{i=1}^n \subset Z^n$, the empirical $\{\Psi_c\}$-risk $\mathcal{E}_n(\mathbf{f})$ of a vector function $\mathbf{f}$ is defined by

$$\mathcal{E}_n(\mathbf{f}) = \frac{1}{n} \sum_{i=1}^n \Psi_{Y_i}(\mathbf{f}(X_i)).$$

Clearly, $\mathcal{E}(\mathbf{f}) = \mathbb{E}_{Z^n} \mathcal{E}_n(\mathbf{f})$.

**Lemma 3.2** *Suppose that $\phi$ satisfies the condition 1 of Assumption 2.3. Moreover, suppose that, for any $c \in \{1,\ldots,K\}$, the collection $\mathcal{A}_c$ of all sets*

$$\{(x,c) : h(x) \neq c\}, \qquad h \in \mathcal{H},$$

*has a finite VC dimension $V_{\mathcal{A}_c}$. Then for any $n$ and $\lambda > 0$ we have*

$$\mathbb{E} \sup_{\mathbf{f} \in \mathcal{F}_\lambda} |\mathcal{E}(\mathbf{f}) - \mathcal{E}_n(\mathbf{f})| \leq 4K^2\lambda |\phi'(-\lambda(K-1))| \sqrt{\frac{2V \ln(4n+2)}{n}}, \tag{9}$$

*where $V = \max_{1 \leq c \leq K} V_{\mathcal{A}_c}$. Also, for any $\delta > 0$, with probability at least $1-\delta$,*

$$\begin{aligned} & \sup_{\mathbf{f} \in \mathcal{F}_\lambda} |\mathcal{E}(\mathbf{f}) - \mathcal{E}_n(\mathbf{f})| \\ \leq \;& 4K^2\lambda |\phi'(-\lambda(K-1))| \sqrt{\frac{2V \ln(4n+2)}{n}} + 2\exp\Big(\frac{-n\delta^2}{2(K-1)^2\phi^2(-\lambda K)}\Big). \end{aligned} \tag{10}$$

**Proof.** The proof is similar to that of Lugosi and Vayatis (2004) Lemma 2. Let $\sigma_1,\ldots,\sigma_n$ be the independent symmetric sign variables, that is,

$$\mathbb{P}\{\sigma_i = -1\} = \mathbb{P}\{\sigma_i = 1\} = \frac{1}{2}.$$

Then, by a standard symmetrization argument,

$$\mathbb{E} \sup_{\mathbf{f} \in \mathcal{F}_\lambda} |\mathcal{E}(\mathbf{f}) - \mathcal{E}_n(\mathbf{f})| \leq 2\mathbb{E} \sup_{\mathbf{f} \in \mathcal{F}_\lambda} \Big| \frac{1}{n} \sum_{i=1}^n \sigma_i (\Psi_{Y_i}(\mathbf{f}(X_i)) - (K-1)\phi(0)) \Big|.$$

On the other hand, it is easily seen that

$$\begin{aligned} & \sup_{\mathbf{f} \in \mathcal{F}_\lambda} \Big| \frac{1}{n} \sum_{i=1}^n \sigma_i (\Psi_{Y_i}(\mathbf{f}(X_i)) - (K-1)\phi(0)) \Big| \\ = \;& \sup_{\mathbf{f} \in \mathcal{F}_1} \Big| \frac{1}{n} \sum_{i=1}^n \sigma_i \sum_{c=1, c \neq Y_i}^K (\phi(-f_c(X_i)) - \phi(0)) \Big| \\ \leq \;& \sum_{c=1}^K \sup_{\mathbf{f} \in \mathcal{F}_1} \Big| \frac{1}{n} \sum_{i=1}^n \sigma_i (\phi(-\lambda f_c(X_i)) - \phi(0)) \Big|, \end{aligned}$$

where the equality holds by the definition of $\mathcal{F}_\lambda$.

For any $c \in \{1,\ldots,K\}$, let $g(t) = \phi(-\lambda t) - \phi(0), t \in [-1, K-1]$. Then $g(0) = 0$, and $g$ satisfies Lipschitz condition with Lipschitz constant $L = -\lambda\phi'(-\lambda(K-1))$. We appeal to the "contraction principle" to conclude for any $c \in \{1,\ldots,K\}$

$$\mathbb{E} \sup_{\mathbf{f} \in \mathcal{F}_1} \Big| \frac{1}{n} \sum_{i=1}^n \sigma_i (\phi(-\lambda f_c(X_i)) - \phi(0)) \Big| \leq 2L\mathbb{E} \sup_{\mathbf{f} \in \mathcal{F}_1} \Big| \frac{1}{n} \sum_{i=1}^n \sigma_i f_c(X_i) \Big|,$$

and consequently,

$$\mathbb{E} \sup_{\mathbf{f} \in \mathcal{F}_\lambda} \Big| \frac{1}{n} \sum_{i=1}^n \sigma_i (\Psi_{Y_i}(\mathbf{f}(X_i)) - 1) \Big| \leq 2L \sum_{c=1}^K \mathbb{E} \sup_{\mathbf{f} \in \mathcal{F}_1} \Big| \frac{1}{n} \sum_{i=1}^n \sigma_i f_c(X_i) \Big|. \tag{11}$$

Since any $f_c = \sum_j \alpha_j T_c(h_j)$ is a convex combination of $T_c(h_j)$ with $h_j \in \mathcal{H}$, it follows that

$$\sup_{\mathbf{f}\in\mathcal{F}_1} \Big| \frac{1}{n}\sum_{i=1}^n \sigma_i f_c(X_i) \Big| = \sup_{h\in\mathcal{H}} \Big| \frac{1}{n}\sum_{i=1}^n \sigma_i T_c(h(X_i)) \Big|. \tag{12}$$

With $X_i, i = 1,\ldots,n$, fixed, $\sum_{i=1}^n \sigma_i T_c(h(X_i))$ is a sum of $n$ independent zero mean random variables bounded between $-1$ and $K-1$. The coefficients satisfy $T_c(h(X_i)) = K-1-KI_{\{h(X_i)\neq c\}}$. By a version of the Vapnik-Chervonenkis inequality we conclude

$$\mathbb{E}\sup_{h\in\mathcal{H}} \Big| \frac{1}{n}\sum_{i=1}^n \sigma_i T_c(h(X_i)) \Big| \le (K-1)\sqrt{\frac{2V_{\mathcal{A}_c}\ln(4n+2)}{n}}, \qquad c = 1,\ldots,K.$$

The details are referred to Lugosi and Vayatis (2004). Summing the last inequalities for $c = 1,\ldots,K$ and appealing to (11) and (12) we prove (9).

It is easily seen that the random variable $\sup_{\mathbf{f}\in\mathcal{F}_\lambda} |\mathcal{E}(\mathbf{f}) - \mathcal{E}_n(\mathbf{f})|$ satisfies the bounded difference assumption with constant $c_i = 2(K-1)\phi(-\lambda K)/n, 1 \le i \le n$. Now inequality (10) follows from (9) and McDiarmid's bounded difference inequality (see Lugosi, 2002; McDiarmid, 1989). The proof is complete. ■

We are in a position to establish the consistency.

**Theorem 3.3** *Suppose that the condition of Theorem 2.4 hold for $\phi$ and that $\mathcal{H}$ satisfies $V_{\mathcal{A}_c} < \infty$ for $c = 1,\ldots,K$. Choose $\lambda_n$ such that $\lambda_n \to \infty$ and $\lambda_n\phi'(-\lambda_n(K-1))\sqrt{\frac{\ln n}{n}} \to 0$ as $n \to \infty$. Assume that, for any $n$ samples $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$, there exists an $\hat{\mathbf{f}}_n \in \mathcal{F}_{\lambda_n}$ such that*

$$\mathcal{E}_n(\hat{\mathbf{f}}_n) \le \inf_{\mathbf{f}\in\mathcal{F}_{\lambda_n}} \mathcal{E}_n(\mathbf{f}) + \varepsilon_n, \tag{13}$$

*where $\varepsilon_n$ is a sequence of positive numbers converging to zero. Then under Assumption 3.1, we have the consistency*

$$\lim_{n\to\infty} \mathbb{E}\mathcal{R}(C(\hat{\mathbf{f}}_n)) = \mathcal{R}^*.$$

**Proof.** Denote by $\mathbf{f}_{\lambda_n}$ an element of $\mathcal{F}_{\lambda_n}$ which minimizes $\mathcal{E}(\mathbf{f})$. By (13) we have

$$\begin{aligned} & \mathcal{E}(\hat{\mathbf{f}}_n) - \mathcal{E}(\mathbf{f}_{\lambda_n}) \\ = \;& \mathcal{E}(\hat{\mathbf{f}}_n) - \mathcal{E}_n(\hat{\mathbf{f}}_n) + \mathcal{E}_n(\hat{\mathbf{f}}_n) - \mathcal{E}_n(\mathbf{f}_{\lambda_n}) + \mathcal{E}_n(\mathbf{f}_{\lambda_n}) - \mathcal{E}(\mathbf{f}_{\lambda_n}) \\ \le \;& 2\sup_{\mathbf{f}\in\mathcal{F}_{\lambda_n}} |\mathcal{E}(\mathbf{f}) - \mathcal{E}_n(\mathbf{f})| + \varepsilon_n. \end{aligned}$$

Therefore,

$$\mathbb{E}\mathcal{E}(\hat{\mathbf{f}}_n) \le 2\mathbb{E}\sup_{\mathbf{f}\in\mathcal{F}_{\lambda_n}} |\mathcal{E}(\mathbf{f}) - \mathcal{E}_n(\mathbf{f})| + \mathcal{E}(\mathbf{f}_{\lambda_n}) + \varepsilon_n.$$

With our choice of $\lambda_n$, the first term on the right-hand side converges to zero by (9). Also $\mathcal{E}(\mathbf{f}_{\lambda_n}) \to \mathcal{E}^*$ by Assumption 3.1. Thus we have $\mathbb{E}\mathcal{E}(\hat{\mathbf{f}}_n) \to \mathcal{E}^*$. The proof is complete by Theorem 2.4 and the inequality

$$\mathbb{E}(\mathcal{E}(\hat{\mathbf{f}}_n) - \mathcal{E}^*)^{\frac{1}{\beta+1}} \le (\mathbb{E}\mathcal{E}(\hat{\mathbf{f}}_n) - \mathcal{E}^*)^{\frac{1}{\beta+1}}.$$

■

**Example 3.4** *The most important choice of $\phi$ in Theorem 3.3 is $\phi(x) = e^{-x}$. In this case, we thus choose $\lambda_n$ such that*

$$\lambda_n \to \infty \qquad \text{and} \qquad \lambda_n e^{\lambda_n (K-1)} \sqrt{\frac{\ln n}{n}} \to 0.$$

*If the set $\mathcal{H}$ has a finite VC dimension and, for any samples $\{(X_i, Y_i)\}_{i=1}^n$, (13) holds, then we have the consistency stated in Theorem 3.3.*

## Acknowledgments

The corresponding author is Chen. Sun's current address is Department of Information and Compute Science, Shengli College, China University of Petroleum. The authors thank Prof. P.L. Bartlett, Prof. D.X. Zhou and anonymous referees for their valuable suggestions which improve the paper significantly. The work is supported in part by NSF of China under grant 10571010.

## References

P. L. Bartlett, M. I. Jordan and J. D. Mcauliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

O. Bousquet, S. Boucheron and G. Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 9:323–375. 2005.

D. R. Chen, Q. Wu, Y. Ying and D. X. Zhou. Support vector machine soft margin classifier: error analysis. *J. Machine Learning Reseach*, 5: 1143–1175, 2004.

D. R. Chen and D. H. Xiang. The consistency of multicategory support vector machines. *Adv in Comput. Math.*, 24: 155–169, 2006.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer–Verlag, New York, 1996.

G. Lugosi. *Pattern classification and learning theory, Principles of Nonparametric Learning.* Springer, Wien, New York, pp 1–56, 2002.

G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting metheds. *Ann. Statist.*,32: 30–55, 2004.

C. McDiarmid. On the method of bounded differences. In *Surveys on Combinatorics*, 148–188, Combridge University Press, 1989.

A. Tewari and P. L. Bartlett: On the consistency of multiclass classification methods, In *Proc. 18th International Conference on Computational Learning Theory*, pages 143–157, 2005.

V. N. Vapnik. *Statistical Learning Theory.* John Wiley and sons, New York, 1998.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32: 56–134, 2004a.

T. Zhang. Statistical analysis of some multiclass large margin classification method. *Journal of Machine Learning Research*, 5: 1225–1251, 2004b.