Multicategory ψ-Learning

Yufeng LIU and Xiaotong SHEN

In binary classification, margin-based techniques usually deliver high performance. As a result, a multicategory problem is often treated as a sequence of binary classifications. In the absence of a dominating class, this treatment may be suboptimal and may yield poor performance, such as for support vector machines (SVMs). We propose a novel multicategory generalization of ψ-learning that treats all classes simultaneously. The new generalization eliminates this potential problem while at the same time retaining the desirable properties of its binary counterpart. We develop a statistical learning theory for the proposed methodology and obtain fast convergence rates for both linear and nonlinear learning examples. We demonstrate the operational characteristics of this method through a simulation. Our results indicate that the proposed methodology can deliver accurate class prediction and is more robust against extreme observations than its SVM counterpart.

KEY WORDS: Generalization error; Nonconvex minimization; Supervised learning; Support vectors.

1. INTRODUCTION

Classification has become increasingly important as a means for facilitating information extraction. Among binary classification techniques, significant developments have been seen in margin-based methodologies, including support vector machines (SVMs; Boser, Guyon, and Vapnik 1992; Cortes and Vapnik 1995), penalized logistic regression (PLR; Lin et al. 2000), import vector machines (IVMs; Zhu and Hastie 2005), and distance-weighted discrimination (DWD; Marron and Todd 2002).

Among many margin-based techniques, those that focus on estimating the decision boundary may yield higher performance than those that focus on conditional probabilities. This is because the former is an easier problem than the latter. For instance, binary SVM directly estimates the Bayes classifier $\operatorname{sign}(P(Y = +1|\mathbf{x}) - 1/2)$ rather than $P(Y = +1|\mathbf{x})$ itself, with input vector $\mathbf{x}$ and class label $Y \in \{\pm 1\}$, as shown by Lin (2002). However, this aspect of the methodology makes its generalization to the multicategory case highly nontrivial. One popular approach, known as "one-versus-rest," solves k binary problems through sequential training. As argued by Lee, Lin, and Wahba (2004), an approach of this sort performs poorly in the absence of a dominating class, because the conditional probability of each class is no greater than $1/2$.

Shen, Tseng, Zhang, and Wong (2003) proposed another margin-based technique, ψ-learning, that replaces the convex SVM loss function by a nonconvex ψ-loss function. These authors showed that more accurate class prediction can be achieved while retaining the margin interpretation. The present article generalizes binary ψ-learning to the multicategory case. Because ψ-learning, like SVM, does not directly yield $P(Y = +1|\mathbf{x})$, we need to take a new approach. To treat all classes simultaneously, we generalize the concept of margins and support vectors through multiple comparisons among different classes. Multicategory ψ-learning has the advantage of retaining the desired properties of its binary counterpart but not suffering from the aforementioned difficulty of one-versus-rest SVM with regard to the dominating class.

To provide insight into multicategory ψ-learning, we develop a statistical learning theory that quantifies the performance of multicategory ψ-learning with respect to the choice of tuning parameters, size of the training sample, and number of classes involved in classification. It also indicates that our multicategory ψ-learning directly estimates the true decision boundary regardless of the presence or absence of the dominating class.

Simulation experiments indicate that ψ-learning outperforms its counterpart SVM in generalization, as in the binary case. Moreover, multicategory ψ-learning is more robust against extreme instances that are wrongly classified than is its counterpart SVM. Interestingly, in linear learning problems, it exhibits some behavior that is similar to nonlinear learning problems with respect to the tuning parameter, but that differs from that of the binary case.

Section 2.1 motivates our approach. Section 2.2 describes our proposal for multicategory ψ-learning, and Section 2.3 briefly discusses computational issues. Section 3 studies the statistical properties of the proposed methodology and develops its statistical learning theory. Section 4 presents numerical examples, and Section 5 provides conclusions and discussions. The Appendix contains the lemmas and technical proofs.

2. METHODOLOGY

The primary goal of classification is to predict the class label Y for a given input vector $\mathbf{x} \in S$ through a classifier, where S is an input space. For k-class classification, a classifier partitions S into k disjoint and exhaustive regions $S_1, \ldots, S_k$, with S_j corresponding to class j. A good classifier is one that predicts class index Y for given $\mathbf{x}$ accurately, as measured by its accuracy of prediction.

Before proceeding, we let $\mathbf{x} \in S \subset \mathbb{R}^d$ be an input vector and y be an output (label) variable. We code y as $\{1, \ldots, k\}$ and define $\mathbf{f} = (f_1, \ldots, f_k)$ as a decision function vector. Here f_j, mapping from S to $\mathbb{R}$, represents class j; $j = 1, \ldots, k$. We use classifier $\operatorname{argmax}_{j=1,\ldots,k} f_j(\mathbf{x})$, induced by $\mathbf{f}$, to assign a label to any input vector $\mathbf{x} \in S$. In other words, $\mathbf{x} \in S$ is assigned to a class with the highest value of $f_j(\mathbf{x})$, which indicates the strength of evidence that $\mathbf{x}$ belongs to class j. A classifier is trained through a training sample $\{(\mathbf{x}_i, y_i); i = 1, \ldots, n\}$, independently and identically distributed according to an unknown probability distribution $P(\mathbf{x}, y)$. Throughout the article, we use $\mathbf{X}$ and Y to denote random variables and $\mathbf{x}$ and y to represent corresponding observations.

The generalization error (GE) quantifies the accuracy of generalization and is defined as $\operatorname{Err}(\mathbf{f}) = P[Y \neq \operatorname{argmax}_j f_j(\mathbf{X})]$,

Yufeng Liu is Assistant Professor, Department of Statistics and Operations Research, Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, NC 27599 (E-mail: *yfliu@email.unc.edu*). Xiaotong Shen is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (E-mail: *xshen@stat.umn.edu*). This research was supported in part by National Science Foundation grants IIS-0328802 and DMS-00-72635. The authors thank the editor, the associate editor, two anonymous referees, and Professor George Fisherman for their helpful comments and suggestions.

Journal of the American Statistical Association
June 2006, Vol. 101, No. 474, Theory and Methods
DOI 10.1198/016214505000000781

the probability of misclassifying a new input vector $\mathbf{X}$. To simplify the expression, we introduce $\mathbf{g}(f(\mathbf{x}), y) = (f_y(\mathbf{x}) - f_1(\mathbf{x}), \ldots, f_y(\mathbf{x}) - f_{y-1}(\mathbf{x}), f_y(\mathbf{x}) - f_{y+1}(\mathbf{x}), \ldots, f_y(\mathbf{x}) - f_k(\mathbf{x}))$, which performs multiple comparisons of class y versus the rest of the classes. Vector $\mathbf{g}(f(\mathbf{x}), y)$ describes the unique feature of a multicategory problem, that is, directly related to the generalized margins introduced shortly. Furthermore, for $\mathbf{u} = (u_1, \ldots, u_{k-1})$, we define the multivariate sign function, $\text{sign}(\mathbf{u}) = 1$ if $\mathbf{u}_{\min} = \min(u_1, \ldots, u_{k-1}) > 0$ and -1 if $\mathbf{u}_{\min} \le 0$. With $\text{sign}(\cdot)$ and $\mathbf{g}(f(\mathbf{x}), y)$ in place, $\mathbf{f}$ indicates correct classification for any given instance $(\mathbf{x}, y)$ if $\mathbf{g}(f(\mathbf{x}), y) > \mathbf{0}_{k-1}$, where $\mathbf{0}_{k-1}$ is a $(k-1)$-dimensional vector of 0. Consequently, the GE reduces to $\text{Err}(\mathbf{f}) = \frac{1}{2}E[1 - \text{sign}(\mathbf{g}(\mathbf{X}, Y))]$, with the empirical generalization error (EGE) $(2n)^{-1}\sum_{i=1}^{n}(1 - \text{sign}(\mathbf{g}(f(\mathbf{x}_i), y_i)))$.

For motivation, we first discuss our setting in the binary case, then generalize it to the multicategory case. In particular, we review binary ψ-learning with the usual coding $\{-1, 1\}$, and then derive it through coding $\{1, 2\}$.

2.1 Motivation

With $y \in \{\pm 1\}$, a margin-based classifier estimates a single function f and uses $\text{sign}(f)$ as the classification rule. Within the regularization framework, it solves $\text{argmin}_f J(f) + C\sum_{i=1}^{n} l(y_i f(\mathbf{x}_i))$, where $J(f)$, a regularization term, controls the complexity of f, a loss function l measures the data fit, and $C > 0$ is a tuning parameter balancing the two terms. For example, SVM uses the hinge loss with $l(u) = [1 - u]_+$, where $[v]_+ = v$ if $v \ge 0$, and 0 otherwise; PLR and IVM adopt the logistic loss $l(u) = \log(1 + e^{-u})$; and the ψ-loss can be any nonincreasing function satisfying $R \ge \psi(u) > 0$ if $u \in (0, \tau)$ and $\psi(u) = 1 - \text{sign}(u)$ otherwise, where $\tau \in (0, 1]$ and $R > 0$. For simplicity, we discuss the linear case in which $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$, $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ represents a d-dimensional hyperplane. In this case $J(f) = \frac{1}{2}\|\mathbf{w}\|^2$ is defined by the geometric margin $\frac{2}{\|\mathbf{w}\|}$, the vertical Euclidean distance between hyperplanes $f = \pm 1$. Here $y_i f(\mathbf{x}_i)$ is the functional margin of instance $(\mathbf{x}_i, y_i)$.

For linear binary ψ-learning with coding $\{1, 2\}$, we now derive a parallel formulation using the argmax rule, by noting that $\mathbf{x}$ is classified as class 2 if $f_2(\mathbf{x}) > f_1(\mathbf{x})$ and as class 1 otherwise, where $f_j(\mathbf{x}) = \mathbf{w}_j^T\mathbf{x} + b_j$, $j = 1, 2$. Evidently, this rule of classification depends only on $\text{sign}((f_2 - f_1)(\mathbf{x}))$. To eliminate redundancy in (f_1, f_2), we invoke a sum-to-0 constraint $f_1 + f_2 = 0$. This type of constraint was previously used by Guermeur (2002) and Lee et al. (2004) in two different SVM formulations. Under this constraint, $\|\mathbf{w}_1\| = \|\mathbf{w}_2\|$. Binary ψ-learning then solves

$$\min_{b_1, b_2, \mathbf{w}_1, \mathbf{w}_2} \left(\frac{1}{2}\sum_{j=1}^{2}\|\mathbf{w}_j\|^2 + C\sum_{i=1}^{n}\psi\big(\mathbf{g}\big(\mathbf{f}(\mathbf{x}_i), y_i\big)\big)\right) \quad \text{subject to} \quad \sum_{j=1}^{2} f_j(\mathbf{x}) = 0 \qquad \forall \mathbf{x} \in S, \tag{1}$$

where $\mathbf{g}(\mathbf{f}(\mathbf{x}_i), y_i) = f_{y_i}(\mathbf{x}_i) - f_{3-y_i}(\mathbf{x}_i)$.

With coding $\{1, 2\}$, instances from classes 1 and 2 that lie in half-spaces $\{\mathbf{x} : \mathbf{g}(\mathbf{f}(\mathbf{x}), 2) \ge -1\}$ and $\{\mathbf{x} : \mathbf{g}(\mathbf{f}(\mathbf{x}), 2) \le 1\}$ are defined as "support vectors." In the separable case, support vectors are instances on hyperplanes $\mathbf{g}(\mathbf{f}(\mathbf{x}), 2) = \pm 1$. Furthermore, the functional margin of $(\mathbf{x}_i, y_i)$ can be defined as $\mathbf{g}(\mathbf{f}(\mathbf{x}_i), y_i)$, indicating the correctness and strength of classification of $\mathbf{x}_i$ by $\mathbf{f}$.

2.2 Multicategory ψ-Learning

As suggested by Shen et al. (2003), the role of a binary ψ-function is twofold. First, it eliminates the scaling problem of the sign function that is scale-invariant. Second, with a positive penalty defined by the positive value of $\psi(u)$ for $u \in (0, \tau)$, it pushes correctly classified instances away from the boundary. We note that $1 - \text{sign}$ as a loss is numerically undesirable, because the solution $\mathbf{f}$ is approximately 0 under regularization.

Using coding $\{1, \ldots, k\}$, we define multivariate ψ-functions on $k - 1$ arguments as follows:

$$\begin{aligned} R \ge \psi(\mathbf{u}) > 0 \qquad & \text{if } \mathbf{u}_{\min} \in (0, \tau), \\ \psi(\mathbf{u}) = 1 - \text{sign}(\mathbf{u}) \qquad & \text{otherwise,} \end{aligned} \tag{2}$$

where $0 < \tau \le 1$ and $0 < R \le 2$ are some constants and $\psi(\mathbf{u})$ is nonincreasing in $\mathbf{u}_{\min}$. We note that this multivariate version preserves the desired properties of its univariate counterpart. In particular, the multivariate ψ assigns a positive penalty to any instance with $\min(\mathbf{g}(\mathbf{f}(\mathbf{x}_i), y_i)) \in (0, \tau)$ to eliminate the scaling problem. To use our computational strategy based on a difference convex (dc) decomposition, we use a specific ψ in implementation,

$$\psi(\mathbf{u}) = \begin{cases} 0 & \text{if } \mathbf{u}_{\min} \ge 1 \\ 2 & \text{if } \mathbf{u}_{\min} < 0 \\ 2(1 - \mathbf{u}_{\min}) & \text{if } 0 \le \mathbf{u}_{\min} < 1. \end{cases} \tag{3}$$

Figure 1 shows a plot of this ψ function for $k = 3$.

Linear multicategory ψ-learning solves $\min_{\mathbf{b}, \mathbf{w}}(\frac{1}{2} \times \sum_{j=1}^{k}\|\mathbf{w}_j\|^2 + C\sum_{i=1}^{n}\psi(\mathbf{g}(\mathbf{x}_i, y_i)))$ subject to $\sum_{j=1}^{k} f_j(\mathbf{x}) = 0$ for $\forall \mathbf{x} \in S$, where $\mathbf{w} = \text{vec}(\mathbf{w}_1, \ldots, \mathbf{w}_k)$ is a kd-dimensional vector with its $(d(i_2 - 1) + i_1)$th element $\mathbf{w}_{i_2}(i_1)$ and $\mathbf{b} = (b_1, \ldots, b_k)^T \in \mathbb{R}^k$. By theorem 2.1 of Liu, Shen, and Doss (2005), minimization with the sum-to-0 constraint for all $\mathbf{x} \in S$ is equivalent to that with the constraint for n training inputs $\{\mathbf{x}_i; i = 1, \ldots, n\}$ only. That is, the infinite constraint $\sum_{j=1}^{k} f_j(\mathbf{x}) = 0$ for $\forall \mathbf{x} \in S$ can be reduced to $\sum_{j=1}^{k} b_j \mathbf{1}_n + \sum_{j=1}^{k} X\mathbf{w}_j = 0$, where $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ is the design matrix

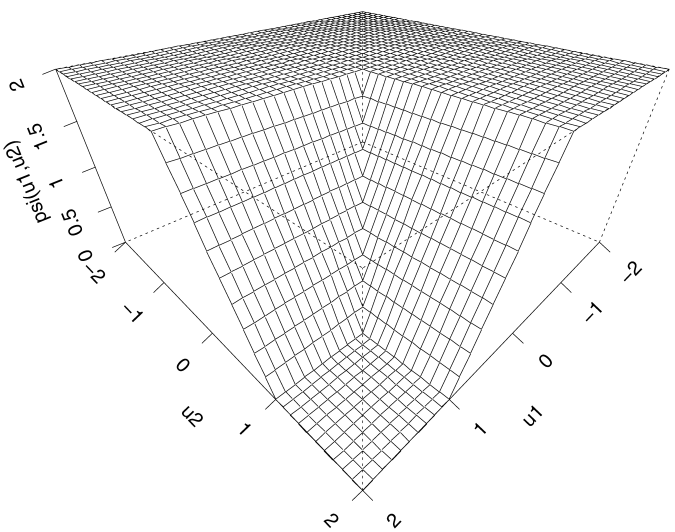

Figure 1. Perspective Plot of the Three-Class ψ Function Defined in (3).

and $\mathbf{1}_n$ is an n-dimensional vector of 1's. This yields linear multicategory ψ-learning,

$$\min_{\mathbf{b},\mathbf{w}}\left(\frac{1}{2}\sum_{j=1}^{k}\|\mathbf{w}_j\|^2 + C\sum_{i=1}^{n}\psi(\mathbf{g}(\mathbf{x}_i, y_i))\right) \quad \text{subject to} \quad \sum_{j=1}^{k} b_j\mathbf{1}_n + X\sum_{j=1}^{k}\mathbf{w}_j = 0, \tag{4}$$

where the value of C ($C > 0$) in (4) reflects relative importance between the geometric margin and the EGE.

In the present context, we define the generalized functional margin of an instance $(\mathbf{x}_i, y_i)$ as $\min(\mathbf{g}(\mathbf{x}_i, y_i))$ and the generalized geometric margin as $\gamma = \min_{1\le j_1<j_2\le k}\gamma_{j_1j_2}$ with $\gamma_{j_1j_2} = \frac{2}{\|\mathbf{w}_{j_1}-\mathbf{w}_{j_2}\|}$ as the vertical Euclidean distance between hyperplanes $f_{j_1} - f_{j_2} = \pm 1$. Here $\gamma_{j_1j_2}$ measures separation between classes i and j; see Figure 2 for an illustration of the role of γ. When $k = 2$, (4) reduces to the binary case of Shen et al. (2003). As a technical remark, we note that (4) uses $\sum_{j=1}^{k}\|\mathbf{w}_j\|^2$ rather than $\max_{1\le j_1<j_2\le k}\|\mathbf{w}_{j_1} - \mathbf{w}_{j_2}\|^2$ in minimization. This is because $\sum_{j=1}^{k}\|\mathbf{w}_j\|^2$ plays a similar role as $\max_{1\le j_1<j_2\le k}\|\mathbf{w}_{j_1} - \mathbf{w}_{j_2}\|^2$ and is easier to implement.

Kernel-based learning can be achieved through a proper kernel $K(\cdot,\cdot)$, mapping from $S \times S$ to $\mathbb{R}$. The kernel is required to satisfy Mercer's condition (Mercer 1909) which ensures the kernel matrix $\mathbf{K}$ to be positive definite, where $\mathbf{K}$ is an $n \times n$ matrix with its i_1i_2th element $K(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$. Then each f_j can be represented as $h_j(\mathbf{x}) + b_j$ with $h_j = \sum_{i=1}^{n} v_{ji}K(\mathbf{x}_i, \mathbf{x})$ by the theory of reproducing kernel Hilbert spaces (cf. Wahba 1998). The kernel-based multicategory ψ-learning then solves

$$\min_{\mathbf{b},\mathbf{v}}\left(\frac{1}{2}\sum_{j=1}^{k}\|h_j\|_{H_K}^2 + C\sum_{i=1}^{n}\psi(\mathbf{g}(\mathbf{x}_i, y_i))\right) \quad \text{subject to} \quad \sum_{j=1}^{k} b_j\mathbf{1}_n + \mathbf{K}\sum_{j=1}^{k}\mathbf{v}_j = 0, \tag{5}$$

where $\mathbf{v}_j = (v_{j1}, \ldots, v_{jn})^T$ and $\mathbf{v} = \mathrm{vec}(\mathbf{v}_1, \ldots, \mathbf{v}_n)$. Using the reproducing kernel property, $\|h_j\|_{H_K}^2$ can be written as $\mathbf{v}_j^T\mathbf{K}\mathbf{v}_j$.

The concept of support vectors can be also extended to multicategory problems. In the separable case, the instances on the boundaries of polyhedrons D_j are the support vectors, where polyhedron D_j is a collection of solutions of a finite system of linear inequalities defined by $\min_j(\mathbf{g}(\mathbf{x}, j)) \ge 1$. In the nonseparable case, the instances belonging to class j that do not fall the inside of D_j are the support vectors.

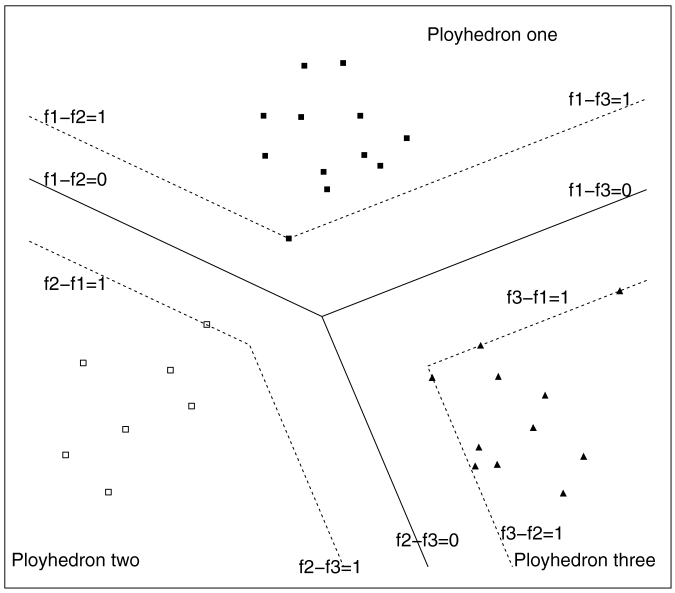

Figure 2. Illustration of the Concept of Margins and Support Vectors in a Three-Class Separable Example. The instances for classes 1–3 fall into the polyhedrons D_j; $j = 1, 2, 3$, where D_1 is $\{\boldsymbol{x}: f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}) \ge 1, f_1(\boldsymbol{x}) - f_3(\boldsymbol{x}) \ge 1\}$, D_2 is $\{\boldsymbol{x}: f_2(\boldsymbol{x}) - f_1(\boldsymbol{x}) \ge 1, f_2(\boldsymbol{x}) - f_3(\boldsymbol{x}) \ge 1\}$, and D_3 is $\{\boldsymbol{x}: f_3(\boldsymbol{x}) - f_1(\boldsymbol{x}) \ge 1, f_3(\boldsymbol{x}) - f_2(\boldsymbol{x}) \ge 1\}$. The generalized geometric margin γ defined as $\min\{\gamma_{12}, \gamma_{13}, \gamma_{23}\}$ is maximized to obtain the decision boundary. There are five support vectors on the boundaries of the three polyhedrons. The five support vectors include one from class 1, one from class 2, and three from class 3.

2.3 Computational Development of ψ-Learning

To treat the nonconvex minimization involved in (4) and (5), we use the state-of-art technology in global optimization—the difference convex algorithm (DCA) of An and Tao (1997). (For a detailed algorithm see Liu, Shen, and Doss 2005.) Other recent computational developments of binary ψ-learning have been given by Liu and Wu (2005) and Liu, Shen, and Wong (2005).

The key to efficient computation is a dc decomposition of $\psi = \psi_1 + \psi_2$, where $\psi_1(\mathbf{u}) = 0$ if $\mathbf{u}_{\min} \ge 1$ and $2(1 - \mathbf{u}_{\min})$ otherwise, and $\psi_2(\mathbf{u}) = 0$ if $\mathbf{u}_{\min} \ge 0$ and $2\mathbf{u}_{\min}$ otherwise. Here ψ_1 can be viewed as a multivariate generalization of the univariate hinge loss. This dc decomposition connects the ψ-loss to the hinge loss ψ_1 of SVM. In fact, the multivariate ψ mimics the GE defined by $1 -$ sign, whereas the generalized hinge loss ψ_1 is a convex upper envelope of $1 -$ sign. With this dc decomposition, ψ corrects the bias introduced by the imposed convexity of ψ_1 and is expected to yield higher generalization accuracy.

3. STATISTICAL LEARNING THEORY

In the literature there has been considerable interest in generalization accuracy of margin-based classifiers. For the binary case, Lin (2000) investigated rates of convergence of SVM with a spline kernel. Bartlett, Jordon, and McAuliffe (2003) studied the rates of convergence for certain convex margin losses. Shen et al. (2003) derived a learning theory for ψ-learning, and Zhang (2004a) obtained consistency for general convex margin-based losses. For the multicategory case, Zhang (2004b) has recently studied the consistency of several large margin classifiers using convex losses. To our knowledge, no results are available for rates of convergence in the multicategory case. In this section we quantify the generalization error rates of the proposed multicategory ψ-learning, as measured by the Bayesian regret.

3.1 Statistical Properties

The generalization performance of a classifier defined by $\mathbf{f}$ is measured by the Bayesian regret $e(\mathbf{f}, \bar{\mathbf{f}}) = \mathrm{Err}(\mathbf{f}) - \mathrm{Err}(\bar{\mathbf{f}}) \ge 0$, the difference between the actual performance and the ideal performance. Here $\bar{\mathbf{f}}$ is the Bayes rule, yielding the ideal performance assuming that the true distribution of $(\mathbf{X}, Y)$ would

have been known in advance, obtained by minimizing $\mathrm{Err}(\mathbf{f}) = \frac{1}{2}E[1 - \mathrm{sign}(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y))]$ with respect to all $\mathbf{f}$, with $\mathbf{g}(\mathbf{f}(\mathbf{x}), j) = \{f_j(\mathbf{x}) - f_l(\mathbf{x}), l \neq j\}$. Note that the Bayes rule is not unique, because any $\bar{\mathbf{f}}$, satisfying $\mathrm{argmax}_j \bar{f}_j(\mathbf{x}) = \mathrm{argmax}_j P_j(\mathbf{x})$ with $P_j(\mathbf{x}) = P(Y = j|\mathbf{x})$ yields the minimum. Without loss of generality, in what follows we use a specific $\bar{\mathbf{f}} = (\bar{f}_1, \ldots, \bar{f}_k)$ with $\bar{f}_j(\mathbf{x}) = \frac{k-1}{k} I(\mathrm{sign}(P_j(\mathbf{x}) - P_l(\mathbf{x}), l \neq j) = 1) - \frac{1}{k} I(\mathrm{sign}(P_j(\mathbf{x}) - P_l(\mathbf{x}), l \neq j) \neq 1)$, that is, $\bar{f}_l(\mathbf{x}) = \frac{k-1}{k}$ if $l = \mathrm{argmax} P_j(\mathbf{x})$, and $-\frac{1}{k}$ otherwise.

Theorem 1 gives expressions of the Bayesian regret, which is critical for establishing our learning theory.

Theorem 1. For any decision function vector $\mathbf{f}$,

$$e(\mathbf{f}, \bar{\mathbf{f}}) = \frac{1}{2}E\Bigg[\sum_{j=1}^{k} P_j(\mathbf{X})\big(\mathrm{sign}\big(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), j)\big) - \mathrm{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), j)\big)\big)\Bigg] \tag{6}$$

$$= E\big[\max_j P_j(\mathbf{X}) - P_{\mathrm{argmax}_j f_j}(\mathbf{X})\big] \geq 0$$

$$= E\Bigg[\sum_{j \neq l} |P_l(\mathbf{X}) - P_j(\mathbf{X})| \times I\big(\mathrm{sign}\big(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), l)\big) = 1, \mathrm{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), j)\big) = 1\big)\Bigg], \tag{7}$$

where $\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{x}), j) = \{\bar{f}_j(\mathbf{x}) - \bar{f}_l(\mathbf{x}), l \neq j\}$.

Equation (6) in Theorem 1 expresses $e(\mathbf{f}, \bar{\mathbf{f}})$ in terms of a weighted sum of the individual misclassification error, weighted by the conditional probability $P_j(\mathbf{X})$. Equation (7) gives an expression of $e(\mathbf{f}, \bar{\mathbf{f}})$ in misclassification resulting from $\binom{k}{2}$ multiple comparisons.

Equation (7) suggests that a multicategory problem differs dramatically from its binary counterpart. For a binary problem, (7) reduces to $e(f_2, \bar{f}_2) = E|P_2(\mathbf{X}) - 1/2||\,\mathrm{sign}(f_2(\mathbf{X})) - \mathrm{sign}(\bar{f}_2(\mathbf{X}))|$ because $P_2(\mathbf{x}) - P_1(\mathbf{x}) = 2(P_2(\mathbf{x}) - 1/2)$. This means that a comparison between $P_1(\mathbf{x})$ and $P_2(\mathbf{x})$ in the binary case is equivalent to examining whether $P_2(\mathbf{x})$ exceeds $1/2$. But for a multicategory problem, this no longer holds because multiple pairwise comparisons are needed to determine the argmax. In fact, there may not exist a dominating class, that is, $\max P_l(\mathbf{x}) < 1/2$ for some $\mathbf{x} \in S$. Therefore, k comparisons of $P_j(\mathbf{x})$ with $1/2$ may not be sufficient to determine the correct classification rule. Indeed, the issue of the existence of a dominating class is important in the multicategory case but not in the binary case.

The ultimate goal of classification is to minimize $E[1 - \mathrm{sign}(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y))]$. To avoid the scale-invariant problem of the sign function, we apply ψ-loss here as a surrogate loss that minimizes $E[\psi(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y))]$. The following theorem says that a ψ-loss yields the same Bayes rule as the $1 -$ sign loss. Thus consistency of multicategory ψ-learning can be established.

Theorem 2. The Bayes decision vector $\bar{\mathbf{f}}$ satisfies $\bar{\mathbf{g}}_{\min}(\bar{\mathbf{f}}(\mathbf{x}), \mathrm{argmax}_{j=1,\ldots,k} P_j(\mathbf{x})) = 1$, where $\bar{\mathbf{g}}_{\min}$ is the minimum of the $k - 1$ elements of vector $\bar{\mathbf{g}}$. For any ψ satisfying (2), $\bar{\mathbf{f}}$ minimizes $E[\psi(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y))]$ and $E[1 - \mathrm{sign}(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y))]$ in the sense that $E[\psi(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y))] \geq E[\psi(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), Y))] = E[1 - \mathrm{sign}(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), Y))] \leq E[1 - \mathrm{sign}(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y))]$ for any $\mathbf{f}$. Furthermore, the minimizers for $E[\psi(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y))]$ and $E[1 - \mathrm{sign}(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y))]$ are not unique; for example, $c\bar{\mathbf{f}}$ is also a minimizer for both quantities for any $c \geq 1$.

Theorem 2 says that ψ-learning estimates the Bayes classifier defined by $\bar{\mathbf{f}}$ as opposed to the conditional probabilities $(P_1(\mathbf{x}), \ldots, P_k(\mathbf{x}))$ and that it plays the same role as $1 -$ sign. Furthermore, the optimal performance of $\bar{\mathbf{f}}$ with $\bar{\mathbf{g}}_{\min}(\bar{\mathbf{f}}(\mathbf{x}), \mathrm{argmax}_j P_j(\mathbf{x})) = 1$ is realized through the ψ-loss function, although it differs from $1 -$ sign.

3.2 Statistical Learning Theory

Let $\mathcal{F}$ be a function class of candidate function vectors that is allowed to depend on n. Note that the Bayes decision function $\bar{\mathbf{f}}$ is not required to belong to $\mathcal{F}$. For any function vector $\mathbf{f} \in \mathcal{F}$, classification is performed by partitioning S into k disjoint sets $(G_{f_1}, \ldots, G_{f_k}) = (\{\mathbf{x} : \mathrm{sign}(\mathbf{g}(\mathbf{f}(\mathbf{x}), 1) = 1\}, \ldots, \{\mathbf{x} : \mathrm{sign}(\mathbf{g}(\mathbf{f}(\mathbf{x}), k) = 1\})$.

In this section we generalize the learning theory of Shen et al. (2003) to the multicategory case. Our learning theory quantifies the magnitude of $e(\mathbf{f}, \bar{\mathbf{f}})$ as a function of n, k, the tuning parameter C, and the complexity of a class of candidate classification partitions $\mathcal{G}(\mathcal{F}) = \{(G_{f_1}, \ldots, G_{f_k}); \mathbf{f} \in \mathcal{F}\}$ induced by $\mathcal{F}$.

Denote the approximation error $e_\psi(\mathbf{f}, \bar{\mathbf{f}}) = \frac{1}{2}(E\psi(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y)) - E\psi(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), Y)))$, which measures the degree of approximation of $\mathcal{G}(\mathcal{F})$ to $(G_{\bar{f}_1}, \ldots, G_{\bar{f}_k})$. Let $J_0 = \max(J(\mathbf{f}_0), 1)$. The following technical assumptions are made:

Assumption A (Approximation error). For some positive sequence $s_n \to 0$ as $n \to \infty$, there exists $\mathbf{f}_0 \in \mathcal{F}$ such that $e_\psi(\mathbf{f}_0, \bar{\mathbf{f}}) \leq s_n$. Equivalently, $\inf_{\{\mathbf{f} \in \mathcal{F}\}} e_\psi(\mathbf{f}, \bar{\mathbf{f}}) \leq s_n$. Similar to $\mathcal{F}$, $\mathbf{f}_0$ may depend on n.

Assumption B (Boundary behavior). There exist some constants $0 < \alpha \leq +\infty$ and $c_1 > 0$ such that $P(\mathbf{X} \in S : (\max P_l(\mathbf{X}) - P_{j \neq \mathrm{argmax} P_l(\mathbf{X})}(\mathbf{X})) < 2\delta) \leq c_1 \delta^\alpha$ for any small $\delta \geq 0$.

Assumption B describes the behavior of the conditional probabilities, P_j's, near the decision boundary $\{\mathbf{x} \in S : \max P_l(\mathbf{x}) = P_j(\mathbf{x})$; for some $l \neq j \in \{1, 2, \ldots, k\}\}$. It is equivalent to $P(\mathbf{X} \in S : (\max P_l(\mathbf{X}) - \text{second } \max P_l(\mathbf{X})) < 2\delta) \leq c_1 \delta^\alpha$ by the fact that $\{\mathbf{X} : \max P_l(\mathbf{X}) - P_{j \neq \mathrm{argmax} P_l(\mathbf{X})}(\mathbf{X})) < 2\delta\} \subset \{\mathbf{X} : \max P_l(\mathbf{X}) - \text{second } \max P_l(\mathbf{X}) < 2\delta\}$.

To specify Assumption C, we define the metric entropy for partitions. For a class of partitions $\boldsymbol{\mathcal{B}} = \{(B_1, \ldots, B_k); B_j \cap B_l = \emptyset \ \forall j \neq l, \bigcup_{1 \leq j \leq k} B_j = S\}$ and any $\epsilon > 0$, call $\{(G_{j1}^v, G_{j1}^u), \ldots, (G_{jm}^v, G_{jm}^u)\}, j = 1, \ldots, k$, an ϵ-bracketing set of $\boldsymbol{\mathcal{B}}$ if for any $(G_1, \ldots, G_k) \in \boldsymbol{\mathcal{B}}$ there exists an h such that $G_{jh}^v \subset G_j \subset G_{jh}^u$ and

$$\max_{1 \leq h \leq m} \max_{1 \leq j \leq k} P(G_{jh}^u \Delta G_{jh}^v) \leq \epsilon, \tag{8}$$

where $G_{jh}^u \Delta G_{jh}^v$ is the set difference between G_{jh}^u and G_{jh}^v. The metric entropy $H_B(\epsilon, \boldsymbol{\mathcal{B}})$ of $\boldsymbol{\mathcal{B}}$ with bracketing is then defined as logarithm of the cardinality of ϵ-bracketing set of $\boldsymbol{\mathcal{B}}$ of the smallest size.

Let $\mathcal{F}(\ell) = \{\mathbf{f} \in \mathcal{F}, J(\mathbf{f}) \leq \ell\} \subset \mathcal{F}$ and $\mathcal{G}(\ell) = \{(G_{f_1}, \ldots, G_{f_k}); \mathbf{f} \in \mathcal{F}(\ell)\} \subset \mathcal{G}(\mathcal{F})$. Then $\mathcal{G}(\ell)$ is the set of classification

partitions under regularization $J(\mathbf{f}) \le \ell$. For instance, $J(\mathbf{f})$ is $\frac{1}{2}\sum_j \|\mathbf{w}_j\|^2$ in (4) or is $\frac{1}{2}\sum_j \|h_j\|^2_{H_K}$ in (5). To measure the complexity of $\mathcal{G}(\ell)$ via the metric entropy, the following assumptions are made.

Assumption C (Metric entropy for partitions). For some positive constants c_i, $i = 2, 3, 4$, there exists some $\epsilon_n > 0$ such that

$$\sup_{\ell \ge 2} \phi(\epsilon_n, \ell) \le c_2 n^{1/2}, \tag{9}$$

where $\phi(\epsilon_n, \ell) = \int_{c_4 L}^{c_3^{1/2} L^{\alpha/2(\alpha+1)}} H_B^{1/2}(u^2/4, \mathcal{G}(\ell))\, du/L$ and $L = L(\epsilon_n, C, \ell) = \min(\epsilon_n^2 + (Cn)^{-1}(\ell/2 - 1)J_0, 1)$.

Assumption D (ψ-function). The ψ-function satisfies (2).

As a technical remark, we note that to simplify the function entropy calculation in Assumption C required in Theorem 4, we may impose an additional condition on the ψ-function. For instance, we may restrict the ψ-loss functions in (2) to satisfy a multivariate Lipschitz condition,

$$|\psi(\mathbf{u}^*) - \psi(\mathbf{u}^{**})| \le D|\mathbf{u}^*_{\min} - \mathbf{u}^{**}_{\min}|, \tag{10}$$

where $D > 0$ is a constant. Condition (10) is satisfied by the specific ψ function in (3), with $D = 2$. This aspect is illustrated in Example 2. However, (10) is irrelevant to the set entropy in Assumption C required in Theorem 3; see Example 1.

Theorem 3. Suppose that Assumptions A–D are met. Then for any classifier of ψ-learning $\operatorname{argmax}(\hat{\mathbf{f}})$, there exists a constant $c_5 > 0$ such that

$$P\big(e(\hat{\mathbf{f}}, \bar{\mathbf{f}}) \ge \delta_n^2\big) \le 3.5 \exp\big(-c_5 n (nC)^{-(\alpha+2)/(\alpha+1)} J_0^{(\alpha+2)/(\alpha+1)}\big),$$

provided that $Cn \ge 2\delta_n^{-2} J_0$, where $\delta_n^2 = \min(\max(\epsilon_n^2, 2s_n), 1)$.

Corollary 1. Under the assumptions of Theorem 3,

$$|e(\hat{\mathbf{f}}, \bar{\mathbf{f}})| = O_p(\delta_n^2), \qquad E|e(\hat{\mathbf{f}}, \bar{\mathbf{f}})| = O(\delta_n^2),$$

provided that $n^{-1/(\alpha+1)}(C^{-1}J_0)^{(\alpha+2)/(\alpha+1)}$ is bounded away from 0.

To obtain the error rate δ_n^2 in Theorem 3, we need to compute the metric entropy for $\mathcal{G}(\ell)$. Computing the metric entropy for partitions, may not be easy because $\mathcal{G}(\ell)$ is induced by the class of functions $\mathcal{F}(\ell)$. Moreover, it is also of interest to establish an upper bound of $e(\hat{\mathbf{f}}, \bar{\mathbf{f}})$ using the corresponding function entropy as opposed to set entropy. In what follows, we develop such results in Theorem 4.

To proceed, we define the L_2-metric entropy with bracketing for $\mathcal{F}$ as follows. For any $\epsilon > 0$, call $\{(g_1^v, g_1^u), \ldots, (g_m^v, g_m^u)\}$ an ϵ-bracketing function if for any $g \in \mathcal{F}$ there is an h such that $g_h^v \le g \le g_h^u$ and $\max_{1 \le h \le m} \|g_h^u - g_h^l\|_2 \le \epsilon$, where $\|\cdot\|_2$ is the usual L_2-norm, defined as $\|g\|_2^2 = \int g^2\, dP$. Then the L_2-metric entropy of $\mathcal{F}$ with bracketing $H_B(\epsilon, \mathcal{F})$ is defined as logarithm of the cardinality of the ϵ-bracketing of the smallest size. Now define a new function set $\mathcal{F}^\psi(\ell) = \{\psi(\mathbf{g}(\mathbf{f}(\mathbf{x}), y)) - \psi(\mathbf{g}_0(\mathbf{f}_0(\mathbf{x}), y)) : \mathbf{f} \in \mathcal{F}(\ell)\}$ and $\phi^*(\epsilon_n^*, \ell) = \int_{c_4 L^*}^{c_3^{1/2} L^{*\alpha/2(\alpha+1)}} H_B^{1/2}(u, \mathcal{F}^\psi(\ell))\, du/L^*$ with $L^* = \min(\epsilon_n^{*2} + (Cn)^{-1}(\ell/2 - 1)J_0, 1)$.

Theorem 4. Suppose that Assumptions A–D are met with $\phi^*(\epsilon_n^*, \ell)$ replacing $\phi(\epsilon_n, \ell)$ in Assumption C. Then for any classifier of ψ-learning $\operatorname{argmax}(\hat{\mathbf{f}})$, there exists a constant $c_5 > 0$ such that

$$\begin{aligned} &P\big(e(\hat{\mathbf{f}}, \bar{\mathbf{f}}) \ge \delta_n^{*2}\big) \\ &\quad \le P\big(e_\psi(\hat{\mathbf{f}}, \bar{\mathbf{f}}) \ge \delta_n^{*2}\big) \\ &\quad \le 3.5 \exp\big(-c_5 n (nC)^{-(\alpha+2)/(\alpha+1)} J_0^{(\alpha+2)/(\alpha+1)}\big), \end{aligned}$$

provided that $Cn \ge 2\delta_n^{*-2} J_0$, where $\delta_n^{*2} = \min(\max(\epsilon_n^{*2}, 2s_n), 1)$.

Corollary 2. Under the assumptions of Theorem 4,

$$|e_\psi(\hat{\mathbf{f}}, \bar{\mathbf{f}})| = O_p(\delta_n^{*2}), \qquad E|e_\psi(\hat{\mathbf{f}}, \bar{\mathbf{f}})| = O(\delta_n^{*2}),$$

provided that $n^{-1/(\alpha+1)}(C^{-1}J_0)^{(\alpha+2)/(\alpha+1)}$ is bounded away from 0.

Note that $e_\psi(\hat{\mathbf{f}}, \bar{\mathbf{f}}) \ge e(\hat{\mathbf{f}}, \bar{\mathbf{f}})$. The rate δ_n^{*2} obtained from Theorem 4 using the metric entropy for functions yields an upper bound of $e(\hat{\mathbf{f}}, \bar{\mathbf{f}})$, and thus $e(\hat{\mathbf{f}}, \bar{\mathbf{f}}) = O(\min(\delta_n^2, \delta_n^{*2}))$ with probability tending to 1 by Corollaries 1 and 2. In applications, one may calculate either δ_n^2 or δ_n^{*2}, depending on which entropy is easier to compute.

Theorems 3 and 4 reveal distinct characteristics of multicategory problems, although they cover the binary case. First, a multicategory problem has a higher level of complexity generally, and hence the number of classes k may have an impact on the performance. In fact, Theorems 3 and 4 permit one to study dependency of $e(\hat{\mathbf{f}}, \bar{\mathbf{f}})$ on k and n simultaneously; see Examples 1 and 2. Second, some properties of binary linear learning no longer hold in the multicategory case when $k > 2$. For instance, the decision boundaries generated by linear learning with $k > 2$ can be piecewise linear hyperplanes.

3.3 Illustrative Examples

To illustrate our learning theory, we study specific learning examples and apply our learning theory to derive error bounds for multicategory ψ-learning.

Example 1 (Linear classification). We consider linear classification involving a class of k hyperplanes $\mathcal{F} = \{\mathbf{f} : f_j(\mathbf{x}) = \mathbf{w}_j^T\mathbf{x} + b_j, \sum_{j=1}^k f_j = 0, \mathbf{x} \in S = [0, 1]^d\}$, where d is a constant. To generate the training sample, we specify $P(Y = j) = 1/k$, $P(\mathbf{x}|Y = j) = k - 1$ for $\{\mathbf{x} : x_1 \in [(j-1)/k, j/k)\}$ and $1/(k-1)$ otherwise, where x_1 is the first coordinate of $\mathbf{x}$. Then the Bayes classifier yields sets $\{\mathbf{x} : x_1 \in [0, 1/k)\}, \ldots, \{\mathbf{x} : x_1 \in [(k-1)/k, 1]\}$ for the corresponding k classes.

We now verify Assumptions A–C. For Assumption A, it is easy to find that $\mathbf{f}_t = (w_{11}x_1 + b_1, \ldots, w_{1k}x_1 + b_k)$ such that the w_{1j}'s are increasing, $\sum_{j=1}^k w_{1j} = 0$, $\sum_{j=1}^k b_j = 0$, and $w_{1j}j/k + b_j = w_{1,j+1}j/k + b_{j+1}$, $j = 1, \ldots, k-1$. Let $\mathbf{f}_0 = n\mathbf{f}_t \in \mathcal{F}$; then $e_\psi(\mathbf{f}_0, \bar{\mathbf{f}}) \le s_n = c_1 n^{-1}$ for some constant $c_1 > 0$. This implies Assumption A with $s_n = c_1 n^{-1}$. Assumption B is satisfied with $\alpha = +\infty$ because $P(\mathbf{X} \in S : (\max P_l(\mathbf{X}) - P_j(\mathbf{X})) < 2\delta) = P(X_1 \in \{1/k, \ldots, (k-1)/k\}) = 0$ for any sufficiently small $\delta > 0$. To verify Assumption C, we note that $H_B(u, \mathcal{G}(\ell)) \le O(k^2 \log(k/u))$ for any given ℓ by Lemma A.1 in the Appendix. Let $\phi_1(\epsilon_n, \ell) = c_3(k^2 \log(k/L^{1/2}))^{1/2}/L^{1/2}$. This in turn

yields $\sup_{\ell\geq 2}\phi(\epsilon_n,\ell)\leq\phi_1(\epsilon_n,2)=c(k^2\log(k/\epsilon_n))^{1/2}/\epsilon_n$ for some $c>0$ and a rate $\epsilon_n=(\frac{k^2\log n}{n})^{1/2}$ when $C/J_0\sim\delta_n^{-2}n^{-1}\sim\frac{1}{k^2\log n}$, provided that $\frac{k^2\log n}{n}\to 0$.

By Corollary 1, we conclude that $e(\hat{\mathbf{f}},\bar{\mathbf{f}})=O(\frac{k^2\log n}{n})$ except for a set of probabilities tending to 0, and $Ee(\hat{\mathbf{f}},\bar{\mathbf{f}})=O(\frac{k^2\log n}{n})$, when $\frac{k^2\log n}{n}\to 0$ as $n\to\infty$. It is interesting to note that $Ee(\hat{\mathbf{f}},\bar{\mathbf{f}})=O(n^{-1}\log n)$ when k is a fixed constant. This conclusion holds generally for any ψ-function satisfying Assumption D.

Example 2 (Gaussian kernel classification). In this example we consider nonlinear learning with the same $P(\mathbf{x},y)$ as in Example 1. Let $\mathcal{F}=\{\mathbf{f}:f_j(\mathbf{x})=\sum_{i=1}^n v_{ji}K(\mathbf{x}_i,\mathbf{x})+b_j,\ \sum_{j=1}^k f_j=0,\mathbf{x}\in S=[0,1]^d\}$ with Gaussian kernel $K(s,t)=\exp(-\|s-t\|^2/\sigma^2)$.

For Assumption A, we note that $\mathcal{F}$ is a rich function space with large n. In fact, any continuous function can be well approximated by Gaussian kernel-based representations under the sup-norm (cf. Steinwart 2001). Thus there exists an $\mathbf{f}_t=(f_{1t},\ldots,f_{kt})\in\mathcal{F}$ such that $f_{jt}(\mathbf{x})\geq 0$ for $x_1\in[(j-1)/k,j/k]$ and <0 otherwise. With a choice of $\mathbf{f}_0=\epsilon_n^{-2}\mathbf{f}_t$, $e_\psi(\mathbf{f}_0,\bar{\mathbf{f}})\leq s_n=c_1\epsilon_n^2$, where c_1 is a constant and ϵ_n is defined later. Assumption B is satisfied with $\alpha=+\infty$ as in Example 1. In this case the metric entropy of $\mathcal{F}^\psi(\ell)$ appears to be easier to compute. We then apply Theorem 4 to obtain the convergence rate. Consider any ψ-function in (2) satisfies (10). Then, by Lemma A.2 in the Appendix, $H_B(u,\mathcal{F}^\psi(\ell))\leq O(k(\log(\ell/u))^{d+1})$ for any given ℓ. Let $\phi_1^*(\epsilon_n^*,\ell)=c_3(k(\log(\ell/L^{*1/2}))^{d+1})^{1/2}/L^{*1/2}$. Then $\sup_{\ell\geq 2}\phi^*(\epsilon_n^*,\ell)\leq\phi_1(\epsilon_n^*,2)=c(k(\log(1/\epsilon_n^*))^{d+1})^{1/2}/\epsilon_n^*$ for some $c>0$. Solving (9) yields a rate $\epsilon_n^*=(\frac{k(\log(nk^{-1}))^{d+1}}{n})^{1/2}$ when $C/J_0\sim\delta_n^{*-2}n^{-1}\sim\frac{1}{k(\log(nk^{-1}))^{d+1}}$ under the condition that $\frac{k(\log(nk^{-1}))^{d+1}}{n}\to 0$ as $n\to\infty$.

By Theorem 4, we conclude that $e(\hat{\mathbf{f}},\bar{\mathbf{f}})\leq e_\psi(\hat{\mathbf{f}},\bar{\mathbf{f}})=O(\frac{k(\log(nk^{-1}))^{d+1}}{n})$ except for a set of probabilities tending to 0. By Corollary 2, $Ee(\hat{\mathbf{f}},\bar{\mathbf{f}})=O(\frac{k(\log(nk^{-1}))^{d+1}}{n})$. This resulting rate reflects the dependence of the rate on the class number k. If k is treated as a fixed constant, then we have $Ee(\hat{\mathbf{f}},\bar{\mathbf{f}})=O(n^{-1}(\log n)^{d+1})$. This conclusion holds generally for any ψ-function satisfying Assumption D and condition (10).

In summary, Examples 1 and 2 provide an insight into the generalization error of the proposed methodology. In view of the lower bound n^{-1} result (cf. Tsybakov 2004) in the binary case, we conjecture the rates obtained in Examples 1 and 2 are nearly optimal, although to our knowledge a lower bound result for any general classifier has not yet been established in the multicategory case. Further investigation is needed.

4. NUMERICAL EXAMPLES

In this section we examine performance of multicategory ψ-learning in terms of generalization and compare it with its counterpart SVM. The literature includes numerous different multicategory SVM generalizations (see, e.g., Lee et al. 2004; Crammer and Singer 2001; Weston and Watkins 1999). To make a fair comparison, we use a version of multicategory SVM that is parallel to our multicategory ψ-learning. Specifically, we replace the ψ function in (4) and (5) by ψ_1. Then, for the linear case, this version of multicategory SVM solves

$$\min_{\mathbf{b},\mathbf{w}}\left(\frac{1}{2}\sum_{j=1}^k\|\mathbf{w}_j\|^2+C\sum_{i=1}^n\psi_1(\mathbf{g}(\mathbf{x}_i,y_i))\right)\quad\text{subject to}\quad\sum_{j=1}^k b_j\mathbf{1}_n+X\sum_{j=1}^k\mathbf{w}_j=0. \tag{11}$$

This version of SVM is closely related to that of Crammer and Singer (2001). In their formulation, all b_j's are set to be 0 rather than using the sum-to-0 constraint, in contrast to (11). As argued by Guermeur (2002), the sum-to-0 constraint is necessary to ensure uniqueness of the solution when a k-dimensional vector of decision functions with intercepts b_j's is used for a k-class problem.

4.1 Simulation

We consider two linear examples in which the GE is approximated by the testing error using a testing sample, independent of training. In what follows, all calculations are carried out using the IMSL C routines.

Three-Class Linear Problem. The training data are generated as follows. First, generate pairs (t_1,t_2) from a bivariate t-distribution with degrees of freedom ν, where $\nu=1,3$ in Examples 1 and 2. Second, randomly assign $\{1,2,3\}$ to its label index for each (t_1,t_2). Third, calculate (x_1,x_2) as $x_1=t_1+a_1$ and $x_2=t_2+a_2$ with three different values of $(a_1,a_2)=(\sqrt{3},1),(-\sqrt{3},1),(0,-2)$ for corresponding classes 1–3. In these examples, the testing and Bayes errors are computed through independent testing samples of size 10^6 for classifiers obtained from training samples of size 150.

To eliminate the dependence on C, the performance of ψ-learning and SVM is maximized by optimizing C over a discrete set in $[10^{-3},10^3]$. For each method, the testing error for the optimal C is averaged over 100 repeated simulations. The simulation results are summarized in Table 1.

As shown in Table 1, ψ-learning usually has a smaller testing error, and thus better generalization, than its counterpart SVM. But the amount of improvement varies across examples. In example 1, the percent of improvement of multicategory ψ-learning over SVM is 43.22% when the corresponding t-distribution has 1 df. In example 2, it decreases to 20.41%

Table 1. Testing, Training Errors, and Their $\hat{e}(\cdot,\bar{\mathbf{f}})$ of SVM and ψ-Learning Using the Best C in Examples 1 and 2 With $n=150$, Averaged Over 100 Simulation Replications and Their Standard Errors in Parentheses

Example	Method	$\text{Training}_{(se)}$	$\text{Testing}_{(se)}$	$\hat{e}(\cdot,\bar{\mathbf{f}})_{(se)}$	No. $\text{SV}_{(se)}$
df = 1	SVM	$.4002_{(.1469)}$	$.4305_{(.1405)}$	$.1835_{(.1405)}$	$141.76_{(10.97)}$
	ψ-L	$.3199_{(.1237)}$	$.3494_{(.1209)}$	$.1024_{(.1209)}$	$64.64_{(15.43)}$
df = 3	SVM	$.1447_{(.0267)}$	$.1505_{(.0045)}$	$.0049_{(.0045)}$	$71.81_{(11.02)}$
	ψ-L	$.1429_{(.0285)}$	$.1495_{(.0033)}$	$.0039_{(.0033)}$	$41.29_{(13.51)}$

NOTE: In example 1, df = 1, the Bayes error is .2470 with the improvement of ψ-learning over SVM 43.22%. In example 2, df = 3, the Bayes error is .1456 with the improvement of ψ-learning over SVM 20.41%. Here, the improvement of ψ-learning over SVM is defined by $(T(\text{SVM})-T(\psi))/\hat{e}(\text{SVM},\bar{\mathbf{f}})$, where $\hat{e}(\cdot,\bar{\mathbf{f}})=T(\cdot)-$ Bayes error, and $T(\cdot)$ denotes the testing error of a given method.

when the t-distribution with 3 df is used. Further, ψ-learning yields a smaller number of support vectors. This suggests that ψ-learning has an even more "sparse" solution than SVM, and hence has stronger ability of data reduction. In a related matter, SVM fails to give data reduction in example 1 because almost all of the instances are support vectors, in contrast to the much smaller number of support vectors for ψ-learning. One plausible explanation is that the first moment of the standard bivariate t-distribution does not exist, and thus the corresponding SVM does not work well. In general, any classifier with an unbounded loss such as SVM may have problems with extreme outliers as in this example. This reinforces our view that ψ-learning is more robust against outliers.

4.2 Application

We now examine the performance of ψ-learning and its counterpart SVM on a benchmark example "letter," obtained from Statlog. In this example, each sample contains 16 primitive numerical attributes converted from its corresponding letter image with a response variable representing 26 categories. The main goal here is to identify each letter image as one of the 26 capital letters in the English alphabet. A detailed description can be found in *www.liacc.up.pt/ML/statlog/datasets/letter/letter.doc.html*.

For illustration, we use the data for letters D, O, and Q with 805, 753, and 783 cases. A random sample of $n = 200$ is selected for training, while leaving the rest for testing. For each training dataset, we seek the best performance of linear ψ-learning and SVM over a set of C-values in $[10^{-3}, 10^{3}]$. Table 2 reports the corresponding results with respect to the smallest testing errors for each method in 10 different cases. Because the Bayes error is unknown, the improvement of ψ-learning over SVM is computed using $(T(\text{SVM}) - T(\psi))/T(\text{SVM})$.

Table 2 indicates that multicategory ψ-learning has a smaller testing error than its counterpart SVM, although the amount of improvement varies from sample to sample. In addition, on average, multicategory ψ-learning has a smaller number of support vectors than SVM. In conclusion, ψ-learning has better generalization and achieves more data reduction than SVM in this example.

5. DISCUSSION

In this article we have proposed a new methodology that generalizes ψ-learning from the binary case to the multicategory case. We have developed a statistical learning theory for ψ-learning in terms of the Bayesian regret. In simulations, we have shown that the proposed methodology performs well and is more robust against outliers than its counterpart SVM. In addition, we have discovered some interesting phenomena that are not with the binary case.

Table 2. Testing Errors for Problem Letter

Case	*SVM*	*ψ-L*	*Improvement (%)*
1	.083	.079	3.39%
2	.073	.063	12.24%
3	.086	.076	11.41%
4	.072	.072	0%
5	.088	.085	3.74%
6	.077	.073	5.45%
7	.075	.072	4.39%
8	.079	.075	5.92%
9	.093	.091	1.51%
10	.090	.086	4.11%
Average no. of SVs	51.1	40.8	

NOTE: Each training dataset is of size 200 and selected from a total of 2,341 samples.

Recently, there has been considerable interest in studying the variable selection problem replacing the conventional L_2 norm with the L_1 norm. In the binary case, Zhu, Hastie, Rosset, and Tibshirani (2003) studied properties of the L_1 SVM and showed that the corresponding regularized solution path is piecewise linear. Thus it is natural to investigate variable selection of the L_1 ψ-learning.

Further developments are needed to make multicategory ψ-learning more useful in practice, particularly methodologies for a data-driven choice of C, variable selection, and regularized solution paths, as well as the nonstandard situation including unequal loss assignments.

APPENDIX: PROOFS

Proof of Theorem 1

By the definition of $\text{Err}(\mathbf{f})$, it is easy to obtain through conditioning that $e(\mathbf{f}, \bar{\mathbf{f}}) = \frac{1}{2}E[\sum_{l=1}^{k} P_l(\mathbf{X})(\text{sign}(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), l)) - \text{sign}(\mathbf{g}(\mathbf{f}(\mathbf{X}), l)))]$. Then it suffices to consider the situation where $\text{sign}(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), l)) - \text{sign}(\mathbf{g}(\mathbf{f}(\mathbf{X}), l))$ is nonzero, that is, when two classifiers disagree. Equivalently, for any given $\mathbf{X} = \mathbf{x}$, we can write $e(\mathbf{f}, \bar{\mathbf{f}})$ using all possible different classifications produced by $\bar{\mathbf{f}}$ and $\mathbf{f}$ jointly, where $\text{sign}(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{x}), l)) = 1$ and $\text{sign}(\mathbf{g}(\mathbf{f}(\mathbf{x}), j)) = 1$ imply that $\bar{\mathbf{f}}$ classifies $\mathbf{x}$ into class l while $\mathbf{f}$ classifies $\mathbf{x}$ into class j for $1 \le l \ne j \le k$. Thus we have

$$
\begin{aligned}
&e(\mathbf{f}, \bar{\mathbf{f}}) \\
&\quad = E\Bigg[\sum_{l=1}^{k}\sum_{j\ne l}(P_l(\mathbf{X}) - P_j(\mathbf{X})) \\
&\qquad\qquad \times I\big(\text{sign}\big(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), l)\big) = 1, \text{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), j)\big) = 1\big)\Bigg] \\
&\quad = E\Bigg[\sum_{l=1}^{k}\sum_{j\ne l}|P_l(\mathbf{X}) - P_j(\mathbf{X})| \\
&\qquad\qquad \times I\big(\text{sign}\big(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), l)\big) = 1, \text{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), j)\big) = 1\big)\Bigg],
\end{aligned}
$$

where the second equality follows from the fact that $\bar{\mathbf{f}}$ is the optimal (Bayes) decision function vector such that $P_l(\mathbf{X}) \ge P_j(\mathbf{X})$ when $\text{sign}(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), l)) = 1$. The desired result then follows.

Proof of Theorem 2

Write $E[1 - \text{sign}(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y))|\mathbf{X} = \mathbf{x}]$ as $\sum_{j=1}^{k}(1 - \text{sign}(\mathbf{g}(\mathbf{f}(\mathbf{x}), j)))$ $P_j(\mathbf{x}) = 1 - \sum_{j=1}^{k}\text{sign}(\mathbf{g}(\mathbf{f}(\mathbf{x}), j))P_j(\mathbf{x})$. Note that for any given $\mathbf{x}$, one and only one of $\text{sign}(\mathbf{g}(\mathbf{f}(\mathbf{x}), j))$ can be 1, and the rest must be equal to -1. Consequently, $E[1 - \text{sign}(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y))]$ is minimized when $\text{sign}(\mathbf{g}(\mathbf{f}(\mathbf{x}), \text{argmax}_j \bar{f}_j(\mathbf{x}))) = 1$, that is, $\mathbf{f} = \bar{\mathbf{f}}$. Evidently, the minimizer is not unique, because $c\bar{\mathbf{f}}$ for $c \ge 1$ is also a minimizer. The desired result then follows from the fact that $\psi(\mathbf{u}) \ge (1 - \text{sign}(\mathbf{u}))$ and $\psi(\bar{\mathbf{g}}) = 1 - \text{sign}(\bar{\mathbf{g}})$.

Proof of Theorem 3

Before proceeding, we introduce some notations that we use later. Let $\tilde{l}_\psi(\mathbf{f}, Z_i) = l_\psi(\mathbf{f}, Z_i) + \lambda J(\mathbf{f})$ be the cost function to be minimized, as in (4) or (5), where $l_\psi(\mathbf{f}, Z_i) = \psi(\mathbf{g}(\mathbf{f}(\mathbf{X}_i), Y_i))$ and $\lambda = 1/(Cn)$.

Let $\tilde{l}(\mathbf{f}, Z_i) = l(\mathbf{f}, Z_i) + \lambda J(\mathbf{f})$, where $l(\mathbf{f}, Z_i) = 1 - \operatorname{sign}(\mathbf{g}(\mathbf{f}(\mathbf{X}_i), Y_i))$. Define the scaled empirical process $E_n(\tilde{l}(\mathbf{f}, Z) - \tilde{l}_\psi(\mathbf{f}_0, Z))$ as

$$n^{-1}\sum_{i=1}^{n}\big(\tilde{l}(\mathbf{f}, Z_i) - \tilde{l}_\psi(\mathbf{f}_0, Z_i) - E[\tilde{l}(\mathbf{f}, Z_i) - \tilde{l}_\psi(\mathbf{f}_0, Z_i)]\big) = E_n[l(\mathbf{f}, Z) - l_\psi(\mathbf{f}_0, Z)],$$

where $Z = (\mathbf{X}, Y)$. Let $A_{i,j} = \{\mathbf{f} \in \mathcal{F} : 2^{i-1}\delta_n^2 \le e(\mathbf{f}, \bar{\mathbf{f}}) < 2^i\delta_n^2, 2^{j-1}J_0 \le J(\mathbf{f}) < 2^jJ_0\}$, and $A_{i,0} = \{\mathbf{f} \in \mathcal{F} : 2^{i-1}\delta_n^2 \le e(\mathbf{f}, \bar{\mathbf{f}}) < 2^i\delta_n^2, J(\mathbf{f}) < J_0\}$, for $j = 1, 2, \ldots$ and $i = 1, 2, \ldots$. Without loss of generality, we assume that $J(\mathbf{f}_0) \ge 1$ and $\max(\epsilon_n^2, 2s_n) < 1$ in the sequel.

The proof uses the treatments of Shen et al. (2003) and Shen (1998), together with the results in Theorem 1 and Assumption B. In what follows, we omit any detail that can be referred to the proof of theorem 1 of Shen et al. (2003).

Using the connection between $e(\hat{\mathbf{f}}, \bar{\mathbf{f}})$ and the cost function of Shen et al. (2003), we have

$$\begin{aligned} &P\big(e(\hat{\mathbf{f}}, \bar{\mathbf{f}}) \ge \delta_n^2\big) \\ &\quad\le P^*\left(\sup_{\{\mathbf{f}\in\mathcal{F}:\, e(\mathbf{f},\bar{\mathbf{f}})\ge\delta_n^2\}} n^{-1}\sum_{i=1}^{n}\big(\tilde{l}_\psi(\mathbf{f}_0, Z_i) - \tilde{l}(\mathbf{f}, Z_i)\big) \ge 0\right) \\ &\quad= I, \end{aligned}$$

where P^* denotes the outer probability measure.

To bound I, we need some inequalities regarding the first and second moments of $\tilde{l}(\mathbf{f}, Z) - \tilde{l}_\psi(\mathbf{f}_0, Z)$ for $\mathbf{f} \in A_{ij}$.

For the first moment, note that $E[l(\mathbf{f}, Z) - l_\psi(\mathbf{f}_0, Z)] = E[l(\mathbf{f}, Z) - l_\psi(\bar{\mathbf{f}}, Z)] - E[l_\psi(\mathbf{f}_0, Z) - l_\psi(\bar{\mathbf{f}}, Z)]$, which is equal to $2(e(\mathbf{f}, \bar{\mathbf{f}}) - e_\psi(\mathbf{f}_0, \bar{\mathbf{f}}))$ because $El_\psi(\bar{\mathbf{f}}, Z) = El(\bar{\mathbf{f}}, Z)$ by Theorem 2. By Assumption A and the definition of δ_n^2, $2e_\psi(\mathbf{f}_0, \bar{\mathbf{f}}) \le 2s_n \le \delta_n^2$. Then, using the assumption that $J_0\lambda \le \delta_n^2/2$, we have, for any integers $i, j \ge 1$,

$$\begin{aligned} \inf_{A_{i,j}} E\big(\tilde{l}(\mathbf{f}, Z) - \tilde{l}_\psi(\mathbf{f}_0, Z)\big) &\ge M(i, j) \\ &= (2^{i-1}\delta_n^2) + \lambda(2^{j-1} - 1)J(\mathbf{f}_0) \end{aligned} \tag{A.1}$$

and

$$\begin{aligned} \inf_{A_{i,0}} E\big(\tilde{l}(\mathbf{f}, Z) - \tilde{l}_\psi(\mathbf{f}_0, Z)\big) &\ge (2^{i-1} - 1/2)\delta_n^2 \\ &\ge M(i, 0) = 2^{i-2}\delta_n^2, \end{aligned} \tag{A.2}$$

where the fact that $2^i - 1 \ge 2^{i-1}$ has been used.

For the second moment, it follows from Theorem 1 and Assumption B that for any $\mathbf{f} \in \mathcal{F}$,

$$\begin{aligned} &e(\mathbf{f}, \bar{\mathbf{f}}) \\ &= E\left[\sum_{l=1}^{k}\sum_{j\ne l}|P_l(\mathbf{X}) - P_j(\mathbf{X})| \right. \\ &\qquad\left. \times I\big(\operatorname{sign}\big(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), l)\big) = 1, \operatorname{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), j)\big) = 1\big)\right] \\ &\ge 2\delta\left(E\left[\sum_{l=1}^{k}\sum_{j\ne l} I\big(\operatorname{sign}\big(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), l)\big) = 1, \operatorname{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), j)\big) = 1\big)\right.\right. \\ &\qquad\left.\left. \times I\big(|P_l(\mathbf{X}) - P_j(\mathbf{X})| \ge 2\delta\big)\right]\right) \\ &\ge \delta\left(E\left[2\sum_{l=1}^{k}\sum_{j\ne l} I\big(\operatorname{sign}\big(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), l)\big) = 1, \operatorname{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), j)\big) = 1\big)\right]\right. \\ &\qquad\left. - 2c_1\delta^\alpha\right) \\ &= \frac{1}{2}(4c_1)^{-1/\alpha} \\ &\qquad\times E\left[2\sum_{l=1}^{k}\sum_{j\ne l} I\big(\operatorname{sign}\big(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), l)\big) = 1, \right. \\ &\qquad\qquad\left.\operatorname{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), j)\big) = 1\big)\right]^{(\alpha+1)/\alpha}, \end{aligned} \tag{A.3}$$

with a choice of $\delta = (E[2\sum_{l=1}^{k}\sum_{j\ne l} I(\operatorname{sign}(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), l)) = 1, \operatorname{sign}(\mathbf{g}(\mathbf{f}(\mathbf{X}), j)) = 1)]/(4c_1))^{1/\alpha}$.

We now establish a connection between the first and second moments. By Theorem 2, $E[\psi(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), Y)) - (1 - \operatorname{sign}(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), Y)))] = 0$. Note that $\psi(\mathbf{u}) \ge 1 - \operatorname{sign}(\mathbf{u})$ for any $\mathbf{u} \in \mathbb{R}^{k-1}$, $E|\psi(\mathbf{g}_0(\mathbf{f}_0(\mathbf{X}), Y)) - (1 - \operatorname{sign}(\mathbf{g}_0(\mathbf{f}_0(\mathbf{X}), Y)))| = E[\psi(\mathbf{g}_0(\mathbf{f}_0(\mathbf{X}), Y)) - (1 - \operatorname{sign}(\mathbf{g}_0(\mathbf{f}_0(\mathbf{X}), Y)))] \le 2e_\psi(\mathbf{f}_0, \bar{\mathbf{f}})$. By the triangular inequality,

$$\begin{aligned} &E[l(\mathbf{f}, Z) - l_\psi(\mathbf{f}_0, Z)]^2 \\ &\quad\le 2E\big|1 - \operatorname{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y)\big) - \psi\big(\mathbf{g}_0(\mathbf{f}_0(\mathbf{X}), Y)\big)\big| \\ &\quad\le 2\big(2e_\psi(\mathbf{f}_0, \bar{\mathbf{f}}) + E\big|\operatorname{sign}\big(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), Y)\big) - \operatorname{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y)\big)\big| \\ &\qquad + E\big|\operatorname{sign}\big(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), Y)\big) - \operatorname{sign}\big(\mathbf{g}_0(\mathbf{f}_0(\mathbf{X}), Y)\big)\big|\big). \end{aligned} \tag{A.4}$$

Note that for any $\mathbf{f} \in \mathcal{F}$,

$$\begin{aligned} &E\big|\operatorname{sign}\big(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), Y)\big) - \operatorname{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y)\big)\big| \\ &\quad= E\left[\sum_{l=1}^{k} I(Y = l)\big|\operatorname{sign}\big(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), l)\big) - \operatorname{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), l)\big)\big|\right] \\ &\quad= E\left[2\sum_{l=1}^{k} I(Y = l) \right. \\ &\qquad\left. \times \sum_{j\ne l} I\big(\operatorname{sign}\big(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), l)\big) = 1, \operatorname{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), j)\big) = 1\big)\right] \\ &\quad\le E\left[2\sum_{l=1}^{k}\sum_{j\ne l} I\big(\operatorname{sign}\big(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), l)\big) = 1, \operatorname{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), j)\big) = 1\big)\right]. \end{aligned}$$

This, together with (A.3), implies that

$$E\big|\operatorname{sign}\big(\bar{\mathbf{g}}(\bar{\mathbf{f}}(\mathbf{X}), Y)\big) - \operatorname{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y)\big)\big| \le c^* e(\mathbf{f}, \bar{\mathbf{f}})^{\alpha/(\alpha+1)}, \tag{A.5}$$

where $c^* = 2^{\alpha/(\alpha+1)}(4c_1)^{1/(\alpha+1)}$. For any $\mathbf{f} \in A_{i,j}$, $e(\mathbf{f}, \bar{\mathbf{f}})^{\alpha/(\alpha+1)} \ge (2^{-1}\delta_n^2)^{\alpha/(\alpha+1)} \ge 2^{-1}\delta_n^2 \ge s_n \ge e_\psi(\mathbf{f}_0, \bar{\mathbf{f}})$, $e(\mathbf{f}, \bar{\mathbf{f}}) \ge e(\mathbf{f}_0, \bar{\mathbf{f}})$, together with (A.4) and (A.5), imply that

$$\begin{aligned} &E[l(\mathbf{f}, Z) - l_\psi(\mathbf{f}_0, Z)]^2 \\ &\quad\le 2\big(2e_\psi(\mathbf{f}_0, \bar{\mathbf{f}}) + c^*(e(\mathbf{f}, \bar{\mathbf{f}})^{\alpha/(\alpha+1)} + e(\mathbf{f}_0, \bar{\mathbf{f}})^{\alpha/(\alpha+1)})\big) \\ &\quad\le c_3^0(e(\mathbf{f}, \bar{\mathbf{f}})/2)^{\alpha/(\alpha+1)}, \end{aligned}$$

with $c_3^0 = 16c_1^{1/(\alpha+1)} + 8$. Consequently, for $i = 1, \ldots$ and $j = 0, 1, \ldots$,

$$\begin{aligned} \sup_{A_{i,j}} E[l_\psi(\mathbf{f}_0, Z) - l(\mathbf{f}, Z)]^2 &\le c_3^0(2^{i-1}\delta_n^2)^{\alpha/(\alpha+1)} \\ &\le c_3 M(i, j)^{\alpha/(\alpha+1)} = v(i, j)^2, \end{aligned}$$

where $c_3 = 2c_3^0$.

We are now ready to bound I. Using the assumption that $J_0\lambda \le \delta_n^2/2$, (A.1) and (A.2), we have $I \le \sum_{i\ge 1, j\ge 0} P^*(\sup_{A_{i,j}} E_n(l_\psi(\mathbf{f}_0, Z) - l(\mathbf{f}, Z)) \ge M(i,j))$. By definition, $l_\psi(\mathbf{f}_0, Z)$ and $l(\mathbf{f}, Z)$ are between 0 and 2. Then $E[l_\psi(\mathbf{f}_0, Z) - l(\mathbf{f}, Z)]^2 \le 4$ and $E_n(l_\psi(\mathbf{f}_0, Z) - l(\mathbf{f}, Z)) \le 4$. For convenience, we scale the empirical process by a constant $t = (4c_3^{1/2})^{-1}$ in what follows. Then

$$\begin{aligned} I &\le \sum_{i,j} P^*\Big(\sup_{A_{i,j}} E_n\big(t[l_\psi(\mathbf{f}_0, Z) - l(\mathbf{f}, Z)]\big) \ge M^c(i,j)\Big) \\ &\quad + \sum_i P^*\Big(\sup_{A_{i,0}} E_n\big(t[l_\psi(\mathbf{f}_0, Z) - l(\mathbf{f}, Z)]\big) \ge M^c(i,0)\Big) \\ &= I_1 + I_2, \end{aligned} \tag{A.6}$$

and $\sup_{A_{i,j}} E[t(l_\psi(\mathbf{f}_0, Z) - l(\mathbf{f}, Z))]^2 \le v^c(i,j)^2$, where $v^c(i,j) = \min(t^{1/2}v(i,j), 1)$, $M^c(i,j) = \min(tM(i,j), c_3^{-1/2})$. Note that $v^c(i,j) < 1$ implies that $M^c(i,j) = tM(i,j)$.

Next, we bound I_i separately. For I_1, we verify the required conditions (4.5)–(4.7) in theorem 3 of Shen and Wong (1994). To compute the metric entropy in (4.7) there, we need to construct a bracketing function of $l_\psi(\mathbf{f}_0, Z) - l(\mathbf{f}, Z)$. Denote an ϵ-bracketing set for $\{(G_{f_1}, \ldots, G_{f_k}); \mathbf{f} \in A_{ij}\}$ to be $\{(G_{p1}^v, \ldots, G_{pm}^v), (G_{p1}^u, \ldots, G_{pm}^u)\}$, $p = 1, \ldots, k$. Let $s_{ph}^v(\mathbf{x})$ be -1 if $\mathbf{x} \in G_{ph}^u$ and 1 otherwise, and $s_{ph}^u(\mathbf{x})$ be -1 if $\mathbf{x} \in G_{ph}^v$ and 1 otherwise, $p = 1, \ldots, k$, $h = 1, \ldots, m$. Then $\{(s_{p1}^v, \ldots, s_{pm}^v), (s_{p1}^u, \ldots, s_{pm}^u)\}$ forms an ϵ-bracketing function of $-\operatorname{sign}(\mathbf{g}(\mathbf{f}(\mathbf{x}), p))$ for $\mathbf{f} \in A_{ij}$ and $p = 1, \ldots, k$. This implies that for any $\epsilon \ge 0$ and $\mathbf{f} \in A_{ij}$, there exists an h $(1 \le h \le m)$ such that $l_h^v(z) \le l(\mathbf{f}, z) - l_\psi(\mathbf{f}_0, z) \le l_h^u(z)$ for any $z = (\mathbf{x}, y)$, where $l_h^u(z) = 1 + \sum_{p=1}^k s_{ph}^u(\mathbf{x})I(y = p) - l_\psi(\mathbf{f}_0, z)$, $l_h^v(z) = 1 + \sum_{p=1}^k s_{ph}^v(\mathbf{x})I(y = p) - l_\psi(\mathbf{f}_0, z)$, and $(E[l_h^u - l_h^v]^2)^{1/2} = (\sum_{p=1}^k E[(s_{ph}^u(\mathbf{x}) - s_{ph}^v(\mathbf{x}))I(y = p)]^2)^{1/2} \le 2(\max_p P(G_{ph}^u \Delta G_{ph}^l))^{1/2} \le 2\epsilon^{1/2}$. So $(E[l_h^u - l_h^v]^2)^{1/2} \le \min(2\epsilon^{1/2}, 2)$. Hence $H_B(\epsilon, \mathcal{F}^*(2^j)) \le H(\epsilon^2/4, \mathcal{G}(2^j))$ for any $\epsilon > 0$ and $j = 0, \ldots,$ where $\mathcal{F}^*(2^j) = \{l(\mathbf{f}, z) - l_\psi(\mathbf{f}_0, z) : \mathbf{f} \in \mathcal{F}, J(\mathbf{f}) \le 2^j\}$. Using the fact that $\int_{aM^c(i,j)}^{v^c(i,j)} H_B^{1/2}(u^2/4, \mathcal{G}(2^j))\, du/M^c(i,j)$ is nonincreasing in i and $M^c(i,j)$; $i = 1, \ldots,$ we have

$$\begin{aligned} &\int_{aM^c(i,j)}^{v^c(i,j)} H_B^{1/2}\big(u^2/4, \mathcal{G}(2^j)\big)\, du/M^c(i,j) \\ &\quad \le \int_{aM^c(1,j)}^{c_3^{1/2}M^c(1,j)^{\alpha/(2(\alpha+1))}} H_B^{1/2}\big(u^2/4, \mathcal{G}(2^j)\big)\, du/M^c(1,j) \\ &\quad \le \phi(\epsilon_n, 2^j), \end{aligned}$$

where $a = \varepsilon/32$ with ε as defined later. Thus (4.7) of Shen and Wong (1994) holds with $M = n^{1/2}M^c(i,j)$ and $v = v^c(i,j)^2$, and so does (4.5). In addition, with $T = 1$,

$$\frac{M^c(i,j)}{v^c(i,j)^2} \le \max\big(c_3^{-1/2}, c_3^{-(2\alpha+3)/(2\alpha+2)}\big) = c_3^{-1/2} \le \frac{\varepsilon}{4T}$$

implies (4.6) with $\varepsilon = 4c_3^{-1/2} < 1$.

Note that $0 < \delta_n \le 1$ and $\lambda J_0 \le \delta_n^2/2$. Using a similar argument as in Shen et al. (2003), an application of theorem 3 of Shen and Wong (1994) yields that

$$I_1 \le 3\exp\big(-c_5 n(\lambda J(\mathbf{f}_0))^{(\alpha+2)/(\alpha+1)}\big) \Big/ \big[1 - \exp\big(-c_5 n(\lambda J(\mathbf{f}_0))^{(\alpha+2)/(\alpha+1)}\big)\big]^2.$$

Here and in the sequel, c_5 is a positive generic constant. I_2 can be bounded similarly.

Finally, $I \le 6\exp(-c_5 n(\lambda J(\mathbf{f}_0))^{(\alpha+2)/(\alpha+1)})/[1 - \exp(-c_5 \times n(\lambda J(\mathbf{f}_0))^{(\alpha+2)/(\alpha+1)})]^2$. This implies that $I^{1/2} \le (5/2 + I^{1/2}) \times \exp(-c_5 n(\lambda J(\mathbf{f}_0)))$. The result then follows from the fact that $I \le I^{1/2} \le 1$.

Proof of Corollary 1

The result follows from Theorem 3 and its proof.

Proof of Theorem 4

This proof is similar to that of Theorem 3. For simplicity, we sketch only the parts that require modifications. Consider the scaled empirical process $E_n(\tilde{l}_\psi(\mathbf{f}, Z) - \tilde{l}_\psi(\mathbf{f}_0, Z))$ and let $A_{i,j} = \{\mathbf{f} \in \mathcal{F} : 2^{i-1}\delta_n^{*2} \le e_\psi(\mathbf{f}, \bar{\mathbf{f}}) < 2^i\delta_n^{*2}, 2^{j-1}J_0 \le J(\mathbf{f}) < 2^jJ_0\}$ and $A_{i,0} = \{\mathbf{f} \in \mathcal{F} : 2^{i-1}\delta_n^{*2} \le e_\psi(\mathbf{f}, \bar{\mathbf{f}}) < 2^i\delta_n^{*2}, J(\mathbf{f}) < J_0\}$, for $j = 1, 2, \ldots$ and $i = 1, 2, \ldots$. Using an analogous argument, we have

$$\begin{aligned} &P\big(e_\psi(\hat{\mathbf{f}}, \bar{\mathbf{f}}) \ge \delta_n^{*2}\big) \\ &\quad \le P^*\left(\sup_{\{\mathbf{f}\in\mathcal{F} : e_\psi(\mathbf{f}, \bar{\mathbf{f}}) \ge \delta_n^{*2}\}} n^{-1}\sum_{i=1}^n \big(\tilde{l}_\psi(\mathbf{f}_0, Z_i) - \tilde{l}_\psi(\mathbf{f}, Z_i)\big) \ge 0\right) \\ &\quad = I. \end{aligned}$$

To bound I, we consider the first and second moments of $\tilde{l}_\psi(\mathbf{f}, Z) - \tilde{l}_\psi(\mathbf{f}_0, Z)$ for $\mathbf{f} \in A_{ij}$. For the first moment, it is straightforward to show that for any integers $i, j \ge 1$, $\inf_{A_{i,j}} E(\tilde{l}_\psi(\mathbf{f}, Z) - \tilde{l}_\psi(\mathbf{f}_0, Z)) \ge M(i,j) = (2^{i-1}\delta_n^{*2}) + \lambda(2^{j-1} - 1)J(\mathbf{f}_0)$, and $\inf_{A_{i,0}} E(\tilde{l}_\psi(\mathbf{f}, Z) - \tilde{l}_\psi(\mathbf{f}_0, Z)) \ge M(i,0) = 2^{i-2}\delta_n^{*2}$.

For the second moment, $e_\psi(\mathbf{f}, \bar{\mathbf{f}}) = e(\mathbf{f}, \bar{\mathbf{f}}) + \frac{1}{2}E[\psi(\mathbf{g}(\mathbf{f}(\mathbf{X}))) \times I(\mathbf{g}(\mathbf{f}(\mathbf{X})) \in (0, \tau))]$ and $e_\psi(\mathbf{f}, \bar{\mathbf{f}}) \le 1$. Thus

$$\begin{aligned} &\frac{1}{2}E\big[\psi\big(\mathbf{g}(\mathbf{f}(\mathbf{X}))\big)I\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y\big) \in (0, \tau)\big)\big] \\ &\quad \le e_\psi(\mathbf{f}, \bar{\mathbf{f}}) \\ &\quad \le (e_\psi(\mathbf{f}, \bar{\mathbf{f}}))^{\alpha/(\alpha+1)}. \end{aligned} \tag{A.7}$$

For any $\mathbf{f} \in A_{i,j}$, $e_\psi(\mathbf{f}, \bar{\mathbf{f}}) \ge 2^{-1}\delta_n^{*2} \ge s_n \ge e_\psi(\mathbf{f}_0, \bar{\mathbf{f}})$, together with (A.5) and (A.7), imply that

$$\begin{aligned} &E[l_\psi(\mathbf{f}, Z) - l_\psi(\mathbf{f}_0, Z)]^2 \\ &\quad \le 2E\big|\operatorname{sign}\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y)\big) - \operatorname{sign}\big(\mathbf{g}_0(\mathbf{f}_0(\mathbf{X}), Y)\big)\big| \\ &\qquad + 2E\big[\psi\big(\mathbf{g}_0(\mathbf{f}_0(\mathbf{X}))\big)I\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y\big) \in (0, \tau)\big)\big] \\ &\qquad + 2E\big[\psi\big(\mathbf{g}(\mathbf{f}(\mathbf{X}))\big)I\big(\mathbf{g}(\mathbf{f}(\mathbf{X}), Y\big) \in (0, \tau)\big)\big] \\ &\quad \le 2\big(c^*\big[e_\psi(\mathbf{f}, \bar{\mathbf{f}})^{\alpha/(\alpha+1)} + e_\psi(\mathbf{f}_0, \bar{\mathbf{f}})^{\alpha/(\alpha+1)}\big]\big) \\ &\qquad + 4\big[e_\psi(\mathbf{f}, \bar{\mathbf{f}})^{\alpha/(\alpha+1)} + e_\psi(\mathbf{f}_0, \bar{\mathbf{f}})^{\alpha/(\alpha+1)}\big] \\ &\quad \le c_3^0\big(e_\psi(\mathbf{f}, \bar{\mathbf{f}})/2\big)^{\alpha/(\alpha+1)}, \end{aligned}$$

with $c_3^0 = 16c_1^{1/(\alpha+1)} + 8$. Therefore, $\sup_{A_{i,j}} E(l_\psi(\mathbf{f}_0, Z) - l_\psi(\mathbf{f}, Z))^2 \le c_3 M(i,j)^{\alpha/(\alpha+1)} = v(i,j)^2$ for $i = 1, \ldots$ and $j = 0, 1, \ldots,$ where $c_3 = 2c_3^0$.

To bound I, note that $I \le I_1 + I_2$, where $I_1 = \sum_{i,j} P^*(\sup_{A_{i,j}} E_n(l_\psi(\mathbf{f}_0, Z) - l_\psi(\mathbf{f}, Z)) \ge M(i,j))$ and $I_2 = \sum_i P^*(\sup_{A_{i,0}} E_n(l_\psi(\mathbf{f}_0, Z) - l_\psi(\mathbf{f}, Z)) \ge M(i,0))$. Thus we can bound I_i separately. Using the fact that $\int_{aM(i,j)}^{v(i,j)} H_B^{1/2}(u, \mathcal{F}^\psi(2^j))\, du/M(i,j)$ is nonincreasing in i and $M(i,j)$, $i = 1, \ldots,$ we have $\int_{aM(i,j)}^{v(i,j)} H_B^{1/2}(u, \mathcal{F}^\psi(2^j))\, du/M(i,j) \le \phi^*(\epsilon_n^*, 2^j)$. The result then follows from the same argument as in the proof of Theorem 3.

Proof of Corollary 2

The result follows from Theorem 4 and its proof.

Lemma 1 (Metric entropy in Example 1). Under the assumptions in Example 1, we have

$$H_B\big(\epsilon, \mathcal{G}(\ell)\big) \le O\big(k^2 \log(k/\epsilon)\big).$$

Proof. Let $(G_1, \ldots, G_k)$ be a classification partition induced by $\mathbf{f}$ and let $G_{j_1 j_2}$ be $\{\mathbf{x}: f_{j_1} - f_{j_2} > 0; \mathbf{x} \in S\}$; $j_1 \ne j_2 \in \{1, \ldots, k\}$. For discussion, we first construct a bracket for $G_{j_1 j_2}$.

Toward this end, we determine d points at which the plane $f_{j_1} - f_{j_2} = 0$ intersects with d out of $d2^{d-1}$ edges of the cube $[0, 1]^d$. For each of these d points, we use a bracket of length ϵ^* to cover, on the edge to which the point belongs. Given an edge, the covering number for this point is no greater than $1/\epsilon^*$. Hence the covering number for the d points on d of $d2^{d-1}$ edges is at most $\binom{d2^{d-1}}{d}(\frac{1}{\epsilon^*})^d$.

After d intersecting points of $f_{j_1} - f_{j_2} = 0$ on the edges of S are covered, we then connect the endpoints of the d brackets to form bracket planes $v_{j_1 j_2} = 0$ and $u_{j_1 j_2} = 0$ such that $\{\mathbf{x}: v_{j_1 j_2} > 0\} \subset \{\mathbf{x}: f_{j_1} - f_{j_2} > 0\} \subset \{\mathbf{x}: u_{j_1 j_2} > 0\}$. Because the longest segment in S has length $\sqrt{d}$ corresponding to the diagonal segment between $(0, \ldots, 0)$ and $(1, \ldots, 1)$, we have $P(\mathbf{x}: v_{j_1 j_2} < 0 < u_{j_1 j_2}) \le (\sqrt{d})^{d-1}\epsilon^*$, because $\mathbf{x}$ is uniformly distributed on S. Consequently, $G^v_{j_1 j_2} \subset G_{j_1 j_2} \subset G^u_{j_1 j_2}$ and $P(G^v_{j_1 j_2} \Delta G^u_{j_1 j_2}) \le (\sqrt{d})^{d-1}\epsilon^*$, where $G^v_{j_1 j_2} = \{\mathbf{x}: v_{j_1 j_2} > 0\}$ and $G^u_{j_1 j_2} = \{\mathbf{x}: u_{j_1 j_2} > 0\}$. Because $G_{j_1} = \bigcap_{j_2} G_{j_1 j_2}$, $G^v_{j_1} \subset G_{j_1} \subset G^u_{j_1}$ and $P(G^v_{j_1} \Delta G^u_{j_1}) \le P(\bigcup_{j_2} G^v_{j_1 j_2} \Delta G^u_{j_1 j_2}) \le (k-1)(\sqrt{d})^{d-1}\epsilon^*$, where $G^v_{j_1} = \bigcap_{j_2} G^v_{j_1 j_2}$ and $G^u_{j_1} = \bigcap_{j_2} G^u_{j_1 j_2}$, $j_1 \ne j_2 \in \{1, \ldots, k\}$.

With $\epsilon = (k-1)(\sqrt{d})^{d-1}\epsilon^*$, $\{(G^v_1, G^u_1), \ldots, (G^v_k, G^u_k)\}$ satisfies $\max_{j_1} P(G^v_{j_1} \Delta G^u_{j_1}) \le \epsilon$, and thus it forms an ϵ-bracketing set for $(G_1, \ldots, G_k)$. Therefore, the ϵ-covering number for all partitions induced by $\mathbf{f}$ is at most $(\binom{d2^{d-1}}{d}(\frac{(k-1)(\sqrt{d})^{d-1}}{\epsilon})^d)^{k(k-1)}$. Because d is a constant, the bracketing metric entropy $H_B(\epsilon, \mathcal{G}(\ell))$ is bounded by $O(k^2 \log(k/\epsilon))$ for any ℓ, yielding the desired result.

Lemma 2 (Metric entropy in Example 2). Under the assumptions in Example 2, we have

$$H_B\big(\epsilon, \mathcal{F}^{\psi}(\ell)\big) \le O\big(k(\log(\ell/\epsilon))^{d+1}\big).$$

Proof. To obtain an upper bound for $H_B(\epsilon, \mathcal{F}^{\psi}(\ell))$, we use the sup-norm entropy bound for a single function set by Zhou (2002), that is, $H_\infty(\epsilon, \mathcal{F}(\ell)) \le O((\log(\ell/\epsilon))^{d+1})$ under the L_∞ metric, $\|g\|_\infty = \sup_{\mathbf{x}\in S} |g(\mathbf{x})|$. Consider an arbitrary function vector $\mathbf{f} = (f_1, \ldots, f_k) \in \mathcal{F}(\ell)$. The metric entropy for all k-dimensional function vectors in $\mathcal{F}(\ell)$ is bounded by $O(k(\log(\ell/\epsilon))^{d+1})$ to cover k functions simultaneously. Let $[f^v_j, f^u_j]$ be an ϵ-bracket for f_j. Then $[f^v_j - f^u_l, f^u_j - f^v_l]$ forms a 2ϵ-bracket for $f_j - f_l$. Denote $g^v_j = \min_{l\in\{1,\ldots,k\}\setminus j}(f^v_j - f^u_l)$ and $g^u_j = \min_{l\in\{1,\ldots,k\}\setminus j}(f^u_j - f^v_l)$. Then $[g^v_j, g^u_j]$ becomes a 2ϵ-bracket for $\mathbf{g}_{\min}(\mathbf{f}, j) = \min_{l\ne j}(f_j - f_l)$. Consequently, $\psi(g^v_j) \ge \psi(\mathbf{g}_{\min}(\mathbf{f}, j)) \ge \psi(g^u_j)$ by the nonincreasing property of ψ function. By (10), we have $|\psi(g^v_j) - \psi(g^u_j)| \le 2D\epsilon$. Because $\mathbf{g}_{\min}(\mathbf{f}, y) = \sum_{j=1}^k I(y = j) \times \mathbf{g}_{\min}(\mathbf{f}, j)$, $\mathbf{g}_{\min}(\mathbf{f}, y) \in [\sum_{j=1}^k I(y = j) g^v_j, \sum_{j=1}^k I(y = j) g^u_j]$ and $|\psi(\sum_{j=1}^k I(y = j) g^v_j) - \psi(\sum_{j=1}^k I(y = j) g^u_j)| \le 2D\epsilon$. Consequently, $[\psi(\sum_{j=1}^k I(y = j) g^u_j(\mathbf{x})) - \psi(\mathbf{g}_0(\mathbf{f}_0(\mathbf{x}), y)), \psi(\sum_{j=1}^k I(y = j) g^v_j(\mathbf{x})) - \psi(\mathbf{g}_0(\mathbf{f}_0(\mathbf{x}), y))]$ forms a bracket of length $2D\epsilon$ for $\psi(\mathbf{g}(\mathbf{f}(\mathbf{x}), y)) - \psi(\mathbf{g}_0(\mathbf{f}_0(\mathbf{x}), y))$. The desired result then follows.

[Received January 2004. Revised June 2005.]

REFERENCES

An, H. L. T., and Tao, P. D. (1997), "Solving a Class of Linearly Constrained Indefinite Quadratic Problems by D.C. Algorithms," *Journal of Global Optimization* 11, 253–285.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2003), "Convexity, Classification, and Risk Bounds," Technical Report 638, University of California, Berkeley, Dept. of Statistics.

Boser, B., Guyon, I., and Vapnik, V. N. (1992), "A Training Algorithm for Optimal Margin Classifiers," in *Proceedings of the Fifth Annual Conference on Computational Learning Theory*, Pittsburgh, PA: ACM Press, pp. 144–152.

Cortes, C., and Vapnik, V. (1995), "Support-Vector Networks," *Machine Learning*, 20, 273–279.

Crammer, K., and Singer, Y. (2001), "On the Algorithmic Implementation of Multiclass Kernel-Based Vector Machines," *Journal of Machine Learning Research*, 2, 265–292.

Guermeur, Y. (2002), "Combining Discriminant Models With New Multiclass SVMS, *Pattern Analysis and Applications*, 5, 168–179.

Lee, Y., Lin, Y., and Wahba, G. (2004), "Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data," *Journal of the American Statistical Association*, 99, 67–81.

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R., and Klein, B. (2000), "Smoothing Spline ANOVA Models for Large Data Sets With Bernoulli Observations and the Randomized GACV," *The Annals of Statistics*, 28, 1570–1600.

Lin, Y. (2000), "Some Asymptotic Properties of the Support Vector Machine," Technical Report 1029, University of Wisconsin–Madison, Dept. of Statistics.

——— (2002), "Support Vector Machines and the Bayes Rule in Classification," *Data Mining and Knowledge Discovery*, 6, 259–275.

Liu, S., Shen, X., and Wong, W. (2005), "Computational Developments of ψ-Learning," in *Proceedings of the Fifth SIAM–ASA International Conference on Data Mining*, Newport, CA: SIAM, pp. 1–12.

Liu, Y., Shen, X., and Doss, H. (2005), "Multicategory ψ-Learning and Support Vector Machine: Computational Tools," *Journal of Computational Graphical Statistics*, 14, 219–236.

Liu, Y., and Wu, Y. (2005), "Optimizing ψ-Learning via Mixed Integer Programming," *Statistica Sinica*, to appear.

Marron, J. S., and Todd, M. J. (2002), "Distance-Weighted Discrimination," Technical Report 1339, Cornell University, School of Operations Research and Industrial Engineering.

Mercer, J. (1909), "Functions of Positive and Negative Type and Their Connection With the Theory of Integral Equations," *Philosophical Transactions of the Royal Society of London*, Ser. A, 209, 415–446.

Shen, X. (1998), "On the Method of Penalization," *Statistica Sinica*, 8, 337–357.

Shen, X., Tseng, G. C., Zhang, X., and Wong, W. H. (2003), "On ψ-Learning," *Journal of the American Statistical Association*, 98, 724–734.

Shen, X., and Wong, W. H. (1994), "Convergence Rate of Sieve Estimates," *The Annals of Statistics*, 22, 580–615.

Steinwart, I. (2001), "On the Influence of the Kernel on the Consistency of Support Vector Machines," *Journal of Machine Learning Research*, 2, 67–93.

Tsybakov, A. B. (2004), "Optimal Aggregation of Classifiers in Statistical Learning," *The Annals of Statistics*, 32, 135–166.

Wahba, G. (1998), "Support Vector Machines, Reproducing Kernel Hilbert Spaces, and Randomized GACV," in *Advances in Kernel Methods: Support Vector Learning*, eds. B. Schölkopf, C. J. C. Burges, and A. J. Smola, Cambridge, MA: MIT Press, pp. 125–143.

Weston, J., and Watkins, C. (1999), "Support Vector Machines for Multi-Class Pattern Recognition," in *Proceedings of the Seventh European Symposium on Artificial Neural Networks*, Bruges, Belgium: D-Fact. Public., pp. 219–224.

Zhang, T. (2004a), "Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization," *The Annals of Statistics*, 32, 56–85.

——— (2004b), "Statistical Analysis of Some Multi-Category Large Margin Classification Methods," *Journal of Machine Learning Research*, 5, 1225–1251.

Zhou, D. X. (2002), "The Covering Number in Learning Theory," *Journal of Complexity*, 18, 739–767.

Zhu, J., and Hastie, T. (2005), "Kernel Logistic Regression and the Import Vector Machine," *Journal of Computational and Graphical Statistics*, 14, 185–205.

Zhu, J., Hastie, T., Rosset, S., and Tibshirani, R. (2003), "1-Norm Support Vector Machines," *Neural Information Processing Systems*, 16, available at *http://books.nips.cc/nips16.html*.