Journal of Machine Learning Research 5 (2004) 1225–1251 Submitted 4/2004; Published 10/04

Statistical Analysis of Some Multi-Category Large Margin Classification Methods

Tong Zhang TZHANG@WATSON.IBM.COM
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA

Editor: Bernhard Schölkopf

Abstract

The purpose of this paper is to investigate statistical properties of risk minimization based multi-category classification methods. These methods can be considered as natural extensions of binary large margin classification. We establish conditions that guarantee the consistency of classifiers obtained in the risk minimization framework with respect to the classification error. Examples are provided for four specific forms of the general formulation, which extend a number of known methods. Using these examples, we show that some risk minimization formulations can also be used to obtain conditional probability estimates for the underlying problem. Such conditional probability information can be useful for statistical inferencing tasks beyond classification.

1. Motivation

Consider a binary classification problem where we want to predict label $y \in \{\pm 1\}$ based on observation x. One of the most significant achievements for binary classification in machine learning is the invention of large margin methods, which include support vector machines and boosting algorithms.

Based on a set of training samples $(X_1,Y_1),\ldots,(X_n,Y_n)$, a large margin binary classification algorithm produces a decision function $\hat{f}(\cdot)$ by minimizing an empirical loss function that is often a convex upper bound of the binary classification error function. Given $\hat{f}(\cdot)$, the binary decision rule is to predict $y = 1$ if $\hat{f}(x) \geq 0$, and to predict $y = -1$ otherwise (the decision rule at $\hat{f}(x) = 0$ is not important).

In the literature, the following form of large margin binary classification is often encountered: we minimize the empirical risk associated with a convex function ϕ in a pre-chosen function class C_n that may depend on the sample size:

$$\hat{f}(\cdot) = \arg\min_{f(\cdot)\in C_n} \frac{1}{n}\sum_{i=1}^{n} \phi(f(X_i)Y_i). \tag{1}$$

Originally such a scheme was regarded as a compromise to avoid computational difficulties associated with direct classification error minimization, which often leads to an NP-hard problem. Some recent works in the statistical literature argued that such methods could be used to obtain conditional probability estimates. For example, see Friedman et al. (2000), Lin (2002), Schapire and Singer (1999), Zhang (2004), Steinwart (2003) for related studies. This point of view allows people to show the consistency of various large margin methods: that is, in the large sample limit, the obtained classifiers achieve the optimal Bayes error rate. For example, see Bartlett et al. (2003), Jiang (2004), Lugosi and Vayatis (2004), Mannor et al. (2003), Steinwart (2002, 2004), Zhang

(2004). The consistency of a learning method is certainly a very desirable property, and one may argue that a good classification method should at least be consistent in the large sample limit.

Although statistical properties of binary classification algorithms based on the risk minimization formulation (1) are quite well-understood due to many recent works such as those mentioned above, there are much fewer studies on risk minimization based multi-category problems which generalizes the binary large margin method (1). The complexity of possible generalizations may be one reason. Another reason may be that one can always estimate the conditional probability for a multi-category problem using the binary classification formulation (1) for each category, and then pick the category with the highest estimated conditional probability (or score).[1]

It is still useful to understand whether there are more natural alternatives, and what risk minimization formulations that generalize (1) can be used to yield consistent classifiers in the large sample limit. An important step toward this direction has recently been taken by Lee et al. (2004), where the authors proposed a multi-category extension of the support vector machine that is infinite-sample Bayes consistent (Fisher consistent). The purpose of this paper is to generalize their study so as to include a much wider class of risk minimization formulations that can lead to consistent classifiers in the infinite-sample limit. Moreover, combined with some relatively simple generalization analysis for kernel methods, we are able to show that with appropriately chosen regularization conditions, classifiers obtained from certain formulations can approach the optimal Bayes error in the large sample limit.

Theoretical analysis of risk minimization based multi-category large margin methods have started to draw more attention recently. For example, in Desyatnikov and Meir (2003), learning bounds for some multi-category convex risk minimization methods were obtained, although the authors did not study possible choices of Bayes consistent formulations. A related study can be found in Liu and Shen (2004), but again only for special formulations.

Although this paper studies a number of multi-category classification methods, we shall not try to argue which one is better practically, or to compare different formulations experimentally. One reason is that some methods investigated in this paper were originally proposed by different researchers, who have much more practical experience with the corresponding algorithms. Due to the scope of this paper, it is simply impossible for us to include a comprehensive empirical study without overlooking some engineering tricks. Casual experimental comparisons can lead to misleading conclusions. Therefore in this paper we only focus on asymptotic theoretical analysis. Although our analysis provides useful statistical insights (especially asymptotically), the performance of a learning algorithm may also be affected by factors which we do not considered here, especially for small-sample problems. We shall refer the readers to Rifkin and Klautau (2004) for a recent experimental study on some multi-category classification algorithms, although the issue of which algorithm may have better practical performance (and under what circumstances) is far from resolved.

We organize the paper as follows. Section 2 introduces the multi-category classification problem, and a general risk minimization based approach. In Section 3, we give conditions that guarantee the infinite-sample consistency of the risk minimization formulation. In Section 4, examples of the general formulation, which extend some existing methods in the literature, will be presented. We shall study their properties such as the associated statistical models and conditions that en-

1. This approach is often called one-versus-all in machine learning. Another main approach is to encode a multi-category classification problem into binary classification sub-problems. The consistency of such encoding schemes cannot be analyzed in our framework, and we shall not discuss them.

sure the infinite-sample consistency (ISC) of the resulting risk minimization estimators. Section 5 contains a relatively simple generalization analysis (which is not necessarily tight) for kernel multi-categorization methods. Our purpose is to demonstrate that with appropriately chosen regularization conditions, classifiers obtained from ISC risk minimization formulations can approach the optimal Bayes classifier in the large sample limit. Concluding remarks will be presented in Section 6.

2. Multi-Category Classification

We consider the following K-category classification problem: given an input vector x, we would like to predict its corresponding label $y \in \{1,\ldots,K\}$. Let $p(x)$ be a predictor of y which is a function of x. In the machine learning framework, the quality of this predictor can be measured by a loss function $L(p(x),y)$, and the data (X,Y) are drawn from an unknown underlying distribution D.

Given a set of training samples $(X_1,Y_1),\ldots,(X_n,Y_n)$, randomly drawn from D, our goal is to find a predictor $\hat{p}(x)$ so that the expected true loss of $\hat{p}$ given below is as small as possible:

$$\mathbf{E}_{X,Y} L(\hat{p}(X),Y),$$

where we use $\mathbf{E}_{X,Y}$ to denote the expectation with respect to the true (but unknown) underlying distribution D.

The loss function $L(p,y)$ can be regarded as a $K \times K$ cost matrix. In this paper, we are mainly interested in the simple but also the most important case of $0-1$ classification loss: we have a loss of 0 for correct prediction, and loss of 1 for incorrect prediction. We consider a slightly more general family of cost matrices, where the classification errors for different classes are penalized differently:

$$L(p,y) = \begin{cases} 0 & \text{if } p = y \\ a_y & \text{if } p \neq y, \end{cases} \tag{2}$$

where $a_y > 0$ ($y = 1,\ldots,K$) are K pre-defined positive numbers. If we let $a_y = 1$ ($y = 1,\ldots,K$), then we have the standard classification error. The more general cost-sensitive classification error in (2) is useful for many applications. For example, in some medical diagnosis applications, classifying a patient with cancer to the *no-cancer* category is much worse than classifying a patient without cancer to the *possible cancer* category (since in the latter case, a more thorough test can be performed to produce a more definite diagnosis).

Let $p(X) \in \{1,\ldots,K\}$ be a classifier. Its classification error under (2) is given by

$$\ell(p(\cdot)) := \mathbf{E}_X \sum_{c=1, c\neq p(X)}^{K} a_c P(Y = c|X). \tag{3}$$

If we know the conditional density $P(Y = c|X)$, then the optimal classification rule with the minimum loss in (3), often referred to as the *Bayes rule*, is given by

$$p_*(X) = \max_{c\in\{1,2,\ldots,K\}} a_c P(Y = c|X). \tag{4}$$

In binary classification with 0-1 classification error, the class rule can be obtained using the sign of a real-valued decision function. This can be generalized to K class classification problem as

follows: we consider K decision functions $f_c(x)$ where $c = 1,\ldots,K$ and we predict the label y of x as

$$T(\mathbf{f}(x)) := \arg\max_{c\in\{1,\ldots,K\}} f_c(x), \tag{5}$$

where we denote by $\mathbf{f}(x)$ the vector function $\mathbf{f}(x) = [f_1(x),\ldots,f_K(x)]$. In the following, we use bold symbols such as $\mathbf{f}$ to denote vectors, and $\mathbf{f}_c$ to denote its c-th component. We also use $\mathbf{f}(\cdot)$ to denote vector functions. If two or more components of $\mathbf{f}$ achieve the same maximum value, then we may choose any of them as $T(\mathbf{f})$. In this framework, $\mathbf{f}_c(x)$ is often regarded as a scoring function for category c that is correlated with how likely x belongs to category c (compared with the remaining $k-1$ categories).

Note that only the relative strength of the component $\mathbf{f}_c$ compared with the alternatives $\mathbf{f}_k$ ($k \neq c$) is important. In particular, the decision rule given in (5) does not change when we add the same numerical quantity to each component of $\mathbf{f}(x)$. This allows us to impose one constraint on the vector $\mathbf{f}(x)$ which decreases the degree of freedom K of the K-component vector $\mathbf{f}(x)$ to $K-1$. For example, in the binary classification case, we can enforce $\mathbf{f}_1(x)+\mathbf{f}_2(x) = 0$, and hence $f(x)$ can be represented as $[\mathbf{f}_1(x),-\mathbf{f}_1(x)]$. The decision rule in (5), which compares $\mathbf{f}_1(x) \geq \mathbf{f}_2(x)$, is equivalent to $\mathbf{f}_1(x) \geq 0$. This leads to the binary classification rule mentioned in the introduction.

In the multi-category case, one may also interpret the possible constraint on the vector function $\mathbf{f}(\cdot)$, which reduces its degree of freedom from K to $K-1$, based on the following observation. In many cases, we seek $\mathbf{f}_c(x)$ as a function of $p(Y = c|x)$. Since we have a constraint $\sum_{c=1}^{K} p(Y = c|x) = 1$ (implying that the degree of freedom for $p(Y = c|x)$ is $K-1$), the degree of freedom for f is also $K-1$ (instead of K). However, we shall point out that in the algorithms we formulate below, we may either enforce such a constraint that reduces the degree of freedom of f, or we do not impose any constraint, which keeps the degree of freedom of f to be K. The advantage of the latter is that it allows the computation of each $\mathbf{f}_c(x)$ to be decoupled. It is thus much simpler both conceptually and numerically. Moreover, it directly handles multiple-label problems where we may assign each x to multiple labels of $y \in \{1,\ldots,K\}$. In this scenario, we do not have a constraint.

In this paper, we consider an empirical risk minimization method to solve a multi-category problem, which is of the following general form:

$$\hat{\mathbf{f}}(\cdot) = \arg\min_{\mathbf{f}(\cdot)\in C_n} \frac{1}{n}\sum_{i=1}^{n} \Psi_{Y_i}(\mathbf{f}(X_i)), \tag{6}$$

where $\mathbf{f}(\cdot)$ is a K-component vector function, and C_n is a vector function class. Each $\Psi_Y(\cdot) : R^K \to R$ (indexed by class label $Y \in \{1,\ldots,K\}$) is a real-valued function that takes a K-component vector as its parameter. As we shall see later, this method is a natural generalization of the binary classification method (1). Note that one may consider an even more general form with $\Psi_Y(\mathbf{f}(X))$ replaced by $\Psi_Y(\mathbf{f}(X),X)$, which we don't study in this paper.

The general formulation (6) covers many traditional and newly proposed multi-category classification methods. Examples will be given in Section 4. Some of them such as some multi-category extensions of support vector machines are directly motivated by margin maximization (in the separable case). In general, as we shall see in Section 4, the function $\Psi_Y(\mathbf{f})$ should be chosen such that it favors a vector predictor $\mathbf{f}$ with the component $\mathbf{f}_Y$ corresponding to the observed class label Y larger than the alternatives $\mathbf{f}_k$ for $k \neq Y$. In this sense, it encourages the correct classification rule in (5) by implicitly maximizes the difference of $\mathbf{f}_Y$ and the remaining components $\mathbf{f}_k$ ($k \neq Y$). One may

interpret this effect as soft margin-maximization, and hence one may consider learning algorithms based on (6) generally as multi-category large margin methods.

Given the estimator $\hat{\mathbf{f}}(\cdot)$ from (6), the classification rule is based on (5) or some variants which we shall discuss later. The main purpose of the paper is to investigate the following two issues:

- Consistency: whether the classification error $\ell(\hat{\mathbf{f}}(\cdot))$ converges to $\ell(p_*(\cdot))$ where $p_*(\cdot)$ is the Bayes rule defined in (4).
- Probability model: the relationship of $\hat{\mathbf{f}}(X)$ and the conditional probability vector $[P(Y = c|X)]_{c=1,\ldots,K}$.

3. Approximation Estimation Decomposition

From the standard learning theory, one can expect that with appropriately chosen C_n, the solution $\hat{\mathbf{f}}(\cdot)$ of (6) approximately minimizes the true Ψ risk $\mathbf{E}_{X,Y}\Psi_Y(\hat{\mathbf{f}}(X))$ with respect to the unknown underlying distribution D within the vector function class C_n. The true risk of a vector function $\mathbf{f}(\cdot)$ can be rewritten as

$$\mathbf{E}_{X,Y}\Psi_Y(\mathbf{f}(X)) = \mathbf{E}_X W(\mathbf{P}(\cdot|X), \mathbf{f}(X)), \tag{7}$$

where $\mathbf{P}(\cdot|X) = [P(Y = 1|X), \ldots, P(Y = K|X)]$ is the conditional probability, and

$$W(\mathbf{q}, \mathbf{f}) := \sum_{c=1}^{K} \mathbf{q}_c \Psi_c(\mathbf{f}). \tag{8}$$

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

$$\Lambda_K := \left\{ \mathbf{q} \in R^K : \sum_{c=1}^{K} \mathbf{q}_c = 1,\ \mathbf{q}_c \geq 0 \right\}.$$

The vector argument $\mathbf{q}$ of $W(\mathbf{q}, \mathbf{f})$ represents the conditional probability vector evaluated at some point x; the argument $\mathbf{f}$ represents the value of our vector predictor evaluated at the same point x. Intuitively, $W(\mathbf{q}, \mathbf{f})$ is the point-wise true loss of $\mathbf{f}$ at some x, with respect to the conditional probability distribution $\mathbf{q} = [P(Y = \cdot|X = x)]$.

In order to understand the large sample behavior of the algorithm based on solving (6), we first need to understand the behavior of a vector function $\mathbf{f}(\cdot)$ that approximately minimizes $\mathbf{E}_{X,Y}\Psi_Y(\mathbf{f}(X))$. We introduce the following definition. The property has also been referred to as *classification calibrated* in Bartlett et al. (2003) or *Fisher consistent* in Lin (2002). In this paper, we explicitly call it as *infinite-sample consistent*.

Definition 1 *Consider $[\Psi_c(\mathbf{f})]$ in (7). We say that the formulation is infinite-sample consistent (ISC) on a set $\Omega \subseteq R^K$ with respect to the classification error loss (3), if the following conditions hold:*

- *For each c, $\Psi_c(\cdot) : \Omega \to R$ is bounded below and continuous.*
- *$\forall \mathbf{q} \in \Lambda_K$ and $c \in \{1, \ldots, K\}$ such that $a_c \mathbf{q}_c < \sup_k a_k \mathbf{q}_k$, we have*

$$W^*(\mathbf{q}) := \inf_{\mathbf{f} \in \Omega} W(\mathbf{q}, \mathbf{f}) < \inf \left\{ W(\mathbf{q}, \mathbf{f}) : \mathbf{f} \in \Omega, \mathbf{f}_c = \sup_k \mathbf{f}_k \right\}.$$

Remark 2 *Among the two conditions, the second is more essential. It says that (point-wisely) for each conditional probability vector $\mathbf{q} \in \Lambda_K$, an exact optimal solution of $W(\mathbf{q},\cdot)$ leads to a Bayes rule with respect to the classification error defined in (3). That is, the exact minimization of (7) leads to the exact minimization of classification error. This condition is clearly necessary for consistency. The first condition (continuity) is needed to show that point-wisely, an approximate (instead of exact) minimizer of (7) also approximately minimizes the classification error.*

The following result relates the approximate minimization of the Ψ risk to the approximate minimization of classification error. The proof is left to Appendix B. A more general but also more abstract theory is presented in Appendix A.

Theorem 3 *Let $\mathcal{B}$ be the set of all vector Borel measurable functions (with respect to some underlying topology on the input space) which take values in R^K. For $\Omega \subset R^K$, let $\mathcal{B}_\Omega = \{\mathbf{f} \in \mathcal{B} : \forall x, \mathbf{f}(x) \in \Omega\}$. If $[\Psi_c(\cdot)]$ is ISC on Ω with respect to (3), then $\forall \varepsilon_1 > 0$, $\exists \varepsilon_2 > 0$ such that for all underlying Borel probability measurable D, and $\mathbf{f}(\cdot) \in \mathcal{B}_\Omega$,*

$$\mathbf{E}_{X,Y}\Psi_Y(\mathbf{f}(X)) \leq \inf_{\mathbf{f}' \in \mathcal{B}_\Omega} \mathbf{E}_{X,y}\Psi_Y(\mathbf{f}'(X)) + \varepsilon_2$$

implies

$$\ell(T(\mathbf{f}(\cdot))) \leq \ell_B + \varepsilon_1,$$

$T(\cdot)$ is defined in (5), and ℓ_B is the optimal Bayes error: $\ell_B = \ell(p_*(\cdot))$, with p_* given in (4).

Based on the above theorem, an ISC risk minimization formulation is suitable for multi-category classification problems. The classifier obtained from minimizing (6) can approach the Bayes error rate if we can show that with appropriately chosen function class C_n, approximate minimization of (6) implies approximate minimization of (7). Learning bounds of this kind have been very well-studied in statistics and machine learning. For example, for binary classification, such bounds can be found in Blanchard et al. (2003), Bartlett et al. (2003), Jiang (2004), Lugosi and Vayatis (2004), Mannor et al. (2003), Steinwart (2002, 2004), Zhang (2004), where they were used to prove the consistency of various large margin classification methods. In order to achieve consistency, it is also necessary to take a sequence of function classes C_n (typically, one takes a sequence $C_1 \subset C_2 \subset \cdots \subset \mathcal{B}_\Omega$) such that $\cup_n C_n$ is dense (e.g. with respect to the uniform-norm topology) in $\mathcal{B}_\Omega$. This method, widely studied in the statistics literature, is often referred to as *the method of sieves* (for example, see Chapter 10 of van de Geer, 2000, and references therein). It is also closely related to the structural risk minimization method of Vapnik (1998). The set C_n has the effect of regularization, which ensures that for large n, $\mathbf{E}_{X,Y}\Psi_Y(\hat{\mathbf{f}}(X)) \approx \inf_{\mathbf{f}(\cdot)\in C_n} \mathbf{E}_{X,Y}\Psi_Y(\mathbf{f}(X))$. It follows that as $n \to \infty$, $\mathbf{E}_{X,Y}\Psi_Y(\hat{\mathbf{f}}(X)) \xrightarrow{P} \inf_{\mathbf{f}(\cdot)\in\mathcal{B}_\Omega} \mathbf{E}_{X,Y}\Psi_Y(\mathbf{f}(X))$. Theorem 3 then implies that $\ell(T(\hat{\mathbf{f}}(\cdot))) \xrightarrow{P} \ell_B$. The above idea, although intuitively clear, is not rigorously stated at this point. A rigorous treatment can be found in Section 5.

We can see that there are two types of errors in this framework. The first type of error, often referred to as *approximation error*, measures how close we are from the optimal Bayes error when we approximately minimize the true risk with respect to the surrogate loss function Ψ in (7). Theorem 3 implies that the approximation error goes to zero when we approximately minimize (7). The second type of error, often referred to as *estimation error*, is how close we are from achieving the minimum of the true Ψ risk in (7), when we obtain a classifier based on the empirical minimization

of (6). The overall statistical error of the risk minimization based classification method (6) is given by the combination of approximation error and estimation error.

Before studying learning bounds that relate approximate minimization of (6) to the approximate minimization of (7), we provide examples of Ψ that lead to ISC formulations. We pay special attention to the case that each $\Psi_c(\mathbf{f})$ is a convex function of $\mathbf{f}$, so that the resulting formulation becomes computationally more tractable (assuming we also use convex function classes C_n).

4. Multi-Category Classification Formulations

We give some examples of ISC multi-category classification formulations. They are motivated from methods proposed in the literature, and will be extended in our framework.

The following simple result says that an ISC formulation for an arbitrary loss of the form (2) can be obtained from an ISC formulation of any particular loss in that family.

Proposition 4 *Assume $[\Psi_c(\mathbf{f})]$ is ISC on $\Omega \subset R^K$ with respect to (3) with $a_c = a'_c$ ($c = 1,\ldots,K$). Then $\forall$ positive numbers a''_c ($c = 1,\ldots,K$), $[\Psi_c(\mathbf{f})a''_c/a'_c]$ is ISC on $\Omega \subset R^K$ with respect to (3) with $a_c = a''_c$ ($c = 1,\ldots,K$).*

Proof The first condition of ISC holds automatically. Now we shall check the second condition. For all $\mathbf{q} \in \Lambda_K$, we define $\mathbf{q}'$ as $\mathbf{q}'_c = \mathbf{q}_c a''_c/a'_c$. Therefore

$$\sum_{c=1}^{K} \mathbf{q}_c \frac{\Psi_c(\mathbf{f})a''_c}{a'_c} = \sum_{c=1}^{K} \mathbf{q}'_c \Psi_c(\mathbf{f}).$$

The ISC condition of $[\Psi_c(\mathbf{f})]$ with respect to $\{a'_c\}$ implies

$$\inf\left\{\sum_{c=1}^{K} \mathbf{q}_c \frac{\Psi_c(\mathbf{f})a''_c}{a'_c} : \mathbf{f} \in \Omega, \mathbf{f}_c = \sup_k \mathbf{f}_k\right\} > \inf_{\mathbf{f}\in\Omega} \sum_{c=1}^{K} \mathbf{q}_c \frac{\Psi_c(\mathbf{f})a''_c}{a'_c}$$

for all c such that $a'_c\mathbf{q}'_c < \sup_k a'_k\mathbf{q}'_k$. That is, for all c such that $a''_c\mathbf{q}_c < \sup_k a''_k\mathbf{q}_k$. This gives the second condition of ISC. ■

Due to the above result, for notational simplicity, we shall focus on the 0-1 classification error in this section, with $a_c = 1$ in (3):

$$\ell(p(\cdot)) = \mathbf{E}_X \sum_{c=1, c\neq p(X)}^{K} P(Y = c|X) = 1 - \mathbf{E}_X P(Y = p(\cdot)|X). \tag{9}$$

4.1 Pairwise Comparison Method

This model is motivated from the multi-class support vector machine in Weston and Watkins (1998).[2] Here we consider a more general formulation with the following choice of Ψ:

$$\Psi_c(\mathbf{f}) = \sum_{k=1}^{K} \phi(\mathbf{f}_c - \mathbf{f}_k), \tag{10}$$

2. According to Schölkopf and Smola. (2002), page 213, an identical method was proposed independently by Blanz et al. (1995) three years earlier in a talk given at AT&T.

where ϕ is an appropriately chosen real-valued function. The choice in Weston and Watkins (1998) is the hinge loss for the SVM formulation: $\phi(p) = (1-p)_+$.

Typically we choose a decreasing function ϕ in (10). Assume that we observe a datum X with its label Y. The intuition behind (10) is to favor a large value $\mathbf{f}_Y(X) - \mathbf{f}_k(X)$ for $k \neq Y$, which encourages the correct classification rule. This approach has some attractive features. Since it makes pairwise comparisons, the penalty term $\phi(\mathbf{f}_c - \mathbf{f}_k)$ can be adjusted in a pairwise fashion. This can be useful for some cost-sensitive classification problems that are more general than the particular form we consider in (3). With a differentiable ϕ (thus excludes the SVM hinge loss), this method also has the very desirable property of *order preserving*, which we state below.

Theorem 5 *Consider the formulation in (10). Let $\phi(\cdot) : R \to R$ be a non-increasing function such that $\phi(z) < \phi(-z)$ for all $z > 0$. Consider any $\mathbf{q} \in \Lambda_K$ and $\mathbf{f}$ such that $W(\mathbf{q},\mathbf{f}) = W^*(\mathbf{q})$. If $\mathbf{q}_i < \mathbf{q}_j$, we have $\mathbf{f}_i \leq \mathbf{f}_j$. Moreover, if $\phi(\cdot)$ is differentiable and $\phi'(0) < 0$, then we have $\mathbf{f}_i < \mathbf{f}_j$.*

Proof We can take $i = 1$ and $j = 2$. Let $\mathbf{f}' = \mathbf{f}_k$ when $k > 2$, $\mathbf{f}'_1 = \mathbf{f}_2$, and $\mathbf{f}'_2 = \mathbf{f}_1$. We now prove the first part by contradiction. Assume $\mathbf{f}_1 > \mathbf{f}_2$. We have

$$\begin{aligned}
&W(\mathbf{q},\mathbf{f}') - W(\mathbf{q},\mathbf{f}) \\
=&(\mathbf{q}_2 - \mathbf{q}_1)\left[\phi(\mathbf{f}_1 - \mathbf{f}_2) - \phi(\mathbf{f}_2 - \mathbf{f}_1) + \sum_{k>2}(\phi(\mathbf{f}_1 - \mathbf{f}_k) - \phi(\mathbf{f}_2 - \mathbf{f}_k))\right] \\
<&(\mathbf{q}_2 - \mathbf{q}_1)[0+0] = 0.
\end{aligned}$$

This is a contradiction to the optimality of $\mathbf{f}$. Therefore we must have $\mathbf{f}_1 \leq \mathbf{f}_2$, which proves the first part.

Now we assume in addition that $\phi(\cdot)$ is differentiable. Then at the optimal solution, we have the first order condition $\frac{\partial}{\partial \mathbf{f}_c} W(\mathbf{q},\mathbf{f}) = 0$:

$$\mathbf{q}_c \sum_{k=1}^{K} \phi'(\mathbf{f}_c - \mathbf{f}_k) = \sum_{k=1}^{K} \mathbf{q}_k \phi'(\mathbf{f}_k - \mathbf{f}_c).$$

Again, we prove the second part by contradiction. To this end let us assume $\mathbf{f}_1 = \mathbf{f}_2 = f$, then the above equality implies that

$$\mathbf{q}_1 \sum_{k=1}^{K} \phi'(f - \mathbf{f}_k) = \mathbf{q}_2 \sum_{k=1}^{K} \phi'(f - \mathbf{f}_k).$$

This is not possible since $\sum_{k=1}^{K} \phi'(f - \mathbf{f}_k) \leq 2\phi'(0) < 0$. ■

Note that for functions that are not differentiable, even if $\mathbf{q}_1 < \mathbf{q}_2$, we may still allow $\mathbf{f}_1 = \mathbf{f}_2$ at an optimal solution. Moreover, it is possible that the formulation is not ISC. We provide such a counter-example for the hinge loss in Appendix C. However, for differentiable functions, the method is infinite-sample consistent.

Theorem 6 *Let $\phi(\cdot) : R \to R$ be a differentiable non-negative and non-increasing function such that $\phi'(0) < 0$. Then the formulation (10) is ISC on $\Omega = R^K$ with respect to (9).*

Proof Consider $\mathbf{q} \in \Lambda_K$, and assume that $\mathbf{q}_1 < \mathbf{q}_2$. We show that

$$\inf\{W(\mathbf{q},\mathbf{f}) : \mathbf{f} \in \Omega, \mathbf{f}_1 \geq \mathbf{f}_2\} > W^*(\mathbf{q}).$$

This will imply ISC. We again prove by contradiction. If the claim is not true, then we can find sequences $\mathbf{f}^{(m)}$ such that $0 = \mathbf{f}_1^{(m)} \geq \mathbf{f}_2^{(m)}$ and $\lim_m W(\mathbf{q},\mathbf{f}^{(m)}) = W^*(\mathbf{q})$. We can further select subsequences such that for each pair i and j, $\mathbf{f}_i^{(m)} - \mathbf{f}_j^{(m)}$ converges (may converge to $\pm\infty$). This gives a limiting vector $\mathbf{f}$, with properly defined $\mathbf{f}_i - \mathbf{f}_j$ even when either $\mathbf{f}_i$ or $\mathbf{f}_j$ is $\pm\infty$. It follows from the assumption that $W(\mathbf{q},\mathbf{f}) = W^*(\mathbf{q})$ and $0 = \mathbf{f}_1 \geq \mathbf{f}_2$. However, this violates Theorem 5 (with trivial modification of the proof to handle the infinity-case), which asserts that $\mathbf{f}_1 < \mathbf{f}_2$. ■

A method closely related to (10) is to employ the following choice of Ψ (see Crammer and Singer, 2001):

$$\Psi_c(\mathbf{f}) = \phi(\mathbf{f}_c - \sup_{k \neq c} \mathbf{f}_k). \tag{11}$$

However, for convex ϕ, this method is usually not infinite-sample consistent. To see this, we assume that ϕ is a convex decreasing function and $\mathbf{q}_1 \geq \mathbf{q}_2 \cdots \geq \mathbf{q}_K$. After some simple algebra, we may choose $\mathbf{f}_1 \geq \mathbf{f}_2 = \cdots = \mathbf{f}_K$, and the corresponding $W(\mathbf{q},\mathbf{f}) = \mathbf{q}_1\phi(\mathbf{f}_1 - \mathbf{f}_2) - \sum_{k=2}^K \mathbf{q}_k\phi(\mathbf{f}_2 - \mathbf{f}_1)$. This means that unless $\mathbf{q}_1 > 0.5$, we can choose $\mathbf{f}_1 = \mathbf{f}_2$ to achieve the optimal value.

It is also worth mentioning that the formulation in (11) has been applied successfully in many practical applications. This may not be surprising since in many practical problems, the most important scenario is when the true label can be predicted relatively accurately. In such case (more precisely, when $\sup_k \mathbf{q}_k > 0.5$), the method is well behaved (ISC). The same reason is also why one may often successfully use (10) with the SVM hinge loss in practical problems, although from Appendix C, we know that the resulting classification method can be inconsistent. However, the analysis given in this section is still useful for the purpose of understanding the limitations of these methods.

4.2 Constrained Comparison Method

As pointed out, one may impose constraints on possible choices of $\mathbf{f}$. In this section, we consider another direct extension of binary large-margin method (1) to multi-category case. The choice given below is motivated by Lee et al. (2004), where an extension of SVM was proposed. For simplicity, we will consider linear equality constraint only:

$$\Psi_c(\mathbf{f}) = \sum_{k=1,k\neq c}^{K} \phi(-\mathbf{f}_k), \qquad \text{s.t.} \quad \mathbf{f} \in \Omega, \tag{12}$$

where we define Ω as

$$\Omega = \left\{ \mathbf{f} \in R^K : \sum_{k=1}^K \mathbf{f}_k = 0 \right\}.$$

Similar to the pairwise comparison model, if we choose a decreasing function ϕ in (10), then this formulation also encourages the correct classification rule. If we observe a datum X with its label Y, then the formulation favors small $\mathbf{f}_k(X)$ for all $k \neq Y$. Due to the sum to zero constraint, this implies a large $\mathbf{f}_Y(X)$.

We may interpret the added constraint in (12) as a restriction on the function class C_n in (6) such that every $\mathbf{f} \in C_n$ satisfies the constraint. Note that with $K = 2$, this leads to the standard binary large margin method.

Using (12), the conditional true Ψ risk (8) can be written as

$$W(\mathbf{q}, \mathbf{f}) = \sum_{c=1}^{K} (1 - \mathbf{q}_c)\phi(-\mathbf{f}_c), \quad \text{s.t. } \mathbf{f} \in \Omega. \tag{13}$$

Similar to the pairwise comparison model, for certain choices of function ϕ, this formulation has the desirable order preserving property.

Theorem 7 *Consider the formulation in (12), and assume that ϕ is strictly decreasing. Consider any $\mathbf{q} \in \Lambda_K$ and $\mathbf{f} \in \Omega$ such that $W(\mathbf{q}, \mathbf{f}) = W^*(\mathbf{q})$. If $\mathbf{q}_i < \mathbf{q}_j$, we have $\mathbf{f}_i \le \mathbf{f}_j$. Moreover, if ϕ is strictly convex and differentiable, then $\mathbf{f}_i < \mathbf{f}_j$.*

Proof The proof is rather straight forward. Let $i = 1$ and $j = 2$. Also let $\mathbf{f}'_k = \mathbf{f}_k$ when $k > 2$, $\mathbf{f}'_1 = \mathbf{f}_2$, and $\mathbf{f}'_2 = \mathbf{f}_1$. From $W(\mathbf{q}, \mathbf{f}') \ge W(\mathbf{q}, \mathbf{f})$, we obtain $(\mathbf{q}_1 - \mathbf{q}_2)(\phi(-\mathbf{f}_1) - \phi(-\mathbf{f}_2)) \ge 0$. This implies that $\phi(-\mathbf{f}_2) \ge \phi(-\mathbf{f}_1)$. Therefore $\mathbf{f}_1 \le \mathbf{f}_2$.

If ϕ is also differentiable, then using the Lagrangian multiplier method for the constraint $\sum_{c=1}^{K} \mathbf{f}_c = 0$, and differentiate at the optimal solution, we have $(1 - \mathbf{q}_1)\phi'(-\mathbf{f}_1) = (1 - \mathbf{q}_2)\phi'(-\mathbf{f}_2) = \lambda < 0$, where λ is the Lagrangian multiplier. The assumption $1 - \mathbf{q}_1 > 1 - \mathbf{q}_2$ implies that $\phi'(-\mathbf{f}_1) > \phi'(-\mathbf{f}_2)$. The strict convexity implies that $\mathbf{f}_1 < \mathbf{f}_2$. ■

The following result provides a simple way to check the infinite-sample consistency of (12). Note that since it only requires the differentiability on $(-\infty, 0]$, the SVM hinge loss is included.

Theorem 8 *If ϕ is a convex function which is bounded below, differentiable on $(-\infty, 0]$, and $\phi'(0) < 0$, then (12) is infinite-sample consistency on Ω with respect to (9).*

Proof The continuity condition is straight-forward to verify. We may also assume that $\phi(\cdot) \ge 0$ without loss of generality.

Consider $\mathbf{q} \in \Lambda_K$. Without loss of generality, we can assume that $\mathbf{q}_1 < \mathbf{q}_2$, and only need to show that $\inf\{W(\mathbf{q}, \mathbf{f}) : \mathbf{f} \in \Omega, \mathbf{f}_1 = \sup_k \mathbf{f}_k\} > W^*(\mathbf{q})$. Now consider a sequence $\mathbf{f}^{(m)}$ such that $\lim_m W(\mathbf{q}, \mathbf{f}^{(m)}) = \inf\{W(\mathbf{q}, \mathbf{f}) : \mathbf{f} \in \Omega, \mathbf{f}_1 = \sup_k \mathbf{f}_k\}$. Note that $(1 - \mathbf{q}_1)\phi(-\mathbf{f}_1^{(m)})$ is bounded.

Now if the sequence $\{\mathbf{f}^{(m)}\}$ is unbounded, then due to the constraint $\sum_k \mathbf{f}_k^{(m)} = 0$ and $\mathbf{f}_1^{(m)} \ge \mathbf{f}_k^{(m)}$, we know that the sequence $\{\mathbf{f}_1^{(m)}\}$ must also be unbounded. It follows that there is a subsequence (which for simplicity, denote as the whole sequence) such that $\mathbf{f}_1^{(m)} \to +\infty$. The boundedness of $(1 - \mathbf{q}_1)\phi(-\mathbf{f}_1^{(m)})$ implies that $\mathbf{q}_1 = 1$, which is not possible since $\mathbf{q}_1 < \mathbf{q}_2$.

Therefore we know that the sequence $\{\mathbf{f}^{(m)}\}$ must be bounded, and thus it contains a convergent subsequence. Denote the limit as $\mathbf{f}$. We have $W(\mathbf{q}, \mathbf{f}) = \lim_m W(\mathbf{q}, \mathbf{f}^{(m)})$. Therefore we only need to show that $W(\mathbf{q}, \mathbf{f}) > W^*(\mathbf{q})$. We consider three cases:

- $\mathbf{f}_1 = \mathbf{f}_2$. Since $\mathbf{f}_1 = \sup_k \mathbf{f}_k$, we have $\mathbf{f}_1 = \mathbf{f}_2 \ge 0$. The convexity assumption implies that $\phi'(-\mathbf{f}_1) = \phi'(-\mathbf{f}'_2) \le \phi'(0) < 0$. Therefore $(1 - \mathbf{q}_1)\phi'(-\mathbf{f}_1) - (1 - \mathbf{q}_2)\phi'(-\mathbf{f}_2) < 0$. It follows that there is a sufficiently small δ such that $(1 - \mathbf{q}_1)\phi(-\mathbf{f}_1 + \delta) + (1 - \mathbf{q}_2)\phi(-\mathbf{f}_2 - \delta) < (1 - \mathbf{q}_1)\phi(-\mathbf{f}_1) + (1 - \mathbf{q}_2)\phi(-\mathbf{f}_2)$. Therefore if we let $\mathbf{f}'_1 = \mathbf{f}_1 - \delta$, $\mathbf{f}'_2 = \mathbf{f}_2 + \delta$, and $\mathbf{f}'_k = \mathbf{f}_k$ when $k > 2$, then $W(\mathbf{q}, \mathbf{f}) > W(\mathbf{q}, \mathbf{f}') \ge W^*(\mathbf{q})$.

- $\mathbf{f}_1 > \mathbf{f}_2$ and $\phi(-\mathbf{f}_1) > \phi(-\mathbf{f}_2)$. In this case, if we let $\mathbf{f}'_1 = \mathbf{f}_2$, $\mathbf{f}'_2 = \mathbf{f}_1$, and $\mathbf{f}'_k = \mathbf{f}_k$ when $k > 2$, then it is easy to check that $W(\mathbf{q},\mathbf{f}) - W^*(\mathbf{q}) \leq W(\mathbf{q},\mathbf{f}) - W(\mathbf{q},\mathbf{f}') = (\mathbf{q}_1 - \mathbf{q}_2)(\phi(-\mathbf{f}_2) - \phi(-\mathbf{f}_1)) > 0$.

- $\mathbf{f}_1 > \mathbf{f}_2$ and $\phi(-\mathbf{f}_1) \leq \phi(-\mathbf{f}_2)$. Using the condition that $-\mathbf{f}_1 < 0$ and hence $\phi'(-\mathbf{f}_1) \leq \phi'(0) < 0$, we know that for a sufficiently small $\delta > 0$, we have $\phi(-\mathbf{f}_1 + \delta) < \phi(-\mathbf{f}_1) \leq \phi(-\mathbf{f}_2)$ and $-\mathbf{f}_2 - \delta > -\mathbf{f}_1$. Since the convexity of ϕ implies that $\phi(z)$ achieves the maximum on $[-\mathbf{f}_1, -\mathbf{f}_2]$ at its end points, we have $\phi(-\mathbf{f}_2) \geq \phi(-\mathbf{f}_2 - \delta)$. Therefore if we let $\mathbf{f}'_1 = \mathbf{f}_1 - \delta$, $\mathbf{f}'_2 = \mathbf{f}_2 + \delta$, and $\mathbf{f}'_k = \mathbf{f}_k$ when $k > 2$, then $W(\mathbf{q},\mathbf{f}) > W(\mathbf{q},\mathbf{f}') \geq W^*(\mathbf{q})$.

Combining the above three cases, we obtain the result. ■

Using the above criterion, we can convert an ISC convex ϕ for the binary formulation (1) into an ISC multi-category classification formulation (12). In Lee et al. (2004) the special case of SVM (with loss function $\phi(z) = (1-z)_+$ which is convex and differentiable on $(-\infty, 0]$) was studied. The authors demonstrated the infinite-sample consistency by direct calculation, although no results similar to Theorem 3, needed for proving consistency, were established. The treatment presented here generalizes their study.

4.3 One-Versus-All Method

The constrained comparison method in (12) is closely related to the one-versus-all approach, where we use the formulation (1) to train one function $\mathbf{f}_c(X)$ for each class c separately but regarding all data (X,Y) such that $Y \neq c$ as negative data, and all data (X,Y) such that $Y = c$ as positive data. It can be easily checked that the resulting formulation is a special case of (6) with

$$\Psi_c(\mathbf{f}) = \phi(\mathbf{f}_c) + \sum_{k=1,k\neq c}^{K} \phi(-\mathbf{f}_k). \tag{14}$$

Note that this formula is similar to (12), but we don't require the sum-of-zero constraint on $\mathbf{f}$ (that is $\Omega = R^K$). Intuitively, with an observation (X,Y), this formulation encourages the correct classification rule in that it favors a large $\mathbf{f}_Y(X)$ and favors small $\mathbf{f}_k(X)$ when $k \neq Y$. However, if a binary classification method (such as SVM) does not estimate the conditional probability, then the one-versus-all approach may not be infinite-sample consistent, while the formulation in (12) can still be. In order to establish the ISC condition for the one-versus-all approach, we can write

$$W(\mathbf{q},\mathbf{f}) = \sum_{c=1}^{K} \left[\mathbf{q}_c \phi(\mathbf{f}_c) + (1-\mathbf{q}_c)\phi(-\mathbf{f}_c)\right]. \tag{15}$$

We have the following order-preserving property.

Theorem 9 *Consider (14). Assume that ϕ is convex, bounded below, differentiable, and $\phi(z) < \phi(-z)$ when $z > 0$. Consider any $\mathbf{q} \in \Lambda_K$ and $\mathbf{f} \in [-\infty, +\infty]^K$ such that $W(\mathbf{q},\mathbf{f}) = W^*(\mathbf{q})$. If $\mathbf{q}_i < \mathbf{q}_j$, we have $\mathbf{f}_i < \mathbf{f}_j$.*

Proof Let f_q (not necessarily unique) minimizes $q\phi(f) + (1-q)\phi(-f)$. We have the first-order optimality condition

$$q\phi'(f_q) = (1-q)\phi'(-f_q).$$

Note that the assumptions imply that $\phi'(0) < 0$. Therefore $f_q \neq 0$ when $q \neq 0.5$ (otherwise, the optimality condition cannot be satisfied). Therefore by the assumption that $\phi(z) < \phi(-z)$ when $z > 0$, we have $f_q > 0$ when $q > 0.5$ and $f_q < 0$ when $q < 0.5$.

Let $i = 1$ and $j = 2$. We have either $\mathbf{q}_1 \in [0, 0.5)$ or $\mathbf{q}_2 \in (0.5, 1]$. Assume the former (due to the symmetry, the latter case can be proved similarly), which implies that $\mathbf{f}_1 < 0$. If $\mathbf{f}_2 \geq 0$, then the claim $\mathbf{f}_1 < \mathbf{f}_2$ holds. Therefore we only need to consider the case $\mathbf{f}_2 < 0$, and thus $0 \leq \mathbf{q}_1 < \mathbf{q}_2 \leq 0.5$. We now prove by contradiction. Note that $\mathbf{f}_2 > -\infty$ (otherwise, $\mathbf{q}_2\phi(\mathbf{f}_2) = +\infty$). If $\mathbf{f}_2 \leq \mathbf{f}_1 < 0$, then the convexity of ϕ implies $\phi'(\mathbf{f}_2) \leq \phi'(\mathbf{f}_1) < 0$. We have

$$\phi'(-\mathbf{f}_1) = \mathbf{q}_1\phi'(\mathbf{f}_1)/(1-\mathbf{q}_1) > \mathbf{q}_2\phi'(\mathbf{f}_1)/(1-\mathbf{q}_2) \geq \mathbf{q}_2\phi'(\mathbf{f}_2)/(1-\mathbf{q}_2) = \phi'(-\mathbf{f}_2).$$

The convexity implies that $-\mathbf{f}_1 > -\mathbf{f}_2$ (thus $\mathbf{f}_1 < \mathbf{f}_2$), which is a contradiction. Therefore we must have $\mathbf{f}_1 < \mathbf{f}_2$. ■

The following result shows that for a (non-flat) differentiable convex function ϕ, the one-versus-all method is infinite-sample consistent. Note that the theorem excludes the standard SVM method, which employs the non-differentiable hinge loss. However, similar to the discussion at the end of Section 4.1, if the true label can be predicted relatively accurately (that is, the dominant class has a conditional probability larger than 0.5), then the SVM one-versus-all method is consistent. Therefore the method may still perform well for some practical problems (see Rifkin and Klautau, 2004, for example).

Theorem 10 *Under the assumptions of Theorem 9. The method (14) is ISC on $\Omega = R^K$ with respect to (9).*

Proof Consider $\mathbf{q} \in \Lambda_K$. Without loss of generality, we can assume that $\mathbf{q}_1 < \mathbf{q}_2$, and only need to show that $\inf\{W(\mathbf{q},\mathbf{f}) : \mathbf{f} \in \Omega, \mathbf{f}_1 = \sup_k \mathbf{f}_k\} > W^*(\mathbf{q})$. Now consider a sequence $\mathbf{f}^{(m)}$ such that $\lim_m W(\mathbf{q},\mathbf{f}^{(m)}) = \inf\{W(\mathbf{q},\mathbf{f}) : \mathbf{f} \in \Omega, \mathbf{f}_1 = \sup_k \mathbf{f}_k\}$. Let $\mathbf{f}$ be a limiting point of $\mathbf{f}^{(m)}$ in $[-\infty,+\infty]^K$, we have $W(\mathbf{q},\mathbf{f}) = \lim_m W(\mathbf{q},\mathbf{f}^{(m)})$ and $\mathbf{f}_1 = \sup_k \mathbf{f}_k$. From Theorem 9, we have $W(\mathbf{q},\mathbf{f}) > W^*(\mathbf{q})$. ■

Using Theorem 24, we can also obtain a more quantitative bound.

Theorem 11 *Under the assumptions of Theorem 9. The function $V_\phi(q) = \inf_{f \in R}[q\phi(f) + (1-q)\phi(-f)]$ is concave on $[0,1]$. Assume that there exists a constant $c_\phi > 0$ such that*

$$(q-q')^2 \leq c_\phi^2 \left(2V_\phi(\frac{q+q'}{2}) - V_\phi(q) - V_\phi(q') \right),$$

then we have $\forall \mathbf{f}(\cdot)$,

$$\ell(T(\mathbf{f}(\cdot))) \leq \ell_B + c_\phi \left(\mathbf{E}_{X,Y}\Phi_Y(\mathbf{f}(X)) - \inf_{\mathbf{f}'} \mathbf{E}_{X,Y}\Phi_Y(\mathbf{f}'(X)) \right)^{1/2},$$

where $\Phi_Y(\mathbf{f})$ is given in (14), $T(\cdot)$ is defined in (5), ℓ is the 0-1 classification error in (9), and ℓ_B is the optimal Bayes error.

Proof $V_\phi(q)$ is the infimum of concave functions $q\phi(f)+(1-q)\phi(-f)$ indexed by $f \in R$, thus concave.

The second part is an application of Theorem 24. We use the notations of Appendix A: let $\mathcal{X}$ be the input space, $\mathcal{Q} = \Lambda_K$ be the space of conditional probability vectors, and $\mathcal{D} = \{1,\ldots,K\}$ be the space of class labels. We let $\ell(\mathbf{q},k) = \sum_{c=1,c\neq k} \mathbf{q}_c$, and thus the classification error of a decision function $p(\cdot)$ in (9) can be expressed as $\ell(p(\cdot)) = \mathbf{E}_X \ell([P(Y=c|X)]_c, p(X))$. The estimation-model space is R^K, with decision T given by (5). The W function is given by (15). Let $v(\mathbf{q}) \equiv 1$. $\forall \varepsilon > 0$, assume $\Delta\ell(\mathbf{q}, T(\mathbf{f})) \geq \varepsilon$.

Define $V_\phi(q,f) = q\phi(f)+(1-q)\phi(-f)$. Without loss of generality, we may assume that $T(\mathbf{f}) = 1$ and $\mathbf{q}_2 = \sup_c \mathbf{q}_c$. Then $\Delta\ell(\mathbf{q}, T(\mathbf{f})) = \mathbf{q}_2 - \mathbf{q}_1 \geq \varepsilon$.

$$\begin{aligned}
\Delta W(\mathbf{q},\mathbf{f}) &\geq \inf_{\mathbf{f}_1 \geq \mathbf{f}_2} \sum_{i=1}^{2} \left[V_\phi(\mathbf{q}_i,\mathbf{f}_i) - V_\phi(\mathbf{q}_i)\right] \\
&= \inf_{\mathbf{f}_1 = \mathbf{f}_2} \sum_{i=1}^{2} \left[V_\phi(\mathbf{q}_i,\mathbf{f}_i) - V_\phi(\mathbf{q}_i)\right] \\
&= 2 \inf_{\mathbf{f}_1} V_\phi\left(\frac{\mathbf{q}_1+\mathbf{q}_2}{2}, \mathbf{f}_1\right) - (V_\phi(\mathbf{q}_1)+V_\phi(\mathbf{q}_2)) \geq c_\phi^{-2}(\mathbf{q}_1-\mathbf{q}_2)^2 \geq c_\phi^{-2}\varepsilon^2.
\end{aligned}$$

The first equality holds because the minimum cannot be achieved at a point $\mathbf{f}_1 < \mathbf{f}_2$ due to the order-preserving property in Theorem 9. The assumption thus implies that $c_\phi^2 \Delta H_{\ell,W,T,v}(\varepsilon) \geq \varepsilon^2$. The desired result is now a direct consequence of Theorem 24. ■

Remark 12 *Using Taylor expansion, it is easy to verify that the condition $V_\phi''(q) \leq -c < 0$ implies that $(2V_\phi((q+q')/2) - V_\phi(q) - V_\phi(q')) \geq c(q-q')^2/4$. In this case, we may take $c_\phi = 2/\sqrt{c}$. As an example, we consider the least squares method and one of its variants: $\phi(z) = (1-v)^2$ or $\phi(z) = (1-v)_+^2$. In both cases, $V_\phi(q) = 4q(1-q)$. Therefore we can let $c_\phi = 1/\sqrt{2}$.*

The bound can also be further refined under the so-called Tsybakov small noise assumption (see Mammen and Tsybakov, 1999).

Theorem 13 *Under the assumptions of Theorem 11. Let*

$$\gamma(X) = \inf\{\sup_c P(Y=c|X) - P(Y=c'|X) : P(Y=c'|X) < \sup_c P(Y=c|X)\}$$

be the margin between the largest conditional probability and the second largest conditional probability (let $\gamma(X) = 1$ if all conditional probabilities are equal). Consider $\alpha \geq 0$ such that $c_\gamma = \mathbf{E}_X \gamma(X)^{-\alpha} < +\infty$, then we have $\forall \mathbf{f}(\cdot)$,

$$\ell(T(\mathbf{f}(\cdot))) \leq \ell_B + c_\phi^{(2\alpha+2)/(\alpha+2)} \left(\mathbf{E}_{X,Y}\Phi_Y(\mathbf{f}(X)) - \inf_{\mathbf{f}'(\cdot)} \mathbf{E}_{X,Y}\Phi_Y(\mathbf{f}'(X))\right)^{(\alpha+1)/(\alpha+2)} c_\gamma^{1/(\alpha+2)}.$$

Proof Using notations in the proof of Theorem 11, but let $v(\mathbf{q}) = \gamma(\mathbf{q})^{-\alpha}/c_\gamma$, where $\gamma(\mathbf{q}) = \inf\{\sup_c \mathbf{q}_c - \mathbf{q}_k : \mathbf{q}_k < \sup_c \mathbf{q}_c\}$. It is clear that $\mathbf{E}_X v(\mathbf{q}(X)) = 1$ with $\mathbf{q}(X) = [P(Y=1|X), \cdots, P(Y=K|X)]$.

Following the proof of Theorem 11, but assume $\mathbf{q}_2 - \mathbf{q}_1 \geq \varepsilon\nu(\mathbf{q})$. From $\mathbf{q}_2 - \mathbf{q}_1 \geq \gamma(\mathbf{q})$, we have $\forall \beta \geq 0$: $(\mathbf{q}_2 - \mathbf{q}_1)^{1+\beta}/\gamma(\mathbf{q})^{-\alpha+\beta} \geq (\mathbf{q}_2 - \mathbf{q}_1)/\gamma(\mathbf{q})^{-\alpha} \geq \varepsilon/c_\gamma$. Let $\beta = \alpha/(\alpha+2)$, we have

$$\left((\mathbf{q}_2 - \mathbf{q}_1)^2/\gamma(\mathbf{q})^{-\alpha}\right)^{(\alpha+1)/(\alpha+2)} \geq \varepsilon/c_\gamma.$$

This implies that (the first inequality follows from the proof of Theorem 11)

$$\Delta W(\mathbf{q},\mathbf{f})/\nu(\mathbf{q}) \geq c_\phi^{-2}(\mathbf{q}_1 - \mathbf{q}_2)^2/\nu(\mathbf{q}) \geq \varepsilon^{(\alpha+2)/(\alpha+1)} c_\gamma^{-1/(\alpha+1)} c_\phi^{-2}.$$

Thus $c_\gamma^{1/(\alpha+1)} c_\phi^2 \Delta H_{\ell,W,T,\nu}(\varepsilon) \geq \varepsilon^{(\alpha+2)/(\alpha+1)}$. The bound now follows directly from Theorem 24. ■

4.4 Unconstrained Background Discriminative Method

We consider the following unconstrained formulation:

$$\Psi_c(\mathbf{f}) = \psi(\mathbf{f}_c) + s\left(\sum_{k=1}^K t(\mathbf{f}_k)\right), \tag{16}$$

where ψ, s and t are appropriately chosen convex functions that are continuously differentiable. As we shall see later, this is a generalization of the maximum-likelihood method, which corresponds to $s(z) = t(z) = 1$ and $\psi(z) = -\ln(z)$.

We shall choose s and t such that the unconstrained background term $s\left(\sum_{k=1}^K t(\mathbf{f}_k)\right)$ penalizes large $\mathbf{f}_k$ for all k. We also choose a decreasing $\psi(\mathbf{f}_c)$ so that it favors a large $\mathbf{f}_c$. That is, it serves the purpose of discriminating $\mathbf{f}_c$ against the background term. The overall effect is to favor a predictor in which $\mathbf{f}_c$ is larger than $\mathbf{f}_k$ ($k \neq c$). In (16), the first term has a relatively simple form that depends only on the label c. The second term is independent of the label, and can be regarded as a normalization term. Note that this function is symmetric with respect to components of $\mathbf{f}$. This choice treats all potential classes equally. It is also possible to treat different classes differently. For example, replacing $\psi(\mathbf{f}_c)$ by $\psi_c(\mathbf{f}_c)$ or replacing $t(\mathbf{f}_k)$ by $t_k(\mathbf{f}_k)$.

4.4.1 Optimality Equation and Probability Model

Using (16), the conditional true Ψ risk (8) can be written as

$$W(\mathbf{q},\mathbf{f}) = \sum_{c=1}^K \mathbf{q}_c \psi(\mathbf{f}_c) + s\left(\sum_{c=1}^K t(\mathbf{f}_c)\right).$$

In the following, we study the property of the optimal vector $\mathbf{f}^*$ that minimizes $W(\mathbf{q},\mathbf{f})$ for a fixed $\mathbf{q}$.

Given $\mathbf{q}$, the optimal solution $\mathbf{f}^*$ that minimizes $W(\mathbf{q},\mathbf{f})$ satisfies the following first order optimality condition:

$$\mathbf{q}_c \psi'(\mathbf{f}_c^*) + \mu_{\mathbf{f}^*} t'(\mathbf{f}_c^*) = 0 \qquad (c = 1,\ldots,K). \tag{17}$$

where the quantity $\mu_{\mathbf{f}^*} = s'(\sum_{k=1}^K t(\mathbf{f}_k^*))$ is independent of c.

Clearly this equation relates $\mathbf{q}_c$ to $\mathbf{f}_c^*$ for each component c. The relationship of $\mathbf{q}$ and $\mathbf{f}^*$ defined by (17) can be regarded as the (infinity sample-size) probability model associated with the learning method (6) with Ψ given by (16). The following result is quite straight-forward. We shall skip the proof.

Theorem 14 *Assume that ψ, t, s are differentiable functions such that $s'(x) > 0$. If for $a \in [0, +\infty)$, the the solution x of $a\psi'(x) + t'(x) = 0$ is an increasing function of a, then the solution of (17) has the order preserving property: $\mathbf{q}_i < \mathbf{q}_j$ implies $\mathbf{f}_i^* < \mathbf{f}_j^*$. Moreover, the method (16) is ISC.*

In the following, we shall present various formulations of (16) which have the order preserving property.

4.4.2 DECOUPLED FORMULATIONS

We let $s(u) = u$ in (16). The optimality condition (17) becomes

$$\mathbf{q}_c \psi'(\mathbf{f}_c^*) + t'(\mathbf{f}_c^*) = 0 \qquad (c = 1, \ldots, K). \tag{18}$$

This means that we have K decoupled equalities, one for each $\mathbf{f}_c$. This is the simplest and in the author's opinion, the most interesting formulation. Since the estimation problem in (6) is also decoupled into K separate equations, one for each component of $\hat{\mathbf{f}}$, this class of methods are computationally relatively simple and easy to parallelize. Although this method seems to be preferable for multi-category problems, it is not the most efficient way for two-class problems (if we want to treat the two classes in a symmetric manner) since we have to solve two separate equations. We only need to deal with one equation in (1) due to the fact that an effective constraint $\mathbf{f}_1 + \mathbf{f}_2 = 0$ can be used to reduce the number of equations. This variable elimination has little impact if there are many categories.

In the following, we list some examples of multi-category risk minimization formulations. They all have the order preserving property, hence are infinite-sample consistent. We focus on the relationship of the optimal optimizer function $\mathbf{f}_*(\mathbf{q})$ and the conditional probability $\mathbf{q}$, which gives the probability model.

$\psi(u) = -u$ AND $t(u) = e^u$

We obtain the following probability model: $\mathbf{q}_c = e^{\mathbf{f}_c^*}$. This formulation is closely related to the maximum-likelihood estimate with conditional model $\mathbf{q}_c = e^{\mathbf{f}_c^*} / \sum_{k=1}^K e^{\mathbf{f}_k^*}$ (logistic regression). In particular, if we choose a function class such that the normalization condition $\sum_{k=1}^K e^{\mathbf{f}_k} = 1$ holds, then the two formulations are identical. However, they become different when we do not impose such a normalization condition.

$\phi(u) = -\ln u$ AND $t(u) = u$

This formulation is closely related to the previous formulation. It is an extension of maximum-likelihood estimate with probability model $\mathbf{q}_c = \mathbf{f}_c^*$. The resulting method is identical to the maximum-likelihood method if we choose our function class such that $\sum_k \mathbf{f}_k = 1$ and $\mathbf{f}_k \geq 0$ for $k = 1, \ldots, K$. However, the formulation also allows us to use function classes that do not satisfy the normalization constraint $\sum_k \mathbf{f}_k = 1$. Therefore this method is more flexible.

$\phi(u) = -\frac{1}{\alpha} u^\alpha$ $(0 < \alpha < 1)$ AND $t(u) = u$

Closely related to the maximum-likelihood method, this formulation replaces $\phi(u) = -\ln(u)$ by $\phi(u) = -u^\alpha$. The solution is $\mathbf{q}_c = (\mathbf{f}_c^*)^{1/(1-\alpha)}$. Similar to the case of $\phi(u) = -\ln(u)$, we may also impose a constraint $\sum_k \mathbf{f}_k^{1/(1-\alpha)} = 1$, which ensures that the estimated probability always sum to one.

$\phi(u) = -u$ AND $t(u) = \ln(1+e^u)$

This version uses binary logistic regression loss, and we have the following probability model: $\mathbf{q}_c = (1+e^{-\mathbf{f}_c^*})^{-1}$. Again this is an unnormalized model.

$\phi(u) = -u$ AND $t(u) = \frac{1}{p}|u|^p$ $(p > 1)$

We obtain the following probability model: $\mathbf{q}_c = \text{sign}(\mathbf{f}_c^*)|\mathbf{f}_c^*|^{p-1}$. This means that at the solution, $\mathbf{f}_c^* \geq 0$. This formulation is not normalized. If we choose a function family such that $\sum_k |\mathbf{f}_k|^{p-1} = 1$ and $\mathbf{f}_k \geq 0$, then we have a normalized model for which the estimated conditional probability always sum to one. One can also modify this method such that we can use $\mathbf{f}_c^* \leq 0$ to model the condition probability $\mathbf{q}_c = 0$.

$\phi(u) = -u$ AND $t(u) = \frac{1}{p}\max(u,0)^p$ $(p > 1)$

In this probability model, we have the following relationship: $\mathbf{q}_c = \max(\mathbf{f}_c^*, 0)^{p-1}$. The equation implies that we allow $\mathbf{f}_c^* \leq 0$ to model the conditional probability $\mathbf{q}_c = 0$. Therefore, with a fixed function class, this model is more powerful than the previous one. However, at the optimal solution, we still require that $\mathbf{f}_c^* \leq 1$. This restriction can be further alleviated with the following modification.

$\phi(u) = -u$ AND $t(u) = \frac{1}{p}\min(\max(u,0)^p, p(u-1)+1)$ $(p > 1)$

In this model, we have the following relationship at the solution: $\mathbf{q}_c = \min(\max(\mathbf{f}_c^*, 0), 1)^{p-1}$. Clearly this model is more powerful than the previous model since the function value $\mathbf{f}_c^* \geq 1$ can be used to model $\mathbf{q}_c = 1$. For separable problems, at each point there exists a c such that $\mathbf{q}_c = 1$ and $\mathbf{q}_k = 0$ when $k \neq c$. The model requires that $\mathbf{f}_c^* \geq 1$ and $\mathbf{f}_k^* \leq 0$ when $k \neq c$. This is essentially a large margin separation condition, where the function for the true class is separated from the rest by a margin of one.

4.4.3 COUPLED FORMULATIONS

In the coupled formulation with $s(u) \neq u$, the probability model are inherently normalized in some sense. We shall just list a few examples.

$\phi(u) = -u$, AND $t(u) = e^u$, AND $s(u) = \ln(u)$

This is the standard logistic regression model. The probability model is

$$\mathbf{q}_c(x) = \frac{e^{\mathbf{f}_c^*(x)}}{\sum_{c=1}^K e^{\mathbf{f}_c^*(x)}}.$$

The right hand side is always normalized (sum up to 1). One potential disadvantage of this method (at this moment, we don't know whether or not this theoretical disadvantage causes real problems in practice or not) is that it does not model separable data very well. That is, if $\mathbf{q}_c(x) = 0$ or $\mathbf{q}_c(x) = 1$, we require $\mathbf{f}_c^* = \pm\infty$. In comparison, some large margin methods described earlier can model the separable scenario using finite valued $\mathbf{f}^*$.

$\phi(u) = -u$, AND $t(u) = |u|^{p'}$, AND $s(u) = \frac{1}{p}|u|^{p/p'}$ $(p, p' > 1)$

The probability model is

$$\mathbf{q}_c(x) = \left(\sum_{k=1}^{K} |\mathbf{f}_k^*(x)|^{p'}\right)^{(p-p')/p'} \text{sign}(\mathbf{f}_c^*(x))|\mathbf{f}_c^*(x)|^{p'-1}.$$

We may replace $t(u)$ by $t(u) = \max(0,u)^p$, and the probability model becomes

$$\mathbf{q}_c(x) = \left(\sum_{k=1}^{K} \max(\mathbf{f}_k^*(x),0)^{p'}\right)^{(p-p')/p'} \max(\mathbf{f}_c^*(x),0)^{p'-1}.$$

These formulations do not seem to have advantages over the decoupled counterparts (with $s(u) = 1$). For the decoupled counterparts, as explained, the normalization (so that the estimated probability sum to one) can be directly included into the function class. This is more difficult to achieve here due to the more complicated formulations. However, it is unclear whether normalized formulations have practical advantages since one can always explicitly normalize the estimated conditional probability.

5. Consistency of Kernel Multi-Category Classification Methods

In this section, we give conditions that lead to the consistency of kernel methods. It is worth mentioning that generalization bounds obtained in this section are not necessarily tight. We use simple analysis to demonstrate that statistical consistency can be obtained. In order to obtain good rate of convergence results, more sophisticated analysis (such as those used by Blanchard et al., 2004, Bartlett et al., 2003, Mannor et al., 2003, van de Geer, 2000, Scovel and Steinwart, 2003) is needed.

The analysis given in this section is kernel independent. Therefore we can start with an arbitrary reproducing kernel Hilbert space H (for example, see Wahba, 1990, for definition) with inner product $\cdot$ and norm $\|\cdot\|_H$. Each element of H is a function $f(x)$ of the input x. It is well known that for each data point x, we can embed it into H as h_x such that $f(x) = f \cdot h_x$ for all $f \in H$.

In this section, we only consider bounded input distribution D:

$$\sup_x \|h_x\|_H < \infty.$$

We also introduce the following notations:

$$H_A = \{f(\cdot) \in H : \|f\|_H \sup_x \|h_x\|_H \le A\},$$
$$H_{A,K} = H_A^K = \{\mathbf{f}(\cdot) : \mathbf{f}_c(\cdot) \in H_A \text{ for all } c = 1,\ldots,K\}.$$

For notation simplicity, we shall limit our discussion to formulations such that for all $c = 1,\ldots,K$, $\Psi_c(\cdot)$ defined on a subset $\Omega \subset R^K$ can be extended to R^K. For example, for the constrained comparison model with the SVM loss. we require that $\Omega = \{\mathbf{f} \in R^K : \sum_{k=1}^K \mathbf{f}_k = 0\}$, but the formulation itself is well-defined on the entire R^K.

In order to obtain a uniform convergence bound, we shall introduce the following Lipschitz condition. It is clear that all well-behaved formulations such as those considered in this paper satisfy this assumption.

Assumption 15 *Given any $A > 0$, and consider $S_A = \{\mathbf{f} \in R^K : \sup_c |\mathbf{f}_c| \leq A\}$. Then there exists γ_A such that $\forall \mathbf{f}, \mathbf{f}' \in S_A$ and $1 \leq c \leq K$:*

$$|\Psi_c(\mathbf{f}) - \Psi_c(\mathbf{f}')| \leq \gamma_A \sup_k |\mathbf{f}_k - \mathbf{f}'_k|.$$

Definition 16 *Let $Q_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a set of n points. We define the $\ell_\infty(Q_n)$ distance between any two functions $f(x,y)$ and $g(x,y)$ as*

$$\ell_\infty(Q_n)(f,g) = \sup_i |f(X_i, Y_i) - g(X_i, Y_i)|.$$

Let $\mathcal{F}$ be a class of functions of (x,y), the empirical ℓ_∞-covering number *of $\mathcal{F}$, denoted by $N(\varepsilon, \mathcal{F}, \ell_\infty(Q_n))$, is the minimal number of balls $\{g : \ell_\infty(Q_n)(g,f) \leq \varepsilon\}$ of radius ε needed to cover $\mathcal{F}$. The* uniform ℓ_∞ covering number *is given by*

$$N_\infty(\varepsilon, \mathcal{F}, n) = \sup_{Q_n} N(\varepsilon, F, \ell_\infty(Q_n)),$$

where the supremum is over all samples Q_n of size n.

Note that we may also use other covering numbers such as ℓ_2 covering numbers. The ℓ_∞ covering number is more suitable for the specific Lipschitz condition used in Assumption 15. We use the following kernel-independent covering number bound.

Lemma 17 *Consider the function class $\mathcal{F}_{A,K} = \{\Psi_Y(\mathbf{f}(X)) : \mathbf{f} \in H_{A,K}\}$ such that Ψ satisfies Assumption 15. Then there exists a universal constant $C_1 > 0$ such that*

$$\ln N_\infty(\gamma_A \varepsilon, \mathcal{F}_{A,K}, n) \leq K C_1 A^2 \frac{\ln(2 + A/\varepsilon) + \ln n}{n\varepsilon^2}.$$

Proof Note that Theorem 4 of Zhang (2002) implies that there exists C_1 such that

$$\ln N_\infty(\varepsilon, H_A, n) \leq C_1 A^2 \frac{\ln(2 + A/\varepsilon) + \ln n}{n\varepsilon^2}.$$

Therefore with empirical samples $Q_n = \{(X_i, Y_i)\}$, we can find $\exp(K C_1 A^2 \frac{\ln(2+A/\varepsilon)+\ln n}{n\varepsilon^2})$ vectors $\mathbf{f}^j(X_i)$ such that for each $\mathbf{f} \in H_{A,K}$, we have $\inf_j \sup_{i,c} |\mathbf{f}_c(X_i) - \mathbf{f}^j_c(X_i)| \leq \varepsilon$. The assumption implies that this is a cover of $\mathcal{F}_{A,K}$ of radius $\gamma_A \varepsilon$. ■

Remark 18 *For specific kernels, the bound can usually be improved. Moreover, the* log*-covering number (entropy) depends linearly on the number of classes K. This is due to the specific regularization condition we use here. For practical problems, it can be desirable to use other regularization conditions so that the corresponding covering numbers have much weaker dependency (or even independence) on K. For simplicity, we will not discuss such issues in this paper.*

Lemma 19 *Consider function class $\mathcal{F}_{A,K} = \{\Psi_Y(\mathbf{f}(X)) : \mathbf{f} \in H_{A,K}\}$ such that Ψ satisfies Assumption 15. Then there exists a universal constant C such that for all $n \geq 2$:*

$$\mathbf{E}_{Q_n} \sup_{\mathbf{f}\in\mathcal{F}_{A,K}} \left| \frac{1}{n}\sum_{i=1}^{n} \Psi_{Y_i}(\mathbf{f}(X_i)) - \mathbf{E}_{X,Y}\Psi_Y(\mathbf{f}(X)) \right| \leq C\sqrt{K}\frac{\gamma_A A \ln^{3/2} n}{\sqrt{n}},$$

where $\mathbf{E}_{Q_n}$ denotes the expectation over empirical training data $Q_n = \{(X_i, Y_i)\}$.

Proof Let $\mathbf{f}_0 \in H_{A,K}$, and define $\mathcal{F}^0_{A,K} = \{\Psi_Y(\mathbf{f}(X)) - \Psi_Y(\mathbf{f}_0(X)) : \mathbf{f} \in \mathcal{F}_{A,K}\}$. Consider a sequence of binary random variables such that $\sigma_i = \pm 1$ with probability $1/2$. The *Rademacher complexity* of $\mathcal{F}^0_{A,K}$ under empirical sample $Q_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is given by

$$R(\mathcal{F}^0_{A,K}, Q_n) = \mathbf{E}_\sigma \sup_{\mathbf{f}\in H_{A,K}} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i(\Psi_{Y_i}(\mathbf{f}(X_i)) - \Psi_{Y_i}(\mathbf{f}_0(X_i))) \right|.$$

It is well known that there exists a universal constant C_2 (a variant of Corollary 2.2.8 in van der Vaart and Wellner, 1996):

$$R(\mathcal{F}^0_{A,K}, Q_n) \leq C_2 \inf_{\varepsilon_0} \left[\varepsilon_0 + \frac{1}{\sqrt{n}} \int_{\varepsilon_0}^{\infty} \sqrt{\log N_\infty(\varepsilon, \mathcal{F}, Q_n)}\, d\varepsilon \right].$$

Using the bound in Lemma 17, and perform the integration with $\varepsilon_0 = \gamma_A A\sqrt{1/n}$, we obtain

$$R(\mathcal{F}^0_{A,K}, Q_n) \leq \frac{C\sqrt{K}}{2}\frac{\gamma_A A \ln^{3/2} n}{\sqrt{n}},$$

where C is a universal constant.

Now using the standard symmetrization argument (for example, see Lemma 2.3.1 of van der Vaart and Wellner, 1996), we have

$$\mathbf{E}_{Q_n} \sup_{\mathbf{f}\in\mathcal{F}_{A,K}} \left| \frac{1}{n}\sum_{i=1}^{n} \Psi_{Y_i}(\mathbf{f}(X_i)) - \mathbf{E}_{X,Y}\Psi_Y(\mathbf{f}(X)) \right| \leq 2\mathbf{E}_{Q_n} R(\mathcal{F}^0_{A,K}, Q_n) \leq C\sqrt{K}\frac{\gamma_A A \ln^{3/2} n}{\sqrt{n}}.$$

■

Theorem 20 *Consider Ψ that satisfies Assumption 15. Choose A_n such that $A_n \to \infty$ and $\gamma_{A_n} A_n \ln^{3/2} n/\sqrt{n} \to 0$. Let $C_n = H_{A_n,K} \cap \mathcal{B}_\Omega$ (see Theorem 3 for the definition of $\mathcal{B}_\Omega$), where $\Omega \subset R^K$ is a constraint set. Consider the estimator $\hat{\mathbf{f}}(\cdot)$ in (6). We have*

$$\lim_{n\to\infty} \mathbf{E}_{Q_n} \mathbf{E}_{X,Y}\Psi_Y(\hat{\mathbf{f}}(X)) = \inf_{\mathbf{f}\in H\cap\mathcal{B}_\Omega} \mathbf{E}_{X,Y}\Psi_Y(\mathbf{f}(X)).$$

Proof Consider $\mathbf{f}^{(n)} \in C_n$ that minimizes $\mathbf{E}_{X,Y}\Psi_Y(\mathbf{f}(X))$. Since $\sum_{i=1}^n \Psi_{Y_i}(\hat{\mathbf{f}}(X_i)) \leq \sum_{i=1}^n \Psi_{Y_i}(\mathbf{f}^{(n)}(X_i))$, we have from Lemma 19 that

$$\mathbf{E}_{Q_n} \mathbf{E}_{X,Y}\Psi_Y(\hat{\mathbf{f}}(X)) \leq \mathbf{E}_{X,Y}\Psi_Y(\mathbf{f}^{(n)}(X)) + 2C\sqrt{K}\frac{\gamma_{A_n} A_n \ln^{3/2} n}{\sqrt{n}}.$$

Therefore as $n \to \infty$,

$$\lim_n \mathbf{E}_{Q_n} \mathbf{E}_{X,Y} \Psi_Y(\hat{\mathbf{f}}(X)) \to \lim_n \mathbf{E}_{X,Y} \Psi_Y(\mathbf{f}^{(n)}(X)) = \inf_{\mathbf{f} \in H \cap \mathcal{B}_\Omega} \mathbf{E}_{X,Y} \Psi_Y(\mathbf{f}(X)).$$

■

The following consistency result is a straight-forward consequence of Theorem 20 and Theorem 3.

Corollary 21 *Under the conditions of Theorem 20. Assume that Ψ is ISC on Ω with respect to (3). If H is dense in $\mathcal{B}_\Omega$, that is,*

$$\inf_{\mathbf{f}(\cdot) \in H \cap \mathcal{B}_\Omega} \mathbf{E}_{X,Y} \Psi_Y(\mathbf{f}(X)) = \inf_{\mathbf{f}(\cdot) \in \mathcal{B}_\Omega} \mathbf{E}_{X,Y} \Psi_Y(\mathbf{f}(X)),$$

then

$$\lim_{n \to \infty} \mathbf{E}_{Q_n} \mathbf{E}_{X,Y} \Psi_Y(\hat{\mathbf{f}}(X)) = \inf_{\mathbf{f}(\cdot) \in \mathcal{B}_\Omega} \mathbf{E}_{X,Y} \Psi_Y(\mathbf{f}(X)).$$

This implies that the classification error $\ell(\hat{\mathbf{f}})$ converges to the optimal Bayes error in probability.

6. Conclusion

In this paper we investigated a general family of risk minimization based multi-category classification algorithms, which can be considered as natural extensions of binary large margin methods. We established infinite-sample consistency conditions that ensure the statistical consistency of the obtained classifiers in the infinite-sample limit. Several specific forms of the general risk minimization formulation were considered. We showed that some models can be used to estimate conditional class probabilities. As an implication of this work, we see that it is possible to obtain consistent conditional density estimators using various non-maximum likelihood estimation methods. One advantage of some proposed large margin methods is that they allow us to model zero conditional probability directly. Note that for the maximum-likelihood method, near-zero conditional probability may cause robustness problems (at least in theory) due to the unboundedness of the log-loss function. Moreover, combined with some relatively simple generalization analysis, we showed that given appropriately chosen regularization conditions in some reproducing kernel Hilbert spaces, classifiers obtained from some multi-category kernel methods can approach the optimal Bayes error in the large sample limit.

Appendix A. Relationship of True Loss Minimization and Surrogate Loss Minimization

We consider an abstract decision model. Consider input space $\mathcal{X}$, output-model space Q, decision space $\mathcal{D}$, and estimation-model space Ω.

Consider the following functions:

- True loss function: $\ell : Q \times \mathcal{D} \to R$. We also define the corresponding excess loss as

$$\Delta\ell(\mathbf{q}, d) = \ell(\mathbf{q}, d) - \inf_{d' \in \mathcal{D}} \ell(\mathbf{q}, d').$$

- Surrogate loss function: $W : \mathcal{Q} \times \Omega \to R$. We also define the corresponding excess surrogate loss as
$$\Delta W(\mathbf{q}, \mathbf{f}) = W(\mathbf{q}, \mathbf{f}) - \inf_{\mathbf{f}' \in \mathcal{D}} W(\mathbf{q}, \mathbf{f}').$$
- Decision-rule: $T : \Omega \to \mathcal{D}$.

For the multi-category classification problem studied in the main text, $\mathcal{X}$ is the input space, $\mathcal{Q} = \Lambda_K$ is the space of conditional probability vectors $[P(Y = c|\cdot)]_c$, $\mathcal{D} = \{1, \ldots, K\}$ is the space of class labels, and $\Omega \subset R^K$ is the set of possible vector predictors $\mathbf{f} \in R^K$, with T given by (5). The W function is given by (8). With classification error in (2), we let

$$\ell(\mathbf{q}, k) = \sum_{c=1, c \neq k}^{K} a_c \mathbf{q}_c.$$

Therefore the classification error of a decision function $p(\cdot)$ in (3) can be expressed as

$$\ell(p(\cdot)) = \mathbf{E}_X \ell([P(Y = c|X)]_c, p(X)).$$

Definition 22 *Consider function $v : \mathcal{Q} \to R^+$. $\forall \varepsilon \geq 0$, we define*

$$\Delta H_{\ell,W,T,v}(\varepsilon) = \inf \left\{ \frac{\Delta W(\mathbf{q}, \mathbf{f})}{v(\mathbf{q})} : \Delta \ell(\mathbf{q}, T(\mathbf{f})) \geq \varepsilon v(\mathbf{q}) \right\} \cup \{+\infty\}.$$

The definition is designed so that the following properties hold. They are simple re-interpretations of the definition.

Proposition 23 *We have:*

- $\Delta H_{\ell,W,T,v}(\varepsilon) \geq 0$.
- $\Delta H_{\ell,W,T,v}(0) = 0$.
- $\Delta H_{\ell,W,T,v}(\varepsilon)$ *is non-decreasing on* $[0, +\infty)$.
- $v(\mathbf{q}) \Delta H_{\ell,W,T,v}(\Delta \ell(\mathbf{q}, T(\mathbf{f}))/v(\mathbf{q})) \leq \Delta W(\mathbf{q}, \mathbf{f})$.

The importance of the above definition is based on the following theorem. It essentially gives a bound on the expected excessive true loss ℓ using the expected excessive surrogate loss W. The idea was used by Bartlett et al. (2003), Zhang (2004) to analyze binary classification problems.

Theorem 24 *Given any distribution on $\mathcal{X}$, and function $v : \mathcal{Q} \to R^+$ such that*

$$\mathbf{E}_X v(\mathbf{q}(X)) = 1.$$

Let $\zeta(\varepsilon)$ be a convex function on $[0, +\infty)$ such that $\zeta(\varepsilon) \leq \Delta H_{\ell,W,T,v}(\varepsilon)$. Then $\forall \mathbf{f} : \mathcal{X} \to \Omega$, we have

$$\zeta(\mathbf{E}_X \Delta \ell(\mathbf{q}(X), T(\mathbf{f}(X)))) \leq \mathbf{E}_X \Delta W(\mathbf{q}(X), \mathbf{f}(X)).$$

Proof Using Jensen's inequality, we have

$$\zeta(\mathbf{E}_X \Delta\ell(\mathbf{q}(X), T(\mathbf{f}(X)))) \leq \mathbf{E}_X v(\mathbf{q}(X)) \zeta\left(\frac{\Delta\ell(\mathbf{q}(X), T(\mathbf{f}(X)))}{v(\mathbf{q}(X))}\right).$$

Now using the assumption and Proposition 23, we can upper-bound the right hand side by $\mathbf{E}_X \Delta W(\mathbf{q}(X), \mathbf{f}(X))$. This proves the theorem. ■

The following proposition is based mostly on Bartlett et al. (2003). We include it here for completeness.

Proposition 25 *Let $\zeta_*(\varepsilon) = \sup_{a\geq 0, b}\{a\varepsilon + b : \forall z \geq 0, az + b \leq \Delta H_{\ell,W,T,v}(z)\}$, then ζ_* is a convex function. It has the following properties:*

- $\zeta_*(\varepsilon) \leq \Delta H_{\ell,W,T,v}(\varepsilon)$.
- $\zeta_*(\varepsilon)$ *is non-decreasing.*
- *For all convex function ζ such that $\zeta(\varepsilon) \leq \Delta H_{\ell,W,T,v}(\varepsilon)$, $\zeta(\varepsilon) \leq \zeta_*(\varepsilon)$.*
- *Assume that $\exists a > 0$ and $b \in R$ such that $a\varepsilon + b \leq \Delta H_{\ell,W,T,v}(\varepsilon)$, and $\forall \varepsilon > 0, \Delta H_{\ell,W,T,v}(\varepsilon) > 0$. Then $\forall \varepsilon > 0, \zeta_*(\varepsilon) > 0$.*

Proof We note that ζ_* is the point-wise supreme of convex functions, thus it is also convex. We now prove the four properties.

- The first property holds by definition.
- The second property follows from the fact that $\Delta H_{\ell,W,T,v}(z)$ is non-decreasing, and $a\varepsilon' + b > a\varepsilon + b$ when $\varepsilon' > \varepsilon$.
- Given a convex function ζ such that $\zeta(\varepsilon) \leq \Delta H_{\ell,W,T,v}(\varepsilon)$. At any ε, we can find a line $az + b \leq \zeta(z) \leq \Delta H_{\ell,W,T,v}(z)$ and $\zeta(\varepsilon) = a\varepsilon + b$. This implies that $\zeta(\varepsilon) \leq \zeta_*(\varepsilon)$.
- Consider $\varepsilon > 0$. Using the fact that when $z \geq \varepsilon/2$, $\Delta H_{\ell,W,T,v}(z) \geq \Delta H_{\ell,W,T,v}(\varepsilon/2) > 0$, and the assumption, we know that there exists $a_\varepsilon \in (0, a)$ such that $a_\varepsilon(z - \varepsilon/2) < \Delta H_{\ell,W,T,v}(z)$. Therefore $\zeta_*(\varepsilon) \geq a_\varepsilon(\varepsilon - \varepsilon/2) > 0$.

■

The following result shows that the approximate minimization of the expected surrogate loss $\mathbf{E}_X \Delta W$ implies the approximate minimization of the expected true loss $\mathbf{E}_X \Delta\ell$.

Corollary 26 *Consider function $v : Q \to R^+$. If the loss function $\ell(\mathbf{q}, d)/v(\mathbf{q})$ is bounded, and $\forall \varepsilon > 0$, $\Delta H_{\ell,W,T,v}(\varepsilon) > 0$, then there exists a concave function ξ on $[0, +\infty)$ that depends only on ℓ, W, T, and v, such that $\xi(0) = 0$ and $\lim_{\delta\to 0^+} \xi(\delta) = 0$. Moreover, for all distribution on $\mathcal{X}$ such that $\mathbf{E}_X v(\mathbf{q}(X)) = 1$, we have*

$$\mathbf{E}_X \Delta\ell(\mathbf{q}(X), T(\mathbf{f}(X))) \leq \xi(\mathbf{E}_X \Delta W(\mathbf{q}(X), \mathbf{f}(X))).$$

Proof Consider $\zeta_*(\varepsilon)$ in Proposition 25. Let $\xi(\delta) = \sup\{\varepsilon : \varepsilon \geq 0, \zeta_*(\varepsilon) \leq \delta\}$. Then $\zeta_*(\varepsilon) \leq \delta$ implies that $\varepsilon \leq \xi(\delta)$. Therefore the desired inequality follows from Theorem 24. Given $\delta_1, \delta_2 \geq 0$: from $\zeta_*(\frac{\xi(\delta_1)+\xi(\delta_2)}{2}) \leq \frac{\delta_1+\delta_2}{2}$, we know that $\frac{\xi(\delta_1)+\xi(\delta_2)}{2} \leq \xi(\frac{\delta_1+\delta_2}{2})$. Therefore ξ is concave.

We now only need to show that ξ is continuous at 0. From the boundedness of $\ell(\mathbf{q},d)/v(\mathbf{q})$, we know that $\Delta H_{\ell,W,T,v}(z) = +\infty$ when $z > \sup \Delta\ell(\mathbf{q},d)/v(\mathbf{q})$. Therefore $\exists a > 0$ and $b \in R$ such that $a\varepsilon + b \leq \Delta H_{\ell,W,T,v}(\varepsilon)$. Now Pick any $\varepsilon' > 0$, and let $\delta' = \zeta_*(\varepsilon')/2$, we know from Proposition 25 that $\delta' > 0$. This implies that $\xi(\delta) < \varepsilon'$ when $\delta < \delta'$. ■

One can always choose $v(\mathbf{q}) \equiv 1$ to obtain a bound that applies to all underlying distributions on $\mathcal{X}$. However, with a more general v, one may obtain better bounds in some scenarios especially the low noise case. For example, see Theorem 13 in the main text.

Appendix B. Proof of Theorem 3

We shall first prove the following lemma.

Lemma 27 *$W^*(\mathbf{q}) := \inf_{\mathbf{f}\in\Omega} W(\mathbf{q},\mathbf{f})$ is a continuous function on Λ_K.*

Proof Consider a sequence $\mathbf{q}^{(m)} \in \Lambda_K$ such that $\lim_m \mathbf{q}^{(m)} = \mathbf{q}$. Without loss of generality, we assume that there exists k such that $\mathbf{q}_1 = \cdots = \mathbf{q}_k = 0$ and $\mathbf{q}_c > 0$ for $c > k$. Moreover, since each Ψ_c is bounded below, we may assume without loss of generality that $\Psi_c \geq 0$ (this condition can be achieved simply by adding a constant to each Ψ_c).

Now, let

$$\bar{W}(\mathbf{q}',\mathbf{f}) = \sum_{c=k+1}^{K} \mathbf{q}'_c \Psi_c(\mathbf{f})$$

and

$$\bar{W}^*(\mathbf{q}') = \inf_{\mathbf{f}\in\Omega} \sum_{c=k+1}^{K} \mathbf{q}'_c \Psi_c(\mathbf{f}).$$

Since $\{\bar{W}^*(\mathbf{q}^{(m)})\}_m$ is bounded, each sequence $\{\mathbf{q}_c^{(m)}\Psi_c(\cdot)\}_m$ is also bounded near the optimal solution. It is clear from the condition $\lim_m \mathbf{q}_c^{(m)} > 0$ $(c > k)$ that

$$\lim_{m\to\infty} \bar{W}^*(\mathbf{q}^{(m)}) = \bar{W}^*(\mathbf{q}) = W^*(\mathbf{q}).$$

Since $W^*(\mathbf{q}^{(m)}) \geq \bar{W}^*(\mathbf{q}^{(m)})$, we have

$$\liminf_{m\to\infty} W^*(\mathbf{q}^{(m)}) \geq W^*(\mathbf{q}). \tag{19}$$

Now for a sufficiently large positive number A, let

$$W_A^*(\mathbf{q}') = \inf_{\mathbf{f}\in\Omega, \|\mathbf{f}\|_1 \leq A} \sum_{c=1}^{K} \mathbf{q}'_c \Psi_c(\mathbf{f}).$$

We have

$$\limsup_{m\to\infty} W^*(\mathbf{q}^{(m)}) \leq \limsup_{m\to\infty} W_A^*(\mathbf{q}^{(m)}) = W_A^*(\mathbf{q}).$$

Since $\lim_{A\to\infty} W_A^*(\mathbf{q}) = W^*(\mathbf{q})$, we have

$$\limsup_{m\to\infty} W^*(\mathbf{q}^{(m)}) \le W^*(\mathbf{q}).$$

Combining this inequality with (19), we obtain the lemma. ■

Lemma 28 $\forall \varepsilon > 0$, $\exists \delta > 0$ *such that* $\forall \mathbf{q} \in \Lambda_k$:

$$\inf\left\{ W(\mathbf{q},\mathbf{f}) : \mathbf{f}_c = \sup_k \mathbf{f}_k, a_c\mathbf{q}_c \le \sup_k a_k\mathbf{q}_k - \varepsilon \right\} \ge W^*(\mathbf{q}) + \delta. \tag{20}$$

Proof We prove this by contradiction. Assume that (20) does not hold, then $\exists \varepsilon > 0$, and a sequence of $(c^{(m)}, \mathbf{f}^{(m)}, \mathbf{q}^{(m)})$ with $\mathbf{f}^{(m)} \in \Omega$ such that $\mathbf{f}^{(m)}_{c^{(m)}} = \sup_k \mathbf{f}^{(m)}_k$, $a_{c^{(m)}}\mathbf{q}^{(m)}_{c^{(m)}} \le \sup_k a_k\mathbf{q}^{(m)}_k - \varepsilon$, and

$$\lim_{m\to\infty} [W(\mathbf{q}^{(m)}, \mathbf{f}^{(m)}) - W^*(\mathbf{q}^{(m)})] = 0.$$

Since Λ_K is compact, we can choose a subsequence (which we still denoted as the whole sequence for simplicity) such that $c^{(m)} \equiv c^{(1)}$ and $\lim_m \mathbf{q}^{(m)} = \mathbf{q} \in \Lambda_K$. Using Lemma 27, we obtain

$$\lim_{m\to\infty} W(\mathbf{q}^{(m)}, \mathbf{f}^{(m)}) = W^*(\mathbf{q}).$$

Similar to the proof of Lemma 27, we assume that $\Psi_c \ge 0$ ($c = 1,\ldots,K$), $\mathbf{q}_1 = \cdots = \mathbf{q}_k = 0$ and $\mathbf{q}_c > 0$ ($c > k$). We obtain

$$\limsup_{m\to\infty} W(\mathbf{q}, \mathbf{f}^{(m)}) = \limsup_{m\to\infty} \sum_{c=k+1}^{K} \mathbf{q}_c^{(m)} \Psi_c(\mathbf{f}^{(m)}) \le \lim_{m\to\infty} W(\mathbf{q}^{(m)}, \mathbf{f}^{(m)}) = W^*(\mathbf{q}).$$

Note that our assumption also implies that $a_{c^{(1)}}\mathbf{q}_{c^{(1)}} \le \sup_k a_k\mathbf{q}_k - \varepsilon$ and $\mathbf{f}^{(m)}_{c^{(1)}} = \sup_k \mathbf{f}^{(m)}_k$. We have thus obtained a contradiction to the second ISC condition of $\Psi_c(\cdot)$. Therefore (20) must be valid. ■

Proof of the Theorem. We use the notations of Appendix A: let $\mathcal{X}$ be the input space, $\mathcal{Q} = \Lambda_K$ be the space of conditional probability vectors, and $\mathcal{D} = \{1,\ldots,K\}$ be the space of class labels. We let $\ell(\mathbf{q},k) = \sum_{c=1,c\neq k} a_c\mathbf{q}_c$, and thus the classification error of a decision function $p(\cdot)$ in (9) can be expressed as $\ell(p(\cdot)) = \mathbf{E}_X \ell([P(Y=c|X)]_c, p(X))$. The estimation-model space is $\Omega \subset R^K$, with decision T given by (5). The W function is given by (8). Let $v(\mathbf{q}) \equiv 1$. Then (20) implies that $\forall \varepsilon > 0, \Delta H_{\ell,W,T,v}(\varepsilon) > 0$. The theorem now follows directly from the claim of Corollary 26.

Appendix C. Infinite-Sample Inconsistency of the SVM Pairwise Comparison Method

Consider the non-differentiable SVM (hinge) loss $\phi(z) = (1-z)_+$. We show that the pairwise comparison method in (10) is not ISC with $K = 3$. More precisely, we have the following counter-example.

Proposition 29 *Let* $\mathbf{q} = [q_1, q_2, q_3]$ *with* $0 < q_3 < q_2 < q_1$ *such that* $q_1 < q_2 + q_3$ *and* $q_2 > 2q_3$. *Then* $W^*(\mathbf{q}) = W(\mathbf{q}, [1,1,0]) = 1 + q_1 + q_2 + 4q_3$.

Proof Consider $\mathbf{f} = [f_1, f_2, f_3]$. Without loss of generality, we can let $f_3 = 0$. Therefore

$$W(\mathbf{q}, \mathbf{f}) = 1 + q_1[\phi(f_1) + \phi(f_1 - f_2)] + q_2[\phi(f_2) + \phi(f_2 - f_1)] + q_3[\phi(-f_1) + \phi(-f_2)].$$

Clearly if $|f_1| > 100/q_3$ or $|f_2| > 100/q_3$, then $W(\mathbf{q}, \mathbf{f}) > 100 > W(\mathbf{q}, [0,0,0])$. Therefore the optimization of $W(\mathbf{q}, \mathbf{f})$ can be restricted to $|f_1|, |f_2| \leq 100/q_3$. It follows that $W^*(\mathbf{q})$ can be achieved at some point, still denote by $\mathbf{f} = [f_1, f_2, 0]$ such that $|f_1|, |f_2| \leq 100/q_3$.

From the order-preserving property of Theorem 5, we have $f_1 \geq f_2$, and $f_1, f_2 \geq f_3 = 0$. We can rewrite $W(\mathbf{q}, \mathbf{f})$ as

$$W(\mathbf{q}, \mathbf{f}) = 1 + q_1[\phi(f_1) + \phi(f_1 - f_2)] + q_2[\phi(f_2) + (f_1 - f_2) + 1] + q_3[f_1 + f_2 + 2].$$

If $f_2 < 1$, then

$$W(\mathbf{q}, [1 + f_1 - f_2, 1, 0]) - W(\mathbf{q}, [f_1, f_2, 0]) \leq -(q_2 - 2q_3)(1 - f_2) < 0.$$

Therefore we can assume that $f_1 \geq f_2 \geq 1$. Now

$$W(\mathbf{q}, \mathbf{f}) = 1 + q_1\phi(f_1 - f_2) + q_2[f_1 - f_2 + 1] + q_3[f_1 + f_2 + 2].$$

Since $q_1 < q_2 + q_3$, we have $q_1\phi(f_1 - f_2) + (q_2 + q_3)[f_1 - f_2] \geq q_1$, and the equality holds only when $f_1 = f_2$. Therefore $W(\mathbf{q}, \mathbf{f}) \geq 1 + q_1 + q_2[0 + 1] + q_3[2f_2 + 2]$, and the minimum can only be achieved at $f_1 = f_2 = 1$. ■

References

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. Technical Report 638, Statistics Department, University of California, Berkeley, 2003.

Gilles Blanchard, Olivier Bousquet, and Pascal Massart. Statistical performance of support vector machines. http://www.kyb.mpg.de/publications/pss/ps2731.ps, 2004.

Gilles Blanchard, Gabor Lugosi, and Nicolas Vayatis. On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4:861–894, 2003.

V. Blanz, V. Vapnik, and C. Burges. Multiclass discrimination with an extended support vector machine. Talk given at AT&T Bell Labs, 1995.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

Ilya Desyatnikov and Ron Meir. Data-dependent bounds for multi-category classification based on convex losses. In *COLT*, 2003.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000. With discussion.

W. Jiang. Process consistency for adaboost. *The Annals of Statistics*, 32:13–29, 2004. with discussion.

Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of American Statistical Association*, 99:67–81, 2004.

Yi Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, pages 259–275, 2002.

Yufeng Liu and Xiaotong Shen. On multicategory ψ-learning and support vector machine. Private Communication, 2004.

G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting methods. *The Annals of Statistics*, 32:30–55, 2004. with discussion.

E. Mammen and A. Tsybakov. Smooth discrimination analysis. *Annals of Statis.*, 27:1808–1829, 1999.

Shie Mannor, Ron Meir, and Tong Zhang. Greedy algorithms for classification - consistency, convergence rates, and adaptivity. *Journal of Machine Learning Research*, 4:713–741, 2003.

Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37:297–336, 1999.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

Clint Scovel and Ingo Steinwart. Fast rates for support vector machines. Technical Report LA-UR 03-9117, Los Alamos National Laboratory, 2003.

Ingo Steinwart. Support vector machines are universally consistent. *J. Complexity*, 18:768–791, 2002.

Ingo Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4: 1071–1105, 2003.

Ingo Steinwart. Consistency of support vector machines and other regularized kernel machines. *IEEE Transactions on Information Theory*, 2004. to appear.

S. A. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3.

V. N. Vapnik. *Statistical learning theory*. John Wiley & Sons, New York, 1998.

Grace Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference series in applied mathematics. SIAM, 1990.

J. Weston and C. Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Royal Holloway, 1998.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statitics*, 32:56–85, 2004. with discussion.