On the Consistency of Multiclass Classification Methods

Ambuj Tewari   AMBUJ@CS.BERKELEY.EDU
Division of Computer Science
University of California
Berkeley, CA 94720-1776, USA

Peter L. Bartlett   BARTLETT@CS.BERKELEY.EDU
Division of Computer Science and Department of Statistics
University of California
Berkeley, CA 94720-1776, USA

Editor: Peter Auer

Abstract

Binary classification is a well studied special case of the classification problem. Statistical properties of binary classifiers, such as consistency, have been investigated in a variety of settings. Binary classification methods can be generalized in many ways to handle multiple classes. It turns out that one can lose consistency in generalizing a binary classification method to deal with multiple classes. We study a rich family of multiclass methods and provide a necessary and sufficient condition for their consistency. We illustrate our approach by applying it to some multiclass methods proposed in the literature.

Keywords: multiclass classification, consistency, Bayes risk

1. Introduction

We consider the problem of classification in a probabilistic setting: n i.i.d. pairs are generated by a probability distribution on $X \times \mathcal{Y}$. We think of y_i in a pair (x_i, y_i) as being the *label* or *class* of the example x_i. The $|\mathcal{Y}| = 2$ case is referred to as binary classification. A number of methods for binary classification involve finding a real valued function f which minimizes an empirical average of the form

$$\frac{1}{n} \sum_i \Psi_{y_i}(f(x_i)) . \tag{1}$$

In addition, some sort of regularization may be used to avoid overfitting. Typically, the sign of $f(x)$ is used to classify an unseen example x. We interpret $\Psi_y(f(x))$ as being the loss associated with predicting the label of x using $f(x)$ when the true label is y. An important special case of these methods is that of large margin methods which use $\{+1, -1\}$ as the set of labels and $\phi(yf(x))$ as the loss. Here ϕ is some function chosen by taking into account both computational and statistical issues. Bayes consistency of these methods has been analyzed in the literature (see Jiang, 2004; Lugosi and Vayatis, 2004; Zhang, 2004c; Steinwart, 2005; Bartlett et al., 2004). In this paper, we investigate the consistency of multiclass ($|\mathcal{Y}| \geq 2$) methods which try to generalize (1) by replacing f with a vector function $\mathbf{f}$. This category includes the methods of Bredensteiner and Bennett (1999); Lee et al. (2004); Weston and Watkins (1998) and Zhang (2004a). Zhang (2004a,b) has already initiated the study of these methods.

Under suitable conditions, minimizing (1) over a sequence of function classes also approximately minimizes the "Ψ-risk" $R_\Psi(\mathbf{f}) = \mathbb{E}_{XY}[\Psi_y(\mathbf{f}(x))]$. However, our aim in classification is to find a function $\mathbf{f}$ whose probability of misclassification $R(\mathbf{f})$ (often called the "risk" of $\mathbf{f}$) is close to the minimum possible (the so called Bayes risk R^*). Thus, it is natural to investigate the conditions which guarantee that if the Ψ-risk of $\mathbf{f}$ gets close to the optimal then the risk of $\mathbf{f}$ also approaches the Bayes risk. If this happens, we say that the classification method based on Ψ is *Bayes consistent*. Bartlett et al. (2004) defined the notion of "classification calibration" to obtain such conditions for binary classification. The authors also gave a simple characterization of classification calibration for convex loss functions. In Section 1.1 below, we provide a different point of view for looking at classification calibration for binary classification in order to motivate our geometric approach to multiclass classification.

The rest of the paper is organized as follows. Section 2 defines classification calibration in the setting of multiclass classification and provides a justification, in the form of Theorem 2, for studying it: classification calibration is equivalent to Bayes consistency. The main result in Section 3 is Theorem 7 which characterizes classification calibration in terms of geometric properties of some sets associated with the loss function Ψ. In Section 4, we provide certain sufficient conditions for classification calibration. These are provided with the hope that, in practice, some of these conditions will hold and checking the full condition of Theorem 7 can be avoided. Section 5 applies the results obtained in the paper to examine the consistency of a few multiclass methods. Interestingly, many seemingly natural generalizations of binary methods do not lead to consistent multiclass methods. Section 6 provides the conclusion.

1.1 Consistency of Binary Classification Methods

If we have a convex loss function $\phi : \mathbb{R} \mapsto [0, \infty)$ which is differentiable at 0 and $\phi'(0) < 0$, then it is known (Bartlett et al., 2004) that any minimizer f^* of

$$\mathbb{E}_{XY}[\phi(yf(x))] = \mathbb{E}_X[E_{Y|x}[\phi(yf(x))]] \tag{2}$$

yields a Bayes consistent classifier, that is, $P(Y = +1|X = x) > 1/2 \Rightarrow f^*(x) > 0$ and $P(Y = -1|X = x) < 1/2 \Rightarrow f^*(x) < 0$. In order to motivate the approach of the next section let us work with a few examples. Let us fix an x and denote the two conditional probabilities by p_+ and p_-. We also omit the argument in $f(x)$. We can then write the inner conditional expectation in (2) as

$$p_+ \phi(f) + p_- \phi(-f) \ .$$

We wish to find an f which minimizes the expression above. If we define the set $\mathcal{R} \in \mathbb{R}^2$ as

$$\mathcal{R} = \{(\phi(f), \phi(-f)) \ : \ f \in \mathbb{R}\} \ , \tag{3}$$

then the above minimization can be written as

$$\min_{\mathbf{z} \in \mathcal{R}} \langle \mathbf{p}, \mathbf{z} \rangle \tag{4}$$

where $\mathbf{p} = (p_+, p_-)$.

The set $\mathcal{R}$ is shown in Figure 1(a) for the squared hinge loss function $\phi(t) = ((1-t)_+)^2$. Geometrically, the solution to (4) is obtained by taking a line whose equation is $\langle \mathbf{p}, \mathbf{z} \rangle = c$ and then
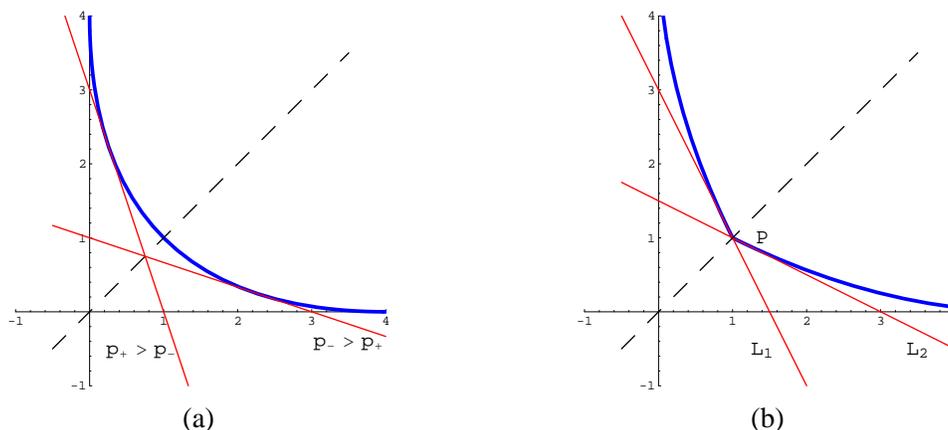
Figure 1: (a) Squared Hinge Loss (b) Inconsistent Case (the thick curve is the set $\mathcal{R}$ in both plots)

sliding it (by varying c) until it just touches $\mathcal{R}$. It is intuitively clear from the figure that if $p_+ > p_-$ then the line is inclined more towards the vertical axis and the point of contact is above the angle bisector of the axes. Similarly, if $p_+ < p_-$ then the line is inclined more towards the horizontal axis and the point is below the bisector. This means that $\text{sign}(\phi(-f) - \phi(f))$ is a consistent classification rule which, because ϕ is a decreasing function, is equivalent to $\text{sign}(f - (-f)) = \text{sign}(f)$. In fact, the condition $\phi'(0) < 0$ is not necessary for the existence of a consistent classification rule based on f. For example, if we had the function $\phi(t) = ((1+t)_+)^2$, we would still get the same set $\mathcal{R}$ but will need to change the classification rule to $\text{sign}(-f)$ in order to preserve consistency.

Why do we need differentiability of ϕ at 0? Figure 1(b) shows the set $\mathcal{R}$ for a convex loss function which is not differentiable at 0. In this case, both lines L_1 and L_2 touch $\mathcal{R}$ at P but L_1 has $p_+ > p_-$ while L_2 has $p_+ < p_-$. Thus we cannot create a consistent classifier based on this loss function. The crux of the problem seems to lie in the fact that there are two distinct supporting lines to the set $\mathcal{R}$ at P and that these two lines are inclined towards different axes.

It seems from the figures that as long as $\mathcal{R}$ is symmetric about the angle bisector of the axes, all supporting lines at a given point are inclined towards the same axis except when the point happens to lie on the angle bisector. To check for consistency, we need to examine the set of supporting lines only at that point. In case the set $\mathcal{R}$ is generated as in (3), this boils down to checking the differentiability of ϕ at 0. In the following sections, we will deal with cases when the set $\mathcal{R}$ is generated in a more general way and the situation possibly involves more than two dimensions. The intuition developed for the binary case will be proven correct by Propositions 9 and 10 in Section 4.

2. Classification Calibration and its Relation to Consistency

Suppose we have $K \geq 2$ classes. For $y \in \{1, \ldots, K\}$, let Ψ_y be a continuous function from $\mathbb{R}^K$ to $\mathbb{R}_+ = [0, \infty)$. Let $\mathcal{F}$ be a class of vector functions $\mathbf{f} : \mathcal{X} \mapsto \mathbb{R}^K$. Let $\{\mathcal{F}_n\}$ be a sequence of function classes such that each $\mathcal{F}_n \subseteq \mathcal{F}$. Suppose $\hat{\mathbf{f}}_n$ is a classifier learned from data. For example, we might obtain $\hat{\mathbf{f}}_n$ by minimizing the empirical Ψ-risk $\hat{R}_\Psi$ over the class $\mathcal{F}_n$,

$$\hat{\mathbf{f}}_n = \arg\min_{\mathbf{f} \in \mathcal{F}_n} \hat{R}_\Psi(\mathbf{f}) = \arg\min_{\mathbf{f} \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} \Psi_{y_i}(\mathbf{f}(x_i)) .$$

There might be some constraints on the set of vector functions in the class $\mathcal{F}$. For example, a common constraint is to have the components of $\mathbf{f}$ sum to zero. More generally, let us assume there is some set $C \in \mathbb{R}^K$ such that

$$\mathcal{F} = \{\mathbf{f} : \forall x, \, \mathbf{f}(x) \in C\}. \tag{5}$$

Let $\Psi(\mathbf{f}(x))$ denote the vector $(\Psi_1(\mathbf{f}(x)), \ldots, \Psi_K(\mathbf{f}(x)))^T$. We predict the label of a new example x to be $\text{pred}(\Psi(\mathbf{f}(x)))$ for some function $\text{pred} : \mathbb{R}^K \mapsto \{1, \ldots, K\}$. The Ψ-risk of a function $\mathbf{f}$ is

$$R_\Psi(\mathbf{f}) = \mathbb{E}_{\mathcal{X}\mathcal{Y}}[\Psi_y(\mathbf{f}(x))],$$

and we denote the least possible Ψ-risk by

$$R_\Psi^* = \inf_{\mathbf{f} \in \mathcal{F}} R_\Psi(\mathbf{f}).$$

In a classification task, we are more interested in the risk of a function $\mathbf{f}$,

$$R(\mathbf{f}) = \mathbb{E}_{\mathcal{X}\mathcal{Y}}[\mathbf{1}[\text{pred}(\Psi(\mathbf{f}(x))) \neq Y]],$$

which is the probability that $\mathbf{f}$ leads to an incorrect prediction on a labeled example drawn from the underlying probability distribution. The least possible risk is

$$R^* = \mathbb{E}_{\mathcal{X}}[1 - \max_y p_y(x)],$$

where $p_y(x) = P(Y = y \mid X = x)$. If $\hat{\mathbf{f}}_n$ is the empirical Ψ-risk minimizer then, under suitable conditions, one would expect $R_\Psi(\hat{\mathbf{f}}_n)$ to converge to R_Ψ^* (in probability). It would be nice if that made $R(\hat{\mathbf{f}}_n)$ converge to R^* (in probability). Theorem 2 below states that classification calibration, a notion that we will soon define, is both necessary and sufficient for this to happen.

In order to understand the behavior of approximate Ψ-risk minimizers, let us write $R_\Psi(\mathbf{f})$ as

$$\mathbb{E}_{\mathcal{X}\mathcal{Y}}[\Psi_y(\mathbf{f}(x))] = \mathbb{E}_{\mathcal{X}}[\mathbb{E}_{\mathcal{Y}|x}[\Psi_y(\mathbf{f}(x))]].$$

The above minimization problem is equivalent to minimizing the inner conditional expectation for each $x \in \mathcal{X}$. Let us fix an arbitrary x for now, so we can write $\mathbf{f}$ instead of $\mathbf{f}(x)$, p_y instead of $p_y(x)$, etc. The minimum might not be achieved and so we consider the infimum[1] $\inf_{\mathbf{f} \in C} \sum_y p_y \Psi_y(\mathbf{f})$ of the conditional expectation above. Define the subsets $\mathcal{R}$ and $\mathcal{S}$ of $\mathbb{R}_+^K$ as

$$\mathcal{R} = \{(\Psi_1(\mathbf{f}), \ldots, \Psi_K(\mathbf{f})) : \mathbf{f} \in C\},$$

$$\mathcal{S} = \text{conv}(\mathcal{R}) = \text{conv}\{(\Psi_1(\mathbf{f}), \ldots, \Psi_K(\mathbf{f})) : \mathbf{f} \in C\}, \tag{6}$$

where $\text{conv}(\mathcal{R})$ denotes the convex hull of $\mathcal{R}$. We then have

$$\inf_{\mathbf{f} \in C} \sum_y p_y \Psi_y(\mathbf{f}) = \inf_{\mathbf{f} \in C} \langle \mathbf{p}, \Psi(\mathbf{f}) \rangle$$
$$= \inf_{\mathbf{z} \in \mathcal{R}} \langle \mathbf{p}, \mathbf{z} \rangle \tag{7}$$
$$= \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle,$$

---

1. Since p_y and $\Psi_y(\mathbf{f})$ are both non-negative, the objective function is bounded below by 0 and hence the existence of an infimum is guaranteed.

where $\mathbf{p} = (p_1, \ldots, p_K)$. The last equality holds because for a fixed $\mathbf{p}$, the function $\mathbf{z} \mapsto \langle \mathbf{p}, \mathbf{z} \rangle$ is a linear function and hence we do not change the infimum by taking the convex hull[2] of $\mathcal{R}$.

Let us define a *symmetric* set to be one with the following property: if a point $\mathbf{z}$ is in the set then so is any point obtained by interchanging any two coordinates of $\mathbf{z}$. We assume that $\mathcal{R}$ is symmetric. This assumption holds whenever the loss function treats all classes equivalently. Note that our assumption about $\mathcal{R}$ implies that $\mathcal{S}$ too is symmetric.

We now define classification calibration of $\mathcal{S}$. The definition intends to capture the property that, for any $\mathbf{p}$, minimizing $\langle \mathbf{p}, \mathbf{z} \rangle$ over $\mathcal{S}$ leads one to $\mathbf{z}$'s which enable us to figure out the index of (one of the) maximum coordinate(s) of $\mathbf{p}$.

Definition 1 *A set $\mathcal{S} \subseteq \mathbb{R}_+^K$ is **classification calibrated** if there exists a predictor function* pred : $\mathbb{R}^K \mapsto \{1, \ldots, K\}$ *such that*

$$\forall \mathbf{p} \in \Delta_K, \quad \inf_{\mathbf{z} \in S \,:\, p_{\text{pred}(\mathbf{z})} < \max_y p_y} \langle \mathbf{p}, \mathbf{z} \rangle > \inf_{\mathbf{z} \in S} \langle \mathbf{p}, \mathbf{z} \rangle \, , \tag{8}$$

where Δ_K is the probability simplex in $\mathbb{R}^K$.

The following theorem tells us that classification calibration is indeed the right property to study, as it is both necessary and sufficient for convergence of Ψ-risk to the optimal Ψ-risk to imply convergence of risk to the Bayes risk. The convergence of the risk of a classifier to the Bayes risk is often referred to as its consistency.

Theorem 2 *Let Ψ be a (vector-valued) loss function and $\mathcal{C}$ be a subset of $\mathbb{R}^K$. Let $\mathcal{F}$ and $\mathcal{S}$ be as defined in (5) and (6) respectively. Then $\mathcal{S}$ is classification calibrated iff the following holds. Whenever $\{\mathcal{F}_n\}$ is a sequence of function classes (where $\mathcal{F}_n \subseteq \mathcal{F}$ and $\cup \mathcal{F}_n = \mathcal{F}$) such that $\hat{\mathbf{f}}_n \in \mathcal{F}_n$ and P is the data generating probability distribution,*

$$R_\Psi(\hat{\mathbf{f}}_n) \xrightarrow{P} R_\Psi^*$$

implies

$$R(\hat{\mathbf{f}}_n) \xrightarrow{P} R^* \, .$$

Proof See Appendix A. ∎

The definition of classification calibration as stated above is not concrete enough to be of any use in checking the consistency of a multiclass method corresponding to a given loss function. We now study the property of classification calibration with the aim of arriving at a characterization expressed in terms of concretely verifiable properties of the loss function. Before we do that, let us observe that it is easy to reformulate the definition in terms of sequences. The exact statement of the reformulation is provided by the following lemma whose proof, being entirely straightforward, is omitted.

Lemma 3 *$\mathcal{S} \subseteq \mathbb{R}_+^K$ is classification calibrated iff there exists a predictor function* pred : $\mathbb{R}^K \mapsto \{1, \ldots, K\}$ *such that the following holds: $\forall \mathbf{p} \in \Delta_K$ and all sequences $\{\mathbf{z}^{(n)}\}$ in $\mathcal{S}$ such that*

$$\langle \mathbf{p}, \mathbf{z}^{(n)} \rangle \to \inf_{\mathbf{z} \in S} \langle \mathbf{p}, \mathbf{z} \rangle \, , \tag{9}$$

---

2. If $\mathbf{z}$ is a convex combination of $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$, then $\langle \mathbf{p}, \mathbf{z} \rangle \geq \min\{\langle \mathbf{p}, \mathbf{z}^{(1)} \rangle, \langle \mathbf{p}, \mathbf{z}^{(2)} \rangle\}$.

we have

$$p_{\mathrm{pred}(\mathbf{z}^{(n)})} = \max_y p_y \tag{10}$$

ultimately.[3]

This makes it easier to see that if $\mathcal{S}$ is classification calibrated then we can find a predictor function such that any sequence achieving the infimum in (7) ultimately predicts the right label (the one having maximum probability). The following lemma shows that symmetry of our set $\mathcal{S}$ allows us to reduce the search space of predictor functions (namely to those functions which map $\mathbf{z}$ to the index of a minimum coordinate).

Lemma 4 *Suppose $\mathcal{S} \subseteq \mathbb{R}_+^K$ is symmetric. If there exists a predictor function* pred *satisfying condition* (8) *in the definition of classification calibration (Definition 1), then any predictor function* pred′ *satisfying*

$$\forall \mathbf{z} \in \mathcal{S}, \ z_{\mathrm{pred}'(\mathbf{z})} = \min_y z_y \tag{11}$$

also satisfies (8).

Proof Consider some $\mathbf{p} \in \Delta_K$ and a sequence $\{\mathbf{z}^{(n)}\}$ such that (9) holds. We have $p_{\mathrm{pred}(\mathbf{z}^{(n)})} = \max_y p_y$ ultimately. In order to derive a contradiction, assume that $p_{\mathrm{pred}'(\mathbf{z}^{(n)})} < \max_y p_y$ infinitely often. Since there are finitely many labels, this implies that there is a subsequence $\{\mathbf{z}^{(n_k)}\}$ and labels M and m such that the following hold,

$$\mathrm{pred}(\mathbf{z}^{(n_k)}) = M \in \{y' : p_{y'} = \max_y p_y\} \ ,$$

$$\mathrm{pred}'(\mathbf{z}^{(n_k)}) = m \in \{y' : p_{y'} < \max_y p_y\} \ ,$$

$$\langle \mathbf{p}, \mathbf{z}^{(n_k)} \rangle \to \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle \ .$$

Because of (11), we also have $z_M^{(n_k)} \geq z_m^{(n_k)}$. Let $\tilde{\mathbf{p}}$ and $\tilde{\mathbf{z}}$ denote the vectors obtained from $\mathbf{p}$ and $\mathbf{z}$ respectively by interchanging the M and m coordinates. Since $\mathcal{S}$ is symmetric, $\mathbf{z} \in \mathcal{S} \Leftrightarrow \tilde{\mathbf{z}} \in \mathcal{S}$. There are two cases depending on whether the inequality in

$$\liminf_k \left(z_M^{(n_k)} - z_m^{(n_k)} \right) \geq 0$$

is strict or not.

If it is strict, denote the value of the lim inf by $\varepsilon > 0$. Then $z_M^{(n_k)} - z_m^{(n_k)} > \varepsilon/2$ ultimately and hence we have

$$\langle \mathbf{p}, \mathbf{z}^{(n_k)} \rangle - \langle \mathbf{p}, \tilde{\mathbf{z}}^{(n_k)} \rangle = (p_M - p_m)(z_M^{(n_k)} - z_m^{(n_k)}) > (p_M - p_m)\varepsilon/2$$

for k large enough. This implies $\liminf_{k \to \infty} \langle \mathbf{p}, \tilde{\mathbf{z}}^{(n_k)} \rangle < \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle$, which is a contradiction.

Otherwise, choose a subsequence[4] $\{\mathbf{z}^{(n_k)}\}$ such that $\lim(z_M^{(n_k)} - z_m^{(n_k)}) = 0$. Multiplying this with $(p_M - p_m)$, we have

$$\lim_{k \to \infty} \left(\langle \tilde{\mathbf{p}}, \tilde{\mathbf{z}}^{(n_k)} \rangle - \langle \tilde{\mathbf{p}}, \mathbf{z}^{(n_k)} \rangle \right) = 0 \ .$$

---

3. Throughout the paper, use of "ultimately" in the context of a sequence $\{\mathbf{z}^{(n)}\}$ means "for all sufficiently large values of n".

4. We do not introduce additional subscripts for simplicity.

We also have

$$\lim_{k\to\infty} \langle \tilde{\mathbf{p}}, \tilde{\mathbf{z}}^{(n_k)} \rangle = \lim_{k\to\infty} \langle \mathbf{p}, \mathbf{z}^{(n_k)} \rangle = \inf_{\mathbf{z}\in\mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle = \inf_{\mathbf{z}\in\mathcal{S}} \langle \tilde{\mathbf{p}}, \tilde{\mathbf{z}} \rangle = \inf_{\mathbf{z}\in\mathcal{S}} \langle \tilde{\mathbf{p}}, \mathbf{z} \rangle ,$$

where the last equality follows because of symmetry. This means

$$\langle \tilde{\mathbf{p}}, \mathbf{z}^{(n_k)} \rangle \to \inf_{\mathbf{z}\in\mathcal{S}} \langle \tilde{\mathbf{p}}, \mathbf{z} \rangle$$

as $k \to \infty$ and therefore

$$\tilde{p}_{\mathrm{pred}(\mathbf{z}^{(n_k)})} = p_M$$

ultimately. This is a contradiction since $\tilde{p}_{\mathrm{pred}(\mathbf{z}^{(n_k)})} = \tilde{p}_M = p_m$. ■

Having identified the class of potentially useful predictor functions, we will henceforth assume that pred is defined as in (11).

3. A Characterization of Classification Calibration

We give a characterization of classification calibration in terms of normals to the convex set $\mathcal{S}$ and its projections onto lower dimensions. For a point $\mathbf{z} \in \partial\mathcal{S}$, we say $\mathbf{p}$ is a normal to $\mathcal{S}$ at $\mathbf{z}$ if[5] $\langle \mathbf{z}' - \mathbf{z}, \mathbf{p} \rangle \geq 0$ for all $\mathbf{z}' \in \mathcal{S}$. Define the set of positive normals at $\mathbf{z}$ as

$$\mathcal{N}(\mathbf{z}) = \{\mathbf{p} : \mathbf{p} \text{ is a normal to } \mathcal{S} \text{ at } \mathbf{z}\} \cap \Delta_K .$$

Definition 5 *A convex set $\mathcal{S} \subseteq \mathbb{R}_+^K$ is* **admissible** *if $\forall \mathbf{z} \in \partial\mathcal{S}, \forall \mathbf{p} \in \mathcal{N}(\mathbf{z})$, we have*

$$\mathrm{argmin}(\mathbf{z}) \subseteq \mathrm{argmax}(\mathbf{p}) \tag{12}$$

where $\mathrm{argmin}(\mathbf{z}) = \{y' : z_{y'} = \min_y z_y\}$ and $\mathrm{argmax}(\mathbf{p}) = \{y' : p_{y'} = \max_y p_y\}$.

Lemma 6 *If $\mathcal{S} \subseteq \mathbb{R}_+^K$ is admissible then for all $\mathbf{p} \in \Delta_K$ and all* bounded *sequences $\{\mathbf{z}^{(n)}\}$ such that $\langle \mathbf{p}, \mathbf{z}^{(n)} \rangle \to \inf_{\mathbf{z}\in\mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle$, we have $p_{\mathrm{pred}(\mathbf{z}^{(n)})} = \max_y p_y$ ultimately.*

Proof Let $Z(\mathbf{p}) = \{\mathbf{z} \in \partial\mathcal{S} : \mathbf{p} \in \mathcal{N}(\mathbf{z})\}$. Taking the limit of a convergent subsequence of the given bounded sequence gives us a point in $\partial\mathcal{S}$ which achieves the infimum of the inner product with $\mathbf{p}$. Thus, $Z(\mathbf{p})$ is not empty. It is easy to see that $Z(\mathbf{p})$ is closed. For a point $\mathbf{z}$ and a set Z, define

$$\mathrm{dist}(z, Z) = \inf_{\mathbf{z}'\in Z} \|\mathbf{z} - \mathbf{z}'\| ,$$

to be the distance of $\mathbf{z}$ from Z. We claim that for all $\varepsilon > 0$, $\mathrm{dist}(\mathbf{z}^{(n)}, Z(\mathbf{p})) < \varepsilon$ ultimately. For if we assume the contrary, boundedness implies that we can find a convergent subsequence $\{\mathbf{z}^{(n_k)}\}$ such that $\forall k$, $\mathrm{dist}(\mathbf{z}^{(n_k)}, Z(\mathbf{p})) \geq \varepsilon$. Let $\mathbf{z}^* = \lim_{k\to\infty} \mathbf{z}^{(n_k)}$. Then $\langle \mathbf{p}, \mathbf{z}^* \rangle = \inf_{\mathbf{z}\in\mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle$ and so $\mathbf{z}^* \in Z(\mathbf{p})$. On the other hand, $\mathrm{dist}(\mathbf{z}^*, Z(\mathbf{p})) \geq \varepsilon$ which gives us a contradiction and our claim is proved.

Now we claim that there exists $\varepsilon' > 0$ such that $\mathrm{dist}(\mathbf{z}^{(n)}, Z(\mathbf{p})) < \varepsilon'$ implies[6]

$$\mathrm{argmin}(\mathbf{z}^{(n)}) \subseteq \mathrm{argmin}(Z(\mathbf{p})) .$$

---

5. Our sign convention is opposite to the usual one (Rockafellar, 1970, see, for example,) because we are dealing with minimum (instead of maximum) problems.

6. For a set Z, $\mathrm{argmin}(Z)$ denotes $\cup_{\mathbf{z}\in Z} \mathrm{argmin}(\mathbf{z})$.

Assume, for sake of a contradiction, that there is a convergent subsequence $\mathbf{z}^{(n_k)}$ such that

$$\text{dist}(\mathbf{z}^{(n_k)}, Z(\mathbf{p})) \to 0 \text{ as } k \to \infty \,,$$

and

$$\forall k, \text{ argmin}(\mathbf{z}^{(n_k)}) \not\subseteq \text{argmin}(Z(\mathbf{p})) \,. \tag{13}$$

Denote the limit by $\mathbf{z}^*$. Since $\text{dist}(\cdot, Z(\mathbf{p}))$ is continuous, $\text{dist}(\mathbf{z}^*, Z(\mathbf{p})) = 0$ which implies that $\mathbf{z}^* \in Z(\mathbf{p})$ as $Z(\mathbf{p})$ is closed. Moreover, for large enough k,

$$\text{argmin}(\mathbf{z}^{(n_k)}) \subseteq \text{argmin}(\mathbf{z}^*) \subseteq \text{argmin}(Z(\mathbf{p})) \,,$$

where the last inclusion holds because $\mathbf{z}^* \in Z(\mathbf{p})$. This contradicts (13). Hence the claim is proved.

Finally, by admissibility of $\mathcal{S}$, $\text{argmin}(Z(\mathbf{p})) \subseteq \text{argmax}(\mathbf{p})$ and so $\text{argmin}(\mathbf{z}^{(n)}) \subseteq \text{argmax}(\mathbf{p})$ ultimately. ∎

The next theorem provides a characterization of classification calibration in terms of normals to $\mathcal{S}$.

Theorem 7 *Let $\mathcal{S} \subseteq \mathbb{R}_+^K$ be a symmetric convex set. Define the projections*

$$\mathcal{S}^{(i)} = \{(z_1, \ldots, z_i)^T : \mathbf{z} \in \mathcal{S}\}$$

for $i \in \{2, \ldots, K\}$. Then $\mathcal{S}$ is classification calibrated iff each $\mathcal{S}^{(i)}$ is admissible.

Proof We prove the easier 'only if' direction first. Suppose some $\mathcal{S}^{(i)}$ is not admissible. Then there exist $\mathbf{z} \in \partial \mathcal{S}^{(i)}$ and $\mathbf{p} \in \mathcal{N}(\mathbf{z})$ and a label y' such that $y' \in \text{argmin}(\mathbf{z})$ and $y' \notin \text{argmax}(\mathbf{p})$. Choose a sequence $\{\mathbf{z}^{(n)}\}$ converging to $\mathbf{z}$. Modify the sequence by replacing, in each $\mathbf{z}^{(n)}$, the coordinates specified by $\text{argmin}(\mathbf{z})$ by their average. The resulting sequence is still in $\mathcal{S}^{(i)}$ (by symmetry and convexity) and has $\text{argmin}(\mathbf{z}^{(n)}) = \text{argmin}(\mathbf{z})$ ultimately. Therefore, if we set $\text{pred}(\mathbf{z}^{(n)}) = y'$, we have $p_{\text{pred}(\mathbf{z}^{(n)})} < \max_y p_y$ ultimately. To get a sequence in $\mathcal{S}$ look at the points whose projections are the $\mathbf{z}^{(n)}$'s and pad $\mathbf{p}$ with $K - i$ zeros.

To prove the other direction, assume each $\mathcal{S}^{(i)}$ is admissible. Consider a sequence $\{\mathbf{z}^{(n)}\}$ with $\langle \mathbf{p}, \mathbf{z}^{(n)} \rangle \to \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle = L$. Without loss of generality, assume that for some $j, 1 \le j \le K$ we have $p_1, \ldots, p_j > 0$ and $p_{j+1}, \ldots, p_K = 0$. We claim that there exists an $M < \infty$ such that $\forall y \le j, z_y^{(n)} \le M$ ultimately. Since $p_j z_j^{(n)} \le L + 1$ ultimately, $M = \max_{1 \le y \le j}\{(L+1)/p_y\}$ works. Consider a set of labels $T \subseteq \{j+1, \ldots, K\}$. Consider the subsequence consisting of those $\mathbf{z}^{(n)}$ for which $z_y^{(n)} \le M$ for $y \in \{1, \ldots, j\} \cup T$ and $z_y^{(n)} > M$ for $y \in \{j+1, \ldots, K\} - T$. The original sequence can be decomposed into finitely many such subsequences corresponding to the $2^{(K-j)}$ choices of the set T. Fix T and convert the corresponding subsequence into a sequence in $\mathcal{S}^{(j+|T|)}$ by dropping the coordinates belonging to the set $\{j+1, \ldots, K\} - T$. Call this sequence $\tilde{\mathbf{z}}^{(n)}$ and let $\tilde{\mathbf{p}}$ be $(p_1, \ldots, p_j, 0, \ldots, 0)^T$. We have a bounded sequence with

$$\langle \tilde{\mathbf{p}}, \tilde{\mathbf{z}}^{(n)} \rangle \to \inf_{\tilde{\mathbf{z}} \in \mathcal{S}^{(j+|T|)}} \langle \tilde{\mathbf{p}}, \tilde{\mathbf{z}} \rangle \,.$$

Thus, by Lemma 6, we have $\tilde{p}_{\text{pred}(\tilde{\mathbf{z}}^{(n)})} = \max_y \tilde{p}_y = \max_y p_y$ ultimately. Since we dropped only those coordinates which were greater than M, $\text{pred}(\tilde{\mathbf{z}}^{(n)})$ picks the same coordinate as $\text{pred}(\mathbf{z}^{(n)})$ where $\mathbf{z}^{(n)}$ is the element from which $\tilde{\mathbf{z}}^{(n)}$ was obtained. Thus we have $p_{\text{pred}(\mathbf{z}^{(n)})} = \max_y p_y$ ultimately and the theorem is proved. ∎

4. Sufficient Conditions for Classification Calibration

In this section, we prove some propositions that reduce the work involved in checking classification calibration of sets. We will see that the assumptions made in the propositions are often satisfied. Our first proposition states that in the presence of symmetry, points having a unique minimum coordinate can never destroy admissibility. Before that, we need the following lemma.

Lemma 8 *Let $S \subseteq \mathbb{R}_+^K$ be a symmetric convex set, $\mathbf{z}$ a point in the boundary of S and $\mathbf{p} \in \mathcal{N}(\mathbf{z})$. Let $y, y' \in \{1, \ldots, K\}$. Then*

$$(z_{y'} - z_y)(p_y - p_{y'}) \geq 0 .$$

Proof We have $\langle \mathbf{z}' - \mathbf{z}, \mathbf{p} \rangle \geq 0$ for all $\mathbf{z}' \in S$. Taking limits, this inequality also holds for $\mathbf{z}' \in \partial S$. Set $\mathbf{z}'$ to be $\mathbf{z}$ with its y, y' coordinates interchanged. By symmetry $\mathbf{z}' \in \partial S$. Therefore $\langle \mathbf{z}' - \mathbf{z}, \mathbf{p} \rangle \geq 0$ which implies, since $\mathbf{z}, \mathbf{z}'$ agree on all but the y, y' coordinates,

$$(z_{y'} - z_y)p_y + (z_y - z_{y'})p_{y'} \geq 0 .$$

∎

Proposition 9 *Let $S \subseteq \mathbb{R}_+^K$ be a symmetric convex set, $\mathbf{z}$ a point in the boundary of S and $\mathbf{p} \in \mathcal{N}(\mathbf{z})$. Then* $\operatorname{argmin}(\mathbf{z}) \subseteq \operatorname{argmax}(\mathbf{p})$ *whenever* $|\operatorname{argmin}(\mathbf{z})| = 1$.

Proof For any y, y', Lemma 8 gives us $(z_{y'} - z_y)(p_y - p_{y'}) \geq 0$. Thus $z_y < z_{y'}$ implies $p_y \geq p_{y'}$. Therefore, if $|\operatorname{argmin}(\mathbf{z})| = 1$ and $z_y = \min_{y'} z_{y'}$ then $p_y \geq p_{y'}$ for all y', and so $\operatorname{argmin}(\mathbf{z}) \subseteq \operatorname{argmax}(\mathbf{p})$. ∎

If the set S possesses a unique normal at every point on its boundary then the next proposition guarantees admissibility.

Proposition 10 *Let $S \subseteq \mathbb{R}_+^K$ be a symmetric convex set, $\mathbf{z}$ a point in the boundary of S and $\mathcal{N}(\mathbf{z}) = \{\mathbf{p}\}$ is a singleton. Then* $\operatorname{argmin}(\mathbf{z}) \subseteq \operatorname{argmax}(\mathbf{p})$. *Thus, S is admissible if $|\mathcal{N}(\mathbf{z})| = 1$ for all $\mathbf{z} \in \partial S$.*

Proof We will assume that there exists a $y, y \in \operatorname{argmin}(\mathbf{z}), y \notin \operatorname{argmax}(\mathbf{p})$ and deduce that there are at least 2 elements in $|\mathcal{N}(\mathbf{z})|$ to get a contradiction. Let $y' \in \operatorname{argmax}(\mathbf{p})$. By Lemma 8 we have $(z_{y'} - z_y)(p_y - p_{y'}) \geq 0$ which implies $z_{y'} \leq z_y$ since $p_y - p_{y'} < 0$. But we already know that $z_y \leq z_{y'}$ and so $z_y = z_{y'}$. Symmetry of S now implies that $\tilde{\mathbf{p}} \in \mathcal{N}(\mathbf{z})$ where $\tilde{\mathbf{p}}$ is obtained from $\mathbf{p}$ by interchanging the y, y' coordinates. Since $p_y \neq p_{y'}$, $\tilde{\mathbf{p}} \neq \mathbf{p}$ which means $|\mathcal{N}(\mathbf{z})| \geq 2$. ∎

Theorem 7 provides a characterization of classification calibration in terms of admissibility of the projections $S^{(k)}$. As we will see in the examples later, proving that a certain set is not classification calibrated simply involves finding a projection $S^{(k)}$ and a "bad" point in $S^{(k)}$ violating the condition of admissibility. On the other hand, the characterization is not so easy to use when we wish to assert, rather than deny, classification calibration. The following lemma shows that under certain assumptions we can ignore the projections $S^{(k)}$ and only check the admissibility of S. That we cannot always ignore projections will become clear when we consider examples in the next section.

Recall that $\mathcal{R}$ is the set of points $(\Psi_1(\mathbf{f}), \ldots, \Psi_K(\mathbf{f}))$ as $\mathbf{f}$ ranges over $\mathcal{C}$. Define the projections $\mathcal{R}^{(k)}$ in the same way as $S^{(k)}$. Since the operations of taking projections and convex hulls commute, it is easy to see that $\operatorname{conv}(\mathcal{R}^{(k)}) = S^{(k)}$.

Theorem 11 *Suppose the set $\mathcal{R}$ is symmetric and the set $\mathcal{S}$ defined in* (6) *is admissible. Then the following holds for any k: if $\mathcal{R}^{(k)}$ is closed then $\mathcal{S}^{(k)}$ is admissible. Furthermore, if $\mathcal{R}^{(k)}$ is closed for all $k \in \{2,\ldots,K\}$ then $\mathcal{S}$ is classification calibrated.*

Proof We only need to prove that $\mathcal{R}^{(k)}$ closed and $\mathcal{S}$ admissible implies $\mathcal{S}^{(k)}$ is admissible since the last claim of the theorem follows from this and Theorem 7. Suppose $\mathcal{R}^{(k)}$ is closed and $\mathcal{S}^{(k)}$ is not admissible. Therefore, there exists positive normal $\mathbf{p}$ to $\mathcal{S}^{(k)}$ at a point $\mathbf{z}'$ in the boundary of $\mathcal{S}^{(k)}$ with

$$\operatorname{argmin}(\mathbf{z}') \not\subseteq \operatorname{argmax}(\mathbf{p}) \ . \tag{14}$$

Since $\mathbf{z}'$ is in the boundary of $\mathcal{S}^{(k)}$ and $\operatorname{conv}(\mathcal{R}^{(k)}) = \mathcal{S}^{(k)}$, there exist points $\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(l)}$ in the closure of $\mathcal{R}^{(k)}$ such that $\mathbf{p} \in \mathcal{N}(\mathbf{z}^{(i)})$ for all i, and

$$\mathbf{z}' = \sum_{i=1}^{l} \lambda_i \mathbf{z}^{(i)}$$

for some $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. It is easy to see that

$$\operatorname{argmin}(\mathbf{z}') \subseteq \bigcup_{i=1}^{l} \operatorname{argmin}(\mathbf{z}^{(i)}) \ .$$

From (14), it now follows that

$$\operatorname{argmin}(\mathbf{z}) \not\subseteq \operatorname{argmax}(\mathbf{p}) \tag{15}$$

holds for some $\mathbf{z} \in \{\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(l)}\}$. Now $\mathbf{z}$ is a limit point of $\mathcal{R}^{(k)}$. Since $\mathcal{R}^{(k)}$ is closed, $\mathbf{z} \in \mathcal{R}^{(k)}$. Let $\tilde{\mathbf{z}}$ be the point in $\mathcal{R}$ such that $\mathbf{z} = \operatorname{proj}(\tilde{\mathbf{z}})$ where proj is projection operator mapping $(z_1,\ldots,z_K)$ to $(z_1,\ldots,z_k)$. Since $\mathcal{R} \subseteq \mathcal{S}$, this gives us $\tilde{\mathbf{z}} \in \mathcal{S}$. Moreover, $\tilde{p}$ is a normal to $\mathcal{S}$ at $\tilde{\mathbf{z}}$ where $\tilde{\mathbf{p}}$ is $\mathbf{p}$ padded with $K-k$ zeros. This is because

$$\langle \tilde{\mathbf{p}}, \mathbf{z}' - \tilde{\mathbf{z}} \rangle = \langle \mathbf{p}, \operatorname{proj}(\mathbf{z}') - \mathbf{z} \rangle \geq 0 \ ,$$

for all $\mathbf{z}' \in \mathcal{S}$. We now claim that $\operatorname{argmin}(\tilde{\mathbf{z}}) \not\subseteq \operatorname{argmax}(\tilde{\mathbf{p}})$, which will give us a contradiction as $\mathcal{S}$ was assumed to be admissible. Since $\tilde{\mathbf{p}}$ is $\mathbf{p}$ padded with zeros, $\operatorname{argmax}(\tilde{\mathbf{p}}) = \operatorname{argmax}(\mathbf{p})$. Also, $\operatorname{argmin}(\tilde{\mathbf{z}}) \subseteq \operatorname{argmin}(\mathbf{z})$ for otherwise applying a suitable permutation to the coordinates of $\tilde{\mathbf{z}}$ will give a point $\hat{\mathbf{z}}$ with $\langle \mathbf{p}, \operatorname{proj}(\hat{\mathbf{z}}) \rangle < \langle \mathbf{p}, \mathbf{z} \rangle$. The claim now follows from (15). ∎

We will now provide a sufficient condition for $\mathcal{R}^{(k)}$ to be closed. To state it, we need the following definition.

Definition 12 *Let $C \subseteq \mathbb{R}^K$ be a set and $\Psi : C \mapsto \mathbb{R}_+^K$ be a loss function. Further, let $\Psi^{(k)} = (\Psi_1,\ldots,\Psi_k)$ be Ψ restricted to the first k coordinates. The mapping $\Psi^{(k)}$ is* **boundedly invertible** *iff for all $M > 0$, there is an $M' > 0$ such that $\|\mathbf{z}\| \leq M$, $\mathbf{z} \in \mathcal{R}^{(k)}$ implies there is $\mathbf{g} \in C$ with $\|\mathbf{g}\| \leq M'$ and $\Psi^{(k)}(\mathbf{g}) = \mathbf{z}$.*

$\Psi^{(k)}$ boundedly invertible roughly means that it has an inverse carrying bounded points in its range to bounded points in the domain.

Lemma 13 *Suppose that C is closed and Ψ is continuous. If $\Psi^{(k)}$ is boundedly invertible then $\mathcal{R}^{(k)}$ is closed.*
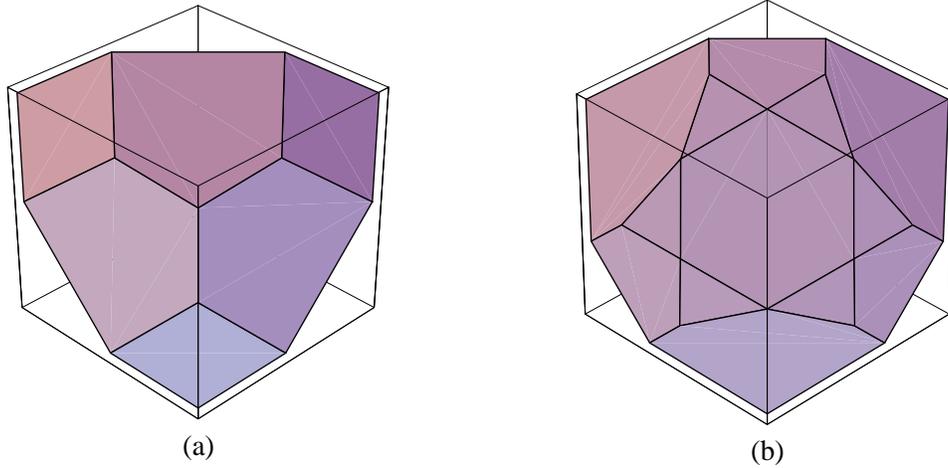
Figure 2: (a) Crammer and Singer (b) Weston and Watkins

Proof Suppose $\mathbf{z}^{(n)} \in \mathcal{R}^{(k)}$ and $\mathbf{z}^{(n)} \to \mathbf{z}$. Then these is an $M > 0$ such that $\|\mathbf{z}^{(n)}\| \leq M$ for all n. Bounded invertibility implies there is a sequence $\{\mathbf{g}^{(n)}\}$ such that, for all n, $\mathbf{g}^{(n)} \in \mathcal{C}$, $\|\mathbf{g}^{(n)}\| < M'$ and $(\Psi_1(\mathbf{g}^{(n)}), \ldots, \Psi_k(\mathbf{g}^{(n)})) = \mathbf{z}^{(n)}$. Being bounded, $\{\mathbf{g}^{(n)}\}$ has a convergent subsequence $\{\mathbf{g}^{(n_k)}\}$. Since $\mathcal{C}$ is closed, the limit $\mathbf{g}^* \in \mathcal{C}$. Now,

$$
\begin{aligned}
(\Psi_1(\mathbf{g}^*), \ldots, \Psi_k(\mathbf{g}^*)) &= \lim_{n \to \infty} (\Psi_1(\mathbf{g}^{(n_k)}), \ldots, \Psi_k(\mathbf{g}^{(n_k)})) \ (\Psi \text{ is continuous}) \\
&= \lim_{n \to \infty} \mathbf{z}^{(n_k)} \\
&= \mathbf{z}
\end{aligned}
$$

and so $\mathbf{z} \in \mathcal{R}^{(k)}$. ∎

5. Examples

We apply the results of the previous section to examine the consistency of several multiclass methods. In all these examples, the functions $\Psi_y(\mathbf{f})$ are obtained from a single real valued function $\psi : \mathbb{R}_+^K \mapsto \mathbb{R}$ as follows

$$
\Psi_y(\mathbf{f}) = \psi(f_y, f_1, \ldots, f_{y-1}, f_{y+1}, \ldots, f_K).
$$

Moreover, the function ψ is symmetric in its last $K - 1$ arguments, that is, interchanging any two of the last $K - 1$ arguments does not change the value of the function. This ensures that the set $\mathcal{S}$ is symmetric. We assume that we predict the label of x to be $\arg\min_y \Psi_y(\mathbf{f})$.

5.1 Example 1

The method of Crammer and Singer (2001) corresponds to

$$
\Psi_y(\mathbf{f}) = \max_{y' \neq y} \phi(f_y - f_{y'}), \ \mathcal{C} = \mathbb{R}^K
$$

with $\phi(t) = (1-t)_+$. For $K = 3$, the boundary of $\mathcal{S}$ is shown in Figure 2(a). At the point $\mathbf{z} = (1,1,1)$, all of these are normals: $(0,1,1)$, $(1,0,1)$, $(1,1,0)$. Thus, there is no y' such that $p_{y'} = \max_y p_y$ for all $\mathbf{p} \in \mathcal{N}(\mathbf{z})$. The method is therefore inconsistent.

Even if we choose an everywhere differentiable convex ϕ with $\phi'(0) < 0$, the three normals mentioned above are still there in $\mathcal{N}(\mathbf{z})$ for $\mathbf{z} = (\phi(0), \phi(0), \phi(0))$. Therefore the method still remains inconsistent.

5.2 Example 2

The method of Weston and Watkins (1998) corresponds to

$$\Psi_y(\mathbf{f}) = \sum_{y' \neq y} \phi(f_y - f_{y'}), \ C = \mathbb{R}^K \tag{16}$$

with $\phi(t) = (1-t)_+$. For $K = 3$, the boundary of $\mathcal{S}$ is shown in Figure 2(b). The central hexagon has vertices (in clockwise order) $(1,1,4)$, $(0,3,3)$, $(1,4,1)$, $(3,3,0)$, $(4,1,1)$ and $(3,0,3)$. At $\mathbf{z} = (1,1,4)$, we have the following normals: $(1,1,0)$, $(1,1,1)$, $(2,3,1)$, $(3,2,1)$ and there is no coordinate which is maximum in all positive normals. The method is therefore inconsistent.

Now assume that ϕ is a positive convex classification calibrated loss function (i.e., $\phi'(0)$ exists and is negative). Note that if ϕ does not satisfy this assumption then we do not get a consistent method even in the binary classification case. Further assume that ϕ achieves its minimum. The following proposition then guarantees that $\Psi^{(k)}$ is boundedly invertible for $k > 1$ and we can ignore projections. In particular, if we choose ϕ differentiable so that $\mathcal{S}$ possesses a unique normal everywhere on its boundary then, by Proposition 9, $\mathcal{S}$ is admissible and hence classification calibrated. Such is the case if, for example, we choose $\phi(t) = ((1-t)_+)^2$.

Proposition 14 *Suppose $\phi : \mathbb{R} \mapsto [0, \infty)$ is a convex function with $\phi'(0) < 0$. Further, suppose that there is a t such that $\phi(t) = \inf_{t' \in \mathbb{R}} \phi(t')$. Then $\Psi_{(k)}$ (for Ψ given by (16)) is boundedly invertible.*

Proof Since ϕ is a positive convex function with $\phi'(0) < 0$ which achieves its minimum on the real line, only two behaviors are possible for ϕ. Either $\phi(t) \to \infty$ as $t \to \infty$ or there exists $t_0 > 0$ such that $\phi(t)$ is constant for $t \geq t_0$. In the first case, boundedness of $(\Psi_1(\mathbf{f}), \ldots, \Psi_k(\mathbf{f})), k > 1$ implies boundedness of all pairwise differences $f_y - f_{y'}$. Let $m^* = \min_y f_y$ and set $g_y = f_y - m^*$. The value of Ψ_y remains the same as it depends only on differences while all components of $\mathbf{g}$ are now bounded.

In the second case, when we only know that $\phi(t) \to \infty$ as $t \to -\infty$ and that $\phi(t)$ is constant for $t \geq t_0$, boundedness of $(\Psi_1(\mathbf{f}), \ldots, \Psi_k(\mathbf{f}))$ ($k > 1$), only implies that differences of the form $f_y - f_{y'}$ are bounded from below for $y \in \{1, \ldots, k\}$, $y' \in \{1, \ldots, K\}$. This implies that $f_y - f_{y'}$ is bounded for $y, y' \leq k$. Let $m^* = \min_{y \leq k} f_y$. Set $h_y = f_y - m^*$ (this doesn't change the value of Ψ as it depends only on differences). Now $h_y \geq 0$ for $y \leq k$ with (at least) one of them exactly zero. Since pairwise differences among these are bounded, the h_y's themselves are bounded for $y \leq k$. This implies $h_{k+1}, \ldots, h_K$ are bounded from above. Now set $g_y = h_y, y \leq k$ and $g_{y'} = \max\{-t_0, h_{y'}\}$ for $y' > k$. To see that we don't change anything by this update, note that when $h_{y'} < -t_0$, both $h_y - h_{y'}$ and $h_y + t_0$ are greater than t_0 (above which ϕ is constant in the case we're considering) for $y \leq k$. Thus, we have managed to make all g_y's bounded without changing the value of Ψ_y for all y. ∎
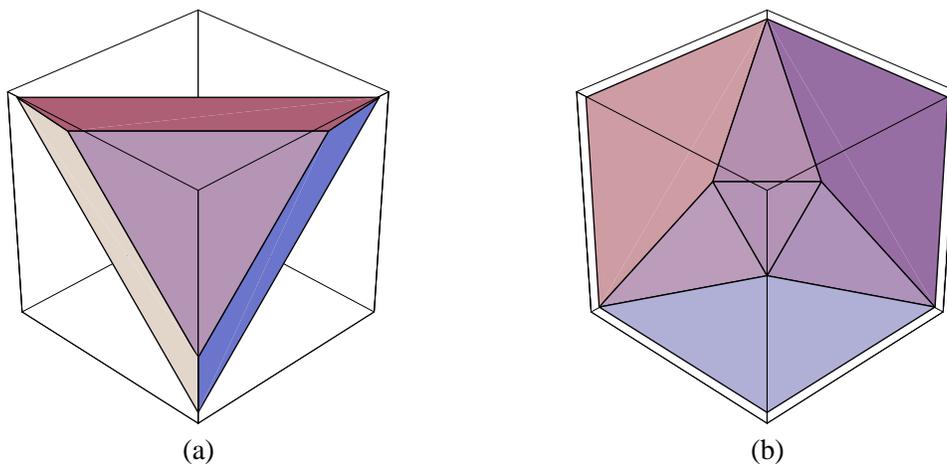
Figure 3: (a) Lee, Lin and Wahba (b) Loss of consistency in multiclass setting

5.3 Example 3

The method of Lee et al. (2004) corresponds to

$$\Psi_y(\mathbf{f}) = \sum_{y' \neq y} \phi(-f_{y'}), \; \mathcal{C} = \{\mathbf{f} : \sum_y f_y = 0\} \tag{17}$$

with $\phi(t) = (1-t)_+$. Figure 3(a) shows the boundary of $\mathcal{S}$ for $K = 3$. In the general K dimensional case, $\mathcal{S}$ is a polyhedron with K vertices where each vertex has a 0 in one of the positions and K's in the rest. It is obvious then when we minimize $\langle \mathbf{p}, \mathbf{z} \rangle$ over $\mathcal{S}$, we will pick the vertex which has a 0 in the same position where $\mathbf{p}$ has its maximum coordinate. But we can also apply our result here. The set of normals is not a singleton only at the vertices. Thus, by Proposition 10, we only need to check the vertices. Since there is a unique minimum coordinate at the vertices, Proposition 9 implies that the method is consistent.

The question which naturally arises is: for which convex loss functions ϕ does (17) lead to a consistent multiclass classification method? Convex loss functions which are classification calibrated for the two class case, that is, differentiable at 0 with $\phi'(0) < 0$, can lead to inconsistent classifiers of this kind in the multiclass setting. An example is provided by the loss function $\phi(t) = \max\{1 - 2t, 2 - t, 0\}$. Figure 3(b) shows the boundary of $\mathcal{S}$ for $K = 3$. The vertices are $(0,3,3)$, $(9,0,0)$ and their permutations. At $(9,0,0)$, the set of normals includes $(0,1,0)$, $(1,2,2)$ and $(0,0,1)$ and therefore condition (12) is violated.

A nice thing about this example is that if we assume ϕ is a positive classification calibrated binary loss function then $\Psi^{(k)}$ is boundedly invertible for $k > 1$. As in the previous example, this implies that differentiability of ϕ is then sufficient to guarantee consistency. In fact, as Zhang (2004b) shows, a convex function ϕ differentiable on $(-\infty, 0]$ with $\phi'(0) < 0$ will yield a consistent method.

Proposition 15 *Suppose $\phi : \mathbb{R} \mapsto [0, \infty)$ is a convex loss function with $\phi'(0) < 0$. Then $\Psi_{(k)}$ (for Ψ given by (17)) is boundedly invertible.*

Proof We have

$$\mathcal{R}^{(k)} = \left\{ \left(\sum_{y' \neq 1} \phi(-f_{y'}), \dots, \sum_{y' \neq k} \phi(-f_{y'}) \right) : \sum_y f_y = 0 \right\} .$$

We claim that $\phi(t) \to \infty$ as $t \to -\infty$. To see this, note that $\phi'(0) < 0$ so the linear function tangent to ϕ at 0 tends to ∞ as $t \to \infty$. Since ϕ is always above the tangent, the same it true for ϕ. This tells us that each f_y is bounded from above. Since $-f_y = \sum_{y' \neq y} f_{y'}$, $-f_y$ is also bounded from above. This means that each f_y is bounded and $\Psi^{(k)}$ is bounded invertible (just take $\mathbf{g} = \mathbf{f}$ in Definition 12). ∎

5.4 Example 4

This is an interesting example because even though we use a differentiable loss function, we still do not have consistency. Let

$$\Psi_y(\mathbf{f}) = \phi(f_y), \ \mathcal{C} = \{\mathbf{f} : \sum_y f_y = 0\}$$

with $\phi(t) = \exp(-\beta t)$ for some $\beta > 0$. One can easily check that

$$\mathcal{R} = \{(z_1, z_2, z_3)^T \in \mathbb{R}_+^3 : z_1 z_2 z_3 = 1\},$$

$$\mathcal{S} = \{(z_1, z_2, z_3)^T \in \mathbb{R}_+^3 : z_1 z_2 z_3 \geq 1\}$$

and so the projection $\mathcal{S}^{(2)}$ is the positive quadrant,

$$\mathcal{S}^{(2)} = \{(z_1, z_2)^T : z_1, z_2 > 0\} .$$

This set is inadmissible and therefore the method is inconsistent.

One can further show that changing ϕ is not of much help. Suppose, ϕ is a positive convex, classification calibrated binary loss function and $K \geq 3$. Then,

$$\mathcal{S}^{(2)} = \{(\phi(f_1), \phi(f_2)) : \sum_y f_y = 0\} .$$

Since $K \geq 3$, (f_1, f_2) is free to assume any value in $\mathbb{R}^2$ and hence $\mathcal{S}^{(2)} = T \times T$ where $T = \{\phi(t) : t \in \mathbb{R}\}$ is the range of ϕ. Let $m = \inf\{\phi(t) : t \in \mathbb{R}\}$. The set T is of the form (m, ∞) or $[m, \infty)$ depending on whether or not ϕ achieves its minimum. Clearly, $\mathcal{S}^{(2)}$ is a translation of the positive quadrant and is not admissible, (m, m) being a point violating the admissibility condition.

5.5 Summary of Examples

Table 1 summarizes how consistency of the four examples treated above depends on the choice of the underlying binary loss function ϕ. The last column gives a condition on ϕ which is sufficient to ensure consistency in case of a convex, positive, classification calibrated ϕ. Note that, for all the examples we considered, mere classification calibration and positivity of ϕ do not suffice to guarantee consistency of the derived multiclass method.

Example	Hinge $(1-t)_+$	Squared Hinge $((1-t)_+)^2$	Exponential $\exp(-t)$	Condition on ϕ
1	×	×	×	Never consistent
2	×	✓	✓[7]	ϕ differentiable & achieves its minimum
3	✓	✓	✓	ϕ differentiable
4	×	×	×	Never consistent

Table 1: Dependence of consistency on the underlying loss function ϕ.

6. Conclusion

We considered multiclass generalizations of classification methods based on convex risk minimization and gave a necessary and sufficient condition for their Bayes consistency. Some examples showed that quite often straightforward generalizations of consistent binary classification methods lead to inconsistent multiclass classifiers. This is especially the case if the original binary method was based on a non-differentiable loss function. Example 4 shows that even differentiable loss functions do not guarantee multiclass consistency. We also showed that in certain cases, one can avoid checking the full set of conditions mentioned in the theorem characterizing classification calibration. The question of consistency then reduces to checking properties of a single convex set. This was illustrated by considering variants of some multiclass methods proposed in the literature and proving their consistency.

Acknowledgments

Appendix A.

We will need the following lemma to prove Theorem 2.

Lemma 16 *The function* $\mathbf{p} \mapsto \inf_{\mathbf{z} \in S} \langle \mathbf{p}, \mathbf{z} \rangle$ *is continuous on* Δ_K.

Proof Let $\{\mathbf{p}^{(n)}\}$ be a sequence converging to $\mathbf{p}$. If B is a bounded subset of $\mathbb{R}^K$, then $\langle \mathbf{p}^{(n)}, \mathbf{z} \rangle \to \langle \mathbf{p}, \mathbf{z} \rangle$ uniformly over $\mathbf{z} \in B$ and therefore

$$\inf_{\mathbf{z} \in B} \langle \mathbf{p}^{(n)}, \mathbf{z} \rangle \to \inf_{\mathbf{z} \in B} \langle \mathbf{p}, \mathbf{z} \rangle \ .$$

Let B_r be a ball of radius r in $\mathbb{R}^K$. Then we have

$$\inf_{\mathbf{z} \in S} \langle \mathbf{p}^{(n)}, \mathbf{z} \rangle \leq \inf_{S \cap B_r} \langle \mathbf{p}^{(n)}, \mathbf{z} \rangle \to \inf_{S \cap B_r} \langle \mathbf{p}, \mathbf{z} \rangle.$$

Therefore

$$\limsup_{n} \inf_{\mathbf{z} \in S} \langle \mathbf{p}^{(n)}, \mathbf{z} \rangle \leq \inf_{\mathbf{z} \in S \cap B_r} \langle \mathbf{p}, \mathbf{z} \rangle \ .$$

---

7. This entry does not follow from the results presented in the paper but is included here for completeness.

Letting $r \to \infty$, we get

$$\limsup_n \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}^{(n)}, \mathbf{z} \rangle \le \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle \ . \tag{18}$$

Without loss of generality, assume that for some $j, 1 \le j \le K$ we have $p_1, \ldots, p_j > 0$ and $p_{j+1}, \ldots, p_K = 0$. For all sufficiently large integers n we can bound the components $p_i^{(n)}$ away from 0 for $i \in \{1, \ldots, j\}$. So, for a sufficiently large ball $B_M \subseteq \mathbb{R}^j$ we have

$$\inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle = \inf_{\mathbf{z} \in \mathcal{S}^{(j)}} \sum_{y=1}^{j} p_y z_y = \inf_{\mathbf{z} \in \mathcal{S}^{(j)} \cap B_M} \sum_{y=1}^{j} p_y z_y \ ,$$

$$\inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}^{(n)}, \mathbf{z} \rangle \ge \inf_{\mathbf{z} \in \mathcal{S}^{(j)}} \sum_{y=1}^{j} p_y^{(n)} z_y = \inf_{\mathbf{z} \in \mathcal{S}^{(j)} \cap B_M} \sum_{y=1}^{j} p_y^{(n)} z_y \ .$$

and thus

$$\liminf_n \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}^{(n)}, \mathbf{z} \rangle \ge \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle \ . \tag{19}$$

Combining (18) and (19), we get

$$\inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}^{(n)}, \mathbf{z} \rangle \to \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle \ .$$

∎

Theorem 2. *Let Ψ be a (vector-valued) loss function and $\mathcal{C}$ be a subset of $\mathbb{R}^K$. Let $\mathcal{F}$ and $\mathcal{S}$ be as defined in (5) and (6) respectively. Then $\mathcal{S}$ is classification calibrated iff the following holds. Whenever $\{\mathcal{F}_n\}$ is a sequence of function classes (where $\mathcal{F}_n \subseteq \mathcal{F}$ and $\cup \mathcal{F}_n = \mathcal{F}$) such that $\hat{\mathbf{f}}_n \in \mathcal{F}_n$ and P is the data generating probability distribution,*

$$R_\Psi(\hat{\mathbf{f}}_n) \xrightarrow{P} R_\Psi^*$$

implies

$$R(\hat{\mathbf{f}}_n) \xrightarrow{P} R^* \ .$$

Proof ('only if') Suppose we could prove that $\forall \varepsilon > 0, \exists \delta > 0$ such that $\forall \mathbf{p} \in \Delta_K$,

$$\max_y p_y - p_{\text{pred}(\mathbf{z})} \ge \varepsilon \Rightarrow \langle \mathbf{p}, \mathbf{z} \rangle - \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle \ge \delta \ . \tag{20}$$

Using this it immediately follows that $\forall \varepsilon, H(\varepsilon) > 0$ where

$$H(\varepsilon) = \inf_{\mathbf{p} \in \Delta_K, \mathbf{z} \in \mathcal{S}} \left\{ \langle \mathbf{p}, \mathbf{z} \rangle - \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle \ : \ \max_y p_y - p_{\text{pred}(\mathbf{z})} \ge \varepsilon \right\} \ .$$

A result of Zhang (2004b, Corollary 26) then guarantees there exists a concave function ξ on $[0, \infty)$ such that $\xi(0) = 0$ and $\xi(\delta) \to 0$ as $\delta \to 0^+$ and

$$R(\mathbf{f}) - R^* \le \xi(R_\Psi(\mathbf{f}) - R_\Psi^*) \ .$$

This inequality combined with $R_\Psi(\hat{\mathbf{f}}_n) \xrightarrow{P} R_\Psi^*$ easily gives $R(\hat{\mathbf{f}}_n) \xrightarrow{P} R^*$. We now prove the implication in (20) by contradiction. Suppose $\mathcal{S}$ is classification calibrated but there exists $\varepsilon > 0$ and a sequence $(\mathbf{z}^{(n)}, \mathbf{p}^{(n)})$ such that

$$p_{\mathrm{pred}(\mathbf{z}^{(n)})}^{(n)} \leq \max_y p_y^{(n)} - \varepsilon \tag{21}$$

and

$$\left(\langle \mathbf{p}^{(n)}, \mathbf{z}^{(n)} \rangle - \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}^{(n)}, \mathbf{z} \rangle \right) \to 0 \ .$$

Since $\mathbf{p}^{(n)}$ come from a compact set, we can choose a convergent subsequence (which we still denote as $\{\mathbf{p}^{(n)}\}$) with limit $\mathbf{p}$. Using Lemma 16, we get

$$\langle \mathbf{p}^{(n)}, \mathbf{z}^{(n)} \rangle \to \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle \ .$$

As before, we assume that precisely the first j coordinates of $\mathbf{p}$ are non-zero. Then the first j coordinates of $\mathbf{z}^{(n)}$ are bounded for sufficiently large n. Hence

$$\limsup_n \langle \mathbf{p}, \mathbf{z}^{(n)} \rangle = \limsup_n \sum_{y=1}^{j} p_y^{(n)} z_y^{(n)} \leq \lim_{n \to \infty} \langle \mathbf{p}^{(n)}, \mathbf{z}^{(n)} \rangle = \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle \ .$$

Now (10) and (21) contradict each other since $\mathbf{p}^{(n)} \to \mathbf{p}$.

('if') If $\mathcal{S}$ is not classification calibrated then by Theorem 7 and Propositions 9 and 10, we have a point in the boundary of some $\mathcal{S}^{(i)}$ where there are at least two normals and which does not have a unique minimum coordinate. Such a point should be there in the projection of $\mathcal{R}$ even without taking the convex hull. Therefore, we must have a sequence $\mathbf{z}^{(n)}$ in $\mathcal{R}$ such that

$$\delta_n = \langle \mathbf{p}, \mathbf{z}^{(n)} \rangle - \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle \to 0 \tag{22}$$

and for all n,

$$p_{\mathrm{pred}(\mathbf{z}^{(n)})} < \max_y p_y \ . \tag{23}$$

Without loss of generality assume that δ_n is a monotonically decreasing sequence. Further, assume that $\delta_n > 0$ for all n. This last assumption might be violated but the following proof then goes through for δ_n replaced by $\max(\delta_n, 1/n)$. Let $\mathbf{g}_n$ be the function that maps every x to one of the pre-images of $\mathbf{z}^{(n)}$ under Ψ. Define $\mathcal{F}_n$ as

$$\mathcal{F}_n = \{\mathbf{g}_n\} \cup \left(\mathcal{F} \cap \{\mathbf{f} : \forall x, \langle \mathbf{p}, \Psi(\mathbf{f}(x)) \rangle - \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle > 4\delta_n \} \right.$$

$$\left. \cap \{\mathbf{f} : \forall x, \forall j, |\Psi_j(\mathbf{f}(x))| < M_n \} \right)$$

where $M_n \uparrow \infty$ is a sequence which we will fix later. Fix a probability distribution P with arbitrary marginal distribution over x and let the conditional distribution of labels be $\mathbf{p}$ for all x. Our choice of $\mathcal{F}_n$ guarantees that the Ψ-risk of $\mathbf{g}_n$ is less than that of other elements of $\mathcal{F}_n$ by at least $3\delta_n$. Suppose, we make sure that

$$P^n \left(|\hat{R}_\Psi(\mathbf{g}_n) - R_\Psi(\mathbf{g}_n)| > \delta_n \right) \to 0 \ , \tag{24}$$

$$P^n \left(\sup_{\mathbf{f} \in \mathcal{F}_n - \{\mathbf{g}_n\}} |\hat{R}_\Psi(\mathbf{f}) - R_\Psi(\mathbf{f})| > \delta_n \right) \to 0 \ . \tag{25}$$

Then, with probability tending to 1, $\hat{\mathbf{f}}_n = \mathbf{g}_n$. By (22), $R_\Psi(\mathbf{g}_n) \to R_\Psi^*$ which implies that $R_\Psi(\hat{\mathbf{f}}_n) \to R_\Psi^*$ in probability. Similarly, (23) implies that $R(\hat{\mathbf{f}}_n) \nrightarrow R^*$ in probability.

We only need to show that we can have (24) and (25) hold. For (24), we apply Chebyshev inequality and use a union bound over the K labels to get

$$P^n\left(\left|\hat{R}_\Psi(\mathbf{g}_n) - R_\Psi(\mathbf{g}_n)\right| > \delta_n\right) \leq \frac{K^3 \|\mathbf{z}^{(n)}\|_\infty}{4n\delta_n^2}$$

The right hand side can be made to go to zero by repeating terms in the sequence $\{\mathbf{z}^{(n)}\}$ to slow down the rate of growth of $\|\mathbf{z}^{(n)}\|_\infty$ and the rate of decrease of δ_n. For (24), we use standard covering number bounds (e.g., see Pollard, 1984, Section II.6).

$$P^n\left(\sup_{\mathbf{f} \in \mathcal{F}_n - \{\mathbf{g}_n\}} \left|\hat{R}_\Psi(\mathbf{f}) - R_\Psi(\mathbf{f})\right| > \delta_n\right) \leq 8\exp\left(\frac{64M_n^2 \log(2n+1)}{\delta_n^2} - \frac{n\delta_n^2}{128M_n^2}\right)$$

Thus M_n/δ_n needs to grow slowly enough such that

$$\frac{n\delta_n^4}{M_n^4 \log(2n+1)} \to \infty \; .$$

∎

References

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Large margin classifiers: Convex loss, low noise and convergence rates. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

Erin J. Bredensteiner and Kristin P. Bennett. Multicategory classification by support vector machines. *Computational Optimization and Applications*, 12:35–46, 1999.

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

Wenxin Jiang. Process consistency for AdaBoost. *Annals of Statistics*, 32(1):13–29, 2004.

Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

Gábor Lugosi and Nicolas Vayatis. On the Bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, 32(1):30–55, 2004.

David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

Ingo Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005.

Jason Weston and Chris Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway College, University of London, 1998.

Tong Zhang. An infinity-sample theory for multi-category large margin classification. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2004a.

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004b.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–85, 2004c.