On the Consistency of Multiclass Classification Methods

Ambuj Tewari[1] and Peter L. Bartlett[2]

[1] Division of Computer Science
University of California, Berkeley
ambuj@cs.berkeley.edu
[2] Division of Computer Science and Department of Statistics
University of California, Berkeley
bartlett@cs.berkeley.edu

Abstract. Binary classification methods can be generalized in many ways to handle multiple classes. It turns out that not all generalizations preserve the nice property of Bayes consistency. We provide a necessary and sufficient condition for consistency which applies to a large class of multiclass classification methods. The approach is illustrated by applying it to some multiclass methods proposed in the literature.

1 Introduction

We consider the problem of classification in a probabilistic setting: n i.i.d. pairs are generated by a probability distribution on $\mathcal{X} \times \mathcal{Y}$. 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 is 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 the so-called *large margin methods* which use $\{+1, -1\}$ as the set of labels and $\phi(yf(x))$ as the loss. Bayes consistency of these methods has been analyzed in the literature (see [1, 4, 6, 9, 13]). 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 found in [2, 5, 10, 11]. Zhang [11, 12] 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" $R_\Psi(\mathbf{f}) = \mathbb{E}_{\mathcal{X}\mathcal{Y}}[\Psi_y(\mathbf{f}(x))]$. 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. Towards this end, a notion of "classification calibration" was defined in [1] for binary classification. The authors also gave a simple characterization of classification calibration for convex loss functions. In Section 2, we provide a different point of view for looking at classification calibration and motivate the geometric approach of Section 3.

Section 3 deals with multiclass classification and defines an analog of classification calibration (Definition 1). A necessary and sufficient condition for classification calibration is provided (Theorem 8). It is not as simple and easy to verify as in the binary case. This helps us realize that the study of multiclass classification is not a simple generalization of results known for the binary case but is much more subtle and involved. Finally, the equivalence of classification calibration and consistency of methods based on empirical Ψ-risk minimization is established (Theorem 10).

In Section 4, we consider a few multiclass methods and apply the result of Section 3 to examine their consistency. Interestingly, many seemingly natural generalizations of binary methods do not lead to consistent multiclass methods. We discuss further work and conclude in Section 5.

2 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 [1] that any minimizer f^* of

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

yields a Bayes consistent classifier, i.e. $P(Y = +1|X = x) > 1/2 \Rightarrow f^*(x) > 0$ 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

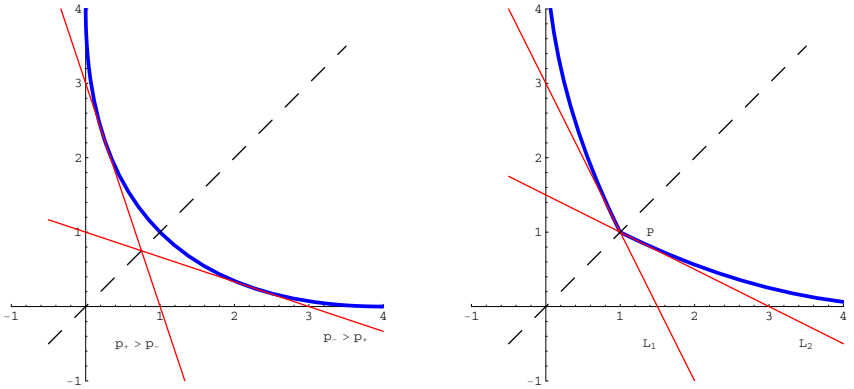
$$\mathcal{R} = \{(\phi(f), \phi(-f)) \ : \ f \in \mathbb{R}\} \ , \tag{3}$$

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 Fig. 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



(a) (b)

Fig. 1. (a) Squared Hinge Loss (b) Inconsistent Case (the thick curve is the set $\mathcal{R}$ in both plots)

equation is $\langle \mathbf{p}, \mathbf{z} \rangle = c$ and then 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 really necessary. 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? Fig. 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. Thus 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 next section, we deal with cases when the set $\mathcal{R}$ is generated in a more general way and the situation possibly involves more than two dimensions.

3 Consistency of Multiclass Classification Methods

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 we obtain a classifier $\hat{\mathbf{f}}_n$ by minimizing the empirical Ψ-risk $\hat{R}_\Psi$ over the class $\mathcal{F}_n$,

$$\hat{\mathbf{f}}_n = \operatorname*{arg\,min}_{\mathbf{f} \in \mathcal{F}_n} \hat{R}_\Psi(\mathbf{f}) = \operatorname*{arg\,min}_{\mathbf{f} \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} \Psi_{y_i}(\mathbf{f}(x_i)) \ . \tag{5}$$

There might be some constraints on the set of vector functions over which we minimize. 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 $\mathcal{C} \in \mathbb{R}^K$ such that

$$\mathcal{F} = \{\mathbf{f} : \forall x, \ \mathbf{f}(x) \in \mathcal{C} \ \}. \tag{6}$$

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 $p_y(x) = P(Y = y \mid X = x)$. 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). 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 of the conditional expectation above[3]

$$\inf_{\mathbf{f} \in \mathcal{C}} \sum_y p_y \Psi_y(\mathbf{f}) \ . \tag{7}$$

[3] 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.

Define the subset $\mathcal{R}$ of $\mathbb{R}_+^K$ as

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

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. We can write (7) in the equivalent form

$$\inf_{\mathbf{z} \in \mathcal{R}} \ \langle \mathbf{p}, \mathbf{z} \rangle \ ,$$

where $\mathbf{p} = (p_1, \ldots, p_K)$. 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[4] of $\mathcal{R}$. Defining $\mathcal{S}$ as

$$\mathcal{S} = \text{conv}\{(\Psi_1(\mathbf{f}), \ldots, \Psi_K(\mathbf{f})) \ : \ \mathbf{f} \in \mathcal{C}\} \ , \tag{8}$$

we finally have

$$\inf_{\mathbf{z} \in \mathcal{S}} \ \langle \mathbf{p}, \mathbf{z} \rangle \ . \tag{9}$$

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* $\text{pred} : \mathbb{R}^K \mapsto \{1, \ldots, K\}$ *such that*

$$\forall \mathbf{p} \in \Delta_K, \ \inf_{\mathbf{z} \in S \, : \, p_{\text{pred}(\mathbf{z})} < \max_y p_y} \langle \mathbf{p}, \mathbf{z} \rangle > \inf_{\mathbf{z} \in \mathcal{S}} \ \langle \mathbf{p}, \mathbf{z} \rangle \ , \tag{10}$$

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

It is easy to reformulate the definition in terms of sequences as the following lemma states.

Lemma 2. *$\mathcal{S} \subseteq \mathbb{R}_+^K$ is classification calibrated iff $\forall \mathbf{p} \in \Delta_K$ and all sequences $\{\mathbf{z}^{(n)}\}$ in $\mathcal{S}$ such that*

$$\langle \mathbf{p}, \mathbf{z}^{(n)} \rangle \to \inf_{\mathbf{z} \in \mathcal{S}} \ \langle \mathbf{p}, \mathbf{z} \rangle \ , \tag{11}$$

we have

$$p_{\text{pred}(\mathbf{z}^{(n)})} = \max_y p_y \tag{12}$$

ultimately.

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 (9) 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).

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

Lemma 3. *If there exists a predictor function* pred *satisfying the condition* (10) *of Definition 1 then any predictor function* pred$'$ *satisfying*

$$\forall \mathbf{z} \in \mathcal{S},\ z_{\mathrm{pred}'(\mathbf{z})} = \min_y z_y \tag{13}$$

also satisfies (10)*.*

Proof. Consider some $\mathbf{p} \in \Delta_K$ and a sequence $\{\mathbf{z}^{(n)}\}$ such that (11) 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,

$$\begin{aligned}
\mathrm{pred}(\mathbf{z}^{(n_k)}) &= M \in \{y' : y' = \max_y p_y\}\ , \\
\mathrm{pred}'(\mathbf{z}^{(n_k)}) &= m \in \{y' : y' < \max_y p_y\}\ , \\
\langle \mathbf{p}, \mathbf{z}^{(n_k)} \rangle &\to \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle\ .
\end{aligned}$$

Because of (13), 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, denote its value by $\epsilon > 0$. Then $z_M^{(n_k)} - z_m^{(n_k)} > \epsilon/2$ ultimately and hence we have

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

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

Otherwise, choose a subsequence[5] $\{\mathbf{z}^{(n_k)}\}$ such that $\lim(z_M^{(n_k)} - z_m^{(n_k)}) = 0$. Multiplying this with $(p_M - p_m)$, we have

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

We also have

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

where the last equality follows because of symmetry. This means

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

[5] We do not introduce additional subscripts for simplicity.

and therefore

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

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

From now on, we assume that pred is defined as in (13). We give another 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 $\langle \mathbf{z}' - \mathbf{z}, \mathbf{p} \rangle \geq 0$[6] 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 4. *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*

$$\operatorname{argmin}(\mathbf{z}) \subseteq \operatorname{argmax}(\mathbf{p}) \tag{14}$$

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

The following lemma states that in the presence of symmetry points having a unique minimum coordinate can never destroy admissibility.

Lemma 5. *Let $\mathcal{S} \subseteq \mathbb{R}^K_+$ be a symmetric convex set, $\mathbf{z}$ a point in the boundary of $\mathcal{S}$ and $\mathbf{p} \in \mathcal{N}(\mathbf{z})$. Then $z_y < z_{y'}$ implies $p_y \geq p_{y'}$ and hence* (14) *holds whenever* $|\operatorname{argmin}(\mathbf{z})| = 1$.

Proof. Consider $\tilde{\mathbf{z}}$ obtained from $\mathbf{z}$ by interchanging the y, y' coordinates. It also is a point in $\partial\mathcal{S}$ by symmetry and thus convexity implies $\mathbf{z}_m = (\mathbf{z}+\tilde{\mathbf{z}})/2 \in \mathcal{S} \cup \partial\mathcal{S}$. Since $\mathbf{p} \in \mathcal{N}(\mathbf{z})$, $\langle \mathbf{z}' - \mathbf{z}, \mathbf{p} \rangle \geq 0$ for all $\mathbf{z}' \in \mathcal{S}$. Taking limits, this inequality also holds for $\mathbf{z}' \in \mathcal{S} \cup \partial\mathcal{S}$. Substituting $\mathbf{z}_m$ for $\mathbf{z}'$, we get $\langle (\tilde{\mathbf{z}} - \mathbf{z})/2, \mathbf{p} \rangle \geq 0$ which simplifies to $(z_{y'} - z_y)(p_y - p_{y'}) \geq 0$ whence the conclusion follows.

If the set $\mathcal{S}$ possesses a unique normal at every point on its boundary then the next lemma guarantees admissibility.

Lemma 6. *Let $\mathcal{S} \subseteq \mathbb{R}^K_+$ be a symmetric convex set, $\mathbf{z}$ a point in the boundary of $\mathcal{S}$ and $\mathcal{N}(\mathbf{z}) = \{\mathbf{p}\}$ is a singleton. Then* $\operatorname{argmin}(\mathbf{z}) \subseteq \operatorname{argmax}(\mathbf{p})$.

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 $y' \in \operatorname{argmax}(\mathbf{p})$. From the proof of Lemma 5 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 $\mathcal{S}$ 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 $|\mathcal{N}(\mathbf{z})| \geq 2$.

[6] Our sign convention is opposite to that of Rockafellar (1970) because we are dealing with minimum (instead of maximum) problems.

Lemma 7. *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 $p_{\text{pred}(\mathbf{z}^{(n)})} = \max_y p_y$ 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. We claim that for all $\epsilon > 0$, $\text{dist}(\mathbf{z}^{(n)}, Z(\mathbf{p})) < \epsilon$ ultimately. For if we assume the contrary, boundedness implies that we can find a convergent subsequence $\{\mathbf{z}^{(n_k)}\}$ such that $\forall k$, $\text{dist}(\mathbf{z}^{(n_k)}, Z(\mathbf{p})) \geq \epsilon$. 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, $\text{dist}(\mathbf{z}^*, Z(\mathbf{p})) \geq \epsilon$ which gives us a contradiction and our claim is proved. Further, there exists $\epsilon' > 0$ such that $\text{dist}(\mathbf{z}^{(n)}, Z(\mathbf{p})) < \epsilon'$ implies $\text{argmin}(\mathbf{z}^{(n)}) \subseteq \text{argmin}(Z(\mathbf{p}))$[7]. 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 8. *Let $\mathcal{S} \subseteq \mathbb{R}^K_+$ be a symmetric convex set. Define the sets*

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

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 sequences 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 $j, 1 \leq j \leq K$ 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 $\forall y \leq j, z_y^{(n)} \leq M$ ultimately. Since $p_j z_j^{(n)} \leq L + 1$ ultimately, $M = \max_{1\leq y\leq j}\{(L+1)/p_y\}$ works. Consider a set of labels $T \subseteq \{j+1, \ldots, K\}$. Consider the subsequence consisting of those $\mathbf{z}^{(n)}$ for which $z_y \leq M$ for $y \in \{1, \ldots, j\} \cup T$ and $z_y > M$ 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

[7] For a set Z, $\text{argmin}(Z)$ denotes $\cup_{\mathbf{z}\in Z} \text{argmin}(\mathbf{z})$.

the set $\{j+1,\ldots,K\}$. 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 7, we have $\tilde{p}_{\mathrm{pred}(\tilde{\mathbf{z}}^{(n)})} = \max_y \tilde{p}_y = \max_y p_y$ ultimately. Since we dropped only those coordinates which were greater than M, $\mathrm{pred}(\tilde{\mathbf{z}}^{(n)})$ picks the same coordinate as $\mathrm{pred}(\mathbf{z}^{(n)})$ where $\mathbf{z}^{(n)}$ is the element from which $\tilde{\mathbf{z}}^{(n)}$ was obtained. Thus we have $p_{\mathrm{pred}(\mathbf{z}^{(n)})} = \max_y p_y$ ultimately and the theorem is proved.

We will need the following lemma to prove our final theorem.

Lemma 9. *The function* $\mathbf{p} \mapsto \inf_{\mathbf{z}\in\mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle$ *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

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

Therefore

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

Letting $r \to \infty$, we get

$$\limsup_n \inf_{\mathbf{z}\in\mathcal{S}} \langle \mathbf{p}^{(n)}, \mathbf{z} \rangle \le \inf_{\mathbf{z}\in\mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle \ . \tag{15}$$

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 and 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

$$\liminf_n \inf_{\mathbf{z}\in\mathcal{S}} \langle \mathbf{p}^{(n)}, \mathbf{z} \rangle \ge \inf_{\mathbf{z}\in\mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle \ . \tag{16}$$

Combining (15) and (16), we get

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

We finally show that classification calibration of $\mathcal{S}$ is equivalent to the consistency of multiclass methods based on (5).

Theorem 10. *Let Ψ be a loss (vector) function and $\mathcal{C}$ be a subset of $\mathbb{R}^K$. Let $\mathcal{F}$ and $\mathcal{S}$ be as defined in* (6) *and* (8) *respectively. Then $\mathcal{S}$ is classification calibrated iff the following holds. For all sequences $\{\mathcal{F}_n\}$ of function classes (where $\mathcal{F}_n \subseteq \mathcal{F}$ and $\cup \mathcal{F}_n = \mathcal{F}$) and for all probability distributions P,*

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

implies

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

Proof. ('only if') We need to prove that $\forall \epsilon > 0, \exists \delta > 0$ such that $\forall \mathbf{p} \in \Delta_K$,

$$\max_y p_y - p_{\mathrm{pred}(\mathbf{z})} \geq \epsilon \Rightarrow \langle \mathbf{p}, \mathbf{z} \rangle - \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle \geq \delta \ . \tag{17}$$

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

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

Corollary 26 in [12] 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

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

We prove (17) by contradiction. Suppose $\mathcal{S}$ is classification calibrated but there exists $\epsilon > 0$ and a sequence $(\mathbf{z}^{(n)}, \mathbf{p}^{(n)})$ such that

$$p^{(n)}_{\mathrm{pred}(\mathbf{z}^{(n)})} \leq \max_y p^{(n)}_y - \epsilon \tag{18}$$

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 9, 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

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

Now (12) and (18) contradict each other since $\mathbf{p}^{(n)} \to \mathbf{p}$.

('if') If $\mathcal{S}$ is not classification calibrated then by Theorem 8 and Lemmas 5 and 6, 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

$$\delta_n = \langle \mathbf{p}, \mathbf{z}^{(n)} \rangle - \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle \to 0 \tag{19}$$

and for all n,

$$p_{\text{pred}(\mathbf{z}^{(n)})} < \max_y p_y \ . \tag{20}$$

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

$$\mathcal{F}_n = \{\mathbf{g}_n\} \ \cup \ (\mathcal{F} \cap \{\mathbf{f} : \forall x, \langle \mathbf{p}, \Psi(\mathbf{f}(x)) \rangle - \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle > 4\delta_n\} \\ \cap \{\mathbf{f} : \forall x, \forall j, |\Psi_j(\mathbf{f}(x)| < M_n\})$$

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

$$P^n \left(\left| \hat{R}_\Psi(\mathbf{g}_n) - R_\Psi(\mathbf{g}_n) \right| > \delta_n \right) \to 0 \ , \tag{21}$$

$$P^n \left(\sup_{\mathbf{f} \in \mathcal{F}_n - \{\mathbf{g}_n\}} \left| \hat{R}_\Psi(\mathbf{f}) - R_\Psi(\mathbf{f}) \right| > \delta_n \right) \to 0 \ . \tag{22}$$

Then, with probability tending to 1, $\hat{\mathbf{f}}_n = \mathbf{g}_n$. By (19), $R_\Psi(\mathbf{g}_n) \to R^*_\Psi$ which implies that $R_\Psi(\hat{\mathbf{f}}_n) \to R^*_\Psi$ in probability. Similarly, (20) implies that $R(\hat{\mathbf{f}}_n) \nrightarrow R^*$ in probability.

We only need to show that we can have (21) and (22) hold. For (21), 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 (21), we use standard covering number bounds (see, for example, Section II.6 on p. 30 in [7]).

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

Thus M_n/δ_n needs to grow slowly enough such that

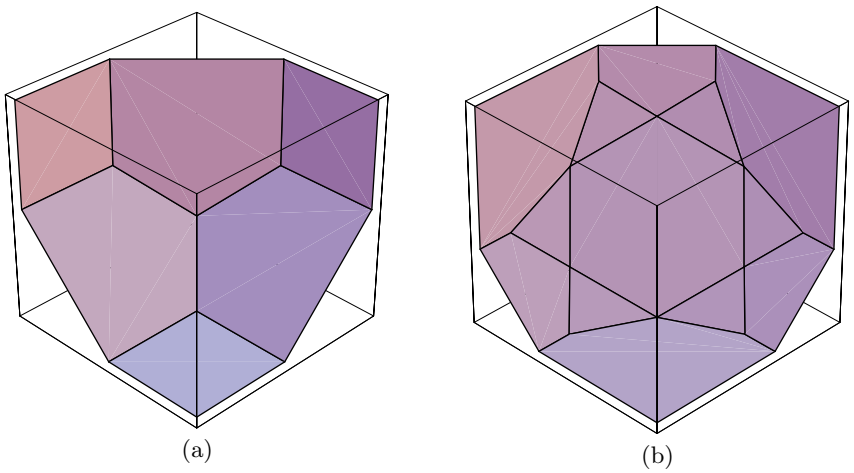
$$\frac{n\delta_n^4}{M_n^4 \log(2n+1)} \to \infty \ .$$

4 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, i.e. 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 $\arg\min_y \Psi_y(\mathbf{f})$.



(a) (b)

Fig. 2. (a) Crammer and Singer (b) Weston and Watkins

4.1 Example 1

The method of Crammer and Singer [3] 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 Fig. 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.

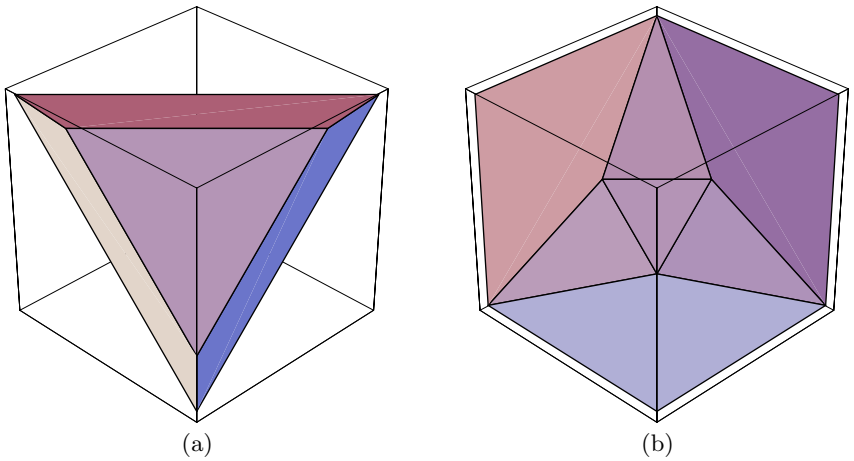
Even if we choose an everywhere differentiable convex ϕ with $\phi'(0) < 0$, the three normals mentioned above are still there in $\mathcal{N}(\mathbf{z})$ for $\mathbf{z} = (\phi(0), \phi(0), \phi(0))$. Therefore the method still remains inconsistent.

4.2 Example 2

The method of Weston and Watkins [10] corresponds to

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

with $\phi(t) = (1-t)_+$. For $K = 3$, the boundary of $\mathcal{S}$ is shown in Fig. 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.



(a) (b)

Fig. 3. (a) Lee, Lin and Wahba (b) Loss of consistency in multiclass setting

4.3 Example 3

The method of Lee, Lin and Wahba [5] corresponds to

$$\Psi_y(\mathbf{f}) = \sum_{y' \neq y} \phi(-f_{y'}),\ \mathcal{C} = \{\mathbf{f} : \sum_y f_y = 0\} \tag{23}$$

with $\phi(t) = (1-t)_+$. Fig. 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 Lemma 6, we only need to check the vertices. Since there is a unique minimum coordinate at the vertices, Lemma 5 implies that the method is consistent.

The question which naturally arises is: for which convex loss functions ϕ does (23) lead to a consistent multiclass classification method? Convex loss functions which are classification calibrated for the two class case, i.e. differentiable at 0 with $\phi'(0) < 0$, can lead to inconsistent classifiers in the multiclass setting. An example is provided by the loss function $\phi(t) = \max\{1 - 2t, 2 - t, 0\}$. Fig. 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 (14) is violated.

As Zhang shows in [12], a convex function ϕ differentiable on $(-\infty, 0]$ with $\phi'(0) < 0$ will yield a consistent method.

4.4 Example 4

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

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

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

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

This set is inadmissible and therefore the method is inconsistent. We point out that this method also does not yield a consistent classifier for the choice $\phi(t) = (1 - t)_+$.

5 Conclusion

We considered multiclass generalizations of classification methods based on convex risk minimization and gave a necessary and sufficient condition for their Bayes consistency. Our 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 are currently trying to find simple and sufficient differentiability conditions that would imply consistency of methods discussed in Examples 2 and 4 (like the one Zhang provides for Example 3).

Acknowledgement

We gratefully acknowledge the support of NSF under award DMS-0434383 and of ARO under MURI grant DAAD 190210383.

References

1. Bartlett, P. L., Jordan, M. I. and McAuliffe, J. D.: Large margin classifiers: Convex Loss, Low Noise and Convergence rates. In *Advances in Neural Information Processing Systems* **16** (2004)
2. Bredensteiner, E.J. and Bennett, K.P.: Multicategory Classification by Support Vector Machines. *Computational Optimization and Applications* **12** (1999) 35–46
3. Crammer, K. and Singer, Y.: On the Algorithmic Implementation of Kernel-based Vector Machines. *Journal of Machine Learning Research* **2** (2001) 265–292
4. Jiang, W.: Process Consistency for AdaBoost. *Annals of Statistics* **32**:1 (2004) 13–29
5. Lee, Y., Li, Y. and Wahba, G.: 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 (2004) 67–81
6. Lugosi, G. and Vayatis, N.: On the Bayes-risk Consistency of Regularized Boosting Methods. *Annals of Statistics* **32**:1 (2004) 30–55
7. Pollard, D.: *Convergence of Stochastic Processes*. Springer-Verlag, New York (1984)
8. Rockafellar, R.T.: *Convex Analysis*. Princeton University Press, Princeton (1970)
9. Steinwart, I.: Consistency of Support Vector Machines and Other Regularized Kernel Classifiers. *IEEE Transactions on Information Theory* **51**:1 (2005) 128–142
10. Weston, J. and and Watkins, C.: Multi-class support vector machines. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway College, University of London (1998)
11. Zhang, T.: An Infinity-sample Theory For Multi-category Large Margin Classification. In *Advances in Neural Information Processing Systems* **16** (2004)
12. Zhang, T.: Statistical Analysis of Some Multi-Category Large Margin Classification Methods. *Journal of Machine Learning Research* **5** (2004) 1225–1251
13. Zhang, T.: Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization. *Annals of Statistics* **32**:1 (2004) 56–85