Fisher Consistency of Multicategory Support Vector Machines

Yufeng Liu
Department of Statistics and Operations Research
Carolina Center for Genome Sciences
University of North Carolina
Chapel Hill, NC 27599-3260
yfliu@email.unc.edu

Abstract

The Support Vector Machine (SVM) has become one of the most popular machine learning techniques in recent years. The success of the SVM is mostly due to its elegant margin concept and theory in binary classification. Generalization to the multicategory setting, however, is not trivial. There are a number of different multicategory extensions of the SVM in the literature. In this paper, we review several commonly used extensions and Fisher consistency of these extensions. For inconsistent extensions, we propose two approaches to make them Fisher consistent, one is to add bounded constraints and the other is to truncate unbounded hinge losses.

1 Background on Binary SVM

The Support Vector Machine (SVM) is a well known large margin classifier and has achieved great success in many applications (Vapnik, 1998, Cristianini and Shawe-Taylor, 2000, and Hastie, Tibshirani, and Friedman, 2000). The basic concept behind the binary SVM is to search a separating hyperplane with maximum separation between the two classes.

Suppose a training dataset containing n training pairs $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, i.i.d realizations from a probability distribution $P(\boldsymbol{x}, y)$, is given. The goal is to search for a linear function $f(\boldsymbol{x}) = \boldsymbol{w}^{'}\boldsymbol{x} + b$ so that $\text{sign}(f(\boldsymbol{x}))$ can be used for prediction of labels for new inputs. The SVM aims to find such an f so that points of class $+1$ and points of class -1 are best separated. In particular, for the separable case, the SVM's solution maximizes the distance between $f(\boldsymbol{x}) = \pm 1$ subject to $y_i f(\boldsymbol{x}_i) \geq 1$; $i = 1, \ldots, n$. This distance can be expressed as $\frac{2}{\|\boldsymbol{w}\|}$ and is known as the geometric margin.

When perfect separation between two classes is not feasible, slack variables ξ_i; $i = 1, \ldots, n$, can be used to measure the amount of violation of the original constraints. Then the SVM solves the following optimization problem:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^n \xi_i \qquad (1)$$
$$\text{s.t.} \qquad y_i f(\boldsymbol{x}_i) \geq 1 - \xi_i, \xi_i \geq 0, \forall i,$$

where $C > 0$ is a tuning parameter which balances the separation and the amount of violation of the constraints.

Optimization formulation in (1) is also known as the primal problem of the SVM. Using the Lagrange multipliers, (1) can be converted into an equivalent dual problem as follows:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{i,j=1}^n y_i y_j \alpha_i \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle - \sum_{i=1}^n \alpha_i \qquad (2)$$
$$\text{s.t.} \qquad \sum_{i=1}^n y_i \alpha_i = 0; 0 \leq \alpha_i \leq C, \forall i.$$

Once the solution of problem (2) is obtained, $\boldsymbol{w}$ can be calculated as $\sum_{i=1}^n y_i \alpha_i \boldsymbol{x}_i$ and the intercept b can be computed using the Karush-Kuhn-Tucker (KKT) complementary conditions of the optimization theory. If nonlinear learning is needed, one can apply the *kernel trick* by replacing the inner product $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$ by $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where the kernel K is a positive definite function. This amounts to applying linear learning in the feature space induced by the kernel K to achieve nonlinear learning in the original input space.

It is now known that the SVM can be fit in the regularization framework (Wahba, 1998) as follows:

$$\min_f J(f) + C\sum_{i=1}^n [1 - y_i f(\boldsymbol{x}_i)]_+, \qquad (3)$$

where the function $[1-u]_+ = 1-u$ if $u \leq 1$ and 0 otherwise and it is known as the hinge loss function (see Figure 1). The term $J(f)$ is a regularization term and in the linear learning setting, it becomes $\frac{1}{2}\|\boldsymbol{w}\|_2^2$ which is related to the geometric margin.

Denote $P(\boldsymbol{x}) = P(Y = 1 | \boldsymbol{X} = \boldsymbol{x})$. Then a binary classifier with loss $V(f(\boldsymbol{x}), y)$ is Fisher consistent if the minimizer of $E[V(f(\boldsymbol{X}), Y | \boldsymbol{X} = \boldsymbol{x}]$ has

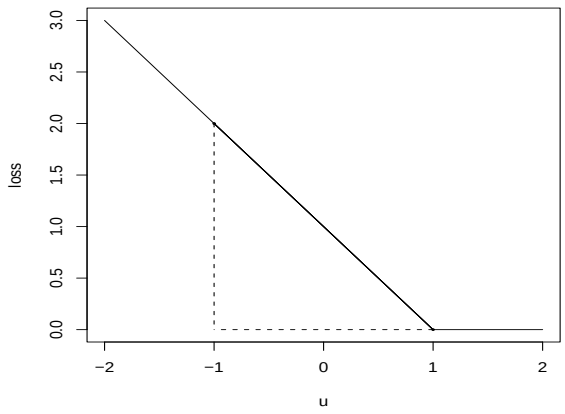

Figure 1: Plot of the hinge loss $H_1(u) = [1-u]_+$.

the same sign as $P(\boldsymbol{x}) - 1/2$. Clearly, Fisher consistency requires the loss function asymptotically yields the Bayes decision boundary. Fisher consistency is also known as "classification-calibration" (Bartlett et al., 2006). For the SVM, Lin (2002) shows that the minimizer of $E[[1-Yf(\boldsymbol{X})]_+|\boldsymbol{X} = \boldsymbol{x}]$ is $\text{sign}(P(\boldsymbol{x}) - 1/2)$ and consequently the hinge loss of the SVM is Fisher consistent. Interestingly, the SVM only targets on the classification set $\{\boldsymbol{x} : P(\boldsymbol{x}) \geq 1/2\}$ without estimating $P(\boldsymbol{x})$ itself.

Extending the binary SVM to multicategory SVM is not trivial. In this paper, we first review four common extensions in Section 2. Fisher consistency of different extensions will be discussed in Secion 3. Among the four extensions, three extensions are not always Fisher consistent. In Sections 4 and 5, we propose to use bounded constraints and truncation to derive Fisher consistent analogs of different losses discussed in Section 2. Some discussions are given in Section 6.

2 Multicategory SVM

The standard SVM only solves binary problems. However, it is very often for one to encounter multicategory problems in practice. To solve a multicategory problem using the SVM, one can use two possible approaches. The first approach is to solve the multicategory problem via a sequence of binary problems, e.g., one-versus-rest and one-versus-one. The second approach is to generalize the binary SVM into a simultaneous multicategory formulation. In this paper, we will focus on the second approach.

Consider a k-class classification problem with $k \geq 2$. When $k = 2$, the methodology to be discussed here reduces to the binary counterpart in Section 1. Let $\boldsymbol{f} = (f_1, f_2, \cdots, f_k)$ be the decision function vector, where each component represents one class and maps from $\mathcal{S}$ to $\Re$. To remove redundant solutions, a sum-to-zero constraint $\sum_{j=1}^k f_j = 0$ is employed. For any new input vector $\boldsymbol{x}$, its label is estimated via a decision rule $\hat{y} = \text{argmax}_{j=1,2,\cdots,k} f_j(\boldsymbol{x})$. Clearly, the argmax rule is equivalent to the sign function used in the binary case in Section 1.

The extension of the SVM from the binary to multicategory case is nontrivial. Before we discuss the detailed formulation of multicategory hinge loss, we first discuss Fisher consistency for multicategory problems. Consider $y \in \{1, \ldots, k\}$ and let $P_j(\boldsymbol{x}) = P(Y = j|\boldsymbol{x})$. Suppose $V(\boldsymbol{f}(\boldsymbol{x}), y)$ is a multicategory loss function. Then in this context, Fisher consistency requires that $\text{argmax}_j f_j^* = \text{argmax}_j P_j$, where $\boldsymbol{f}^*(\boldsymbol{x}) = (f_1^*(\boldsymbol{x}), \ldots, f_k^*(\boldsymbol{x}))$ denotes the minimizer of $E[V(\boldsymbol{f}(\boldsymbol{X}), Y)|\boldsymbol{X} = \boldsymbol{x}]$. Fisher consistency is a desirable condition of a loss function, although a consistent loss may not always translate into better classification accuracy (Hsu and Lin, 2002, Rifkin and Klautau, 2004).

For simplicity, we only focus on standard learning where all types of misclassification are treated equally. The proposed techniques, however, can be extended to more general settings with unequal losses. Note that a point $(\boldsymbol{x}, y)$ is misclassified by $\boldsymbol{f}$ if $y \neq \text{argmax}_j f_j(\boldsymbol{x})$. Thus a sensible loss V should try to force f_y to be the maximum among k functions. Once V is chosen, multicategory SVM solves the following problem

$$\min_{\boldsymbol{f}} \sum_{j=1}^k J(f_j) + C \sum_{i=1}^n V(\boldsymbol{f}(\boldsymbol{x}_i), y_i) \qquad (4)$$

subject to $\sum_{j=1}^k f_j(\boldsymbol{x}) = 0$.

The key of extending the SVM from binary to multicategory is the choice of loss V. There are a number of extensions of the binary hinge loss to the multicategory case proposed in the literature. We consider the following four commonly used extensions and their Fisher consistency or inconsistency will be discussed. For inconsistent losses, we propose two methods to make them Fisher consistent.

(a). (Naive hinge loss, c.f., Zou et al. 2006) $[1 - f_y(\boldsymbol{x})]_+$;

(b). (Lee et al., 2004) $\sum_{j \neq y} [1 + f_j(\boldsymbol{x})]_+$;

(c). (Vapnik, 1998; Weston and Watkins, 1999; Bredensteiner and Bennett, 1999; Guermeur, 2002) $\sum_{j \neq y} [1 - (f_y(\boldsymbol{x}) - f_j(\boldsymbol{x}))]_+$;

(d). (Crammer and Singer, 2001; Liu and Shen, 2006) $[1 - \min_j (f_y(\boldsymbol{x}) - f_j(\boldsymbol{x}))]_+$.

Note that the constant 1 in these losses can be changed to a general positive value. However, the resulting losses will be equivalent to the current ones by re-scaling $\boldsymbol{f}$.

As a remark, we note that the sum-to-zero constraint is essential for Losses (a) and (b), not so for Losses (c) and (d). It is easy to see that all these losses try to encourage f_y to be the maximum among k functions, either explicitly or implicitly. In the next section, we explore Fisher consistency of these four multicategory hinge loss functions.

3 Fisher Consistency of Multicategory Hinge Losses

In this section, we discuss Fisher consistency of all four losses. We would like to clarify here that some of the Fisher consistency results are already available in the literature. There are previous studies on Fisher consistency of multicategory SVMs such as Zhang (2004), Lee et al. (2004), Tewari and Bartlett (2005), and Zou et al. (2006). The goal of this paper is to first review Fisher consistency or inconsistency of these losses and then explore ways, motivated from the discussions in this section, to modify inconsistent extensions to be consistent.

Inconsistency of Loss (a):

Lemma 1. *The minimizer* $\boldsymbol{f}^*$ *of* $E[[1 - f_Y(\boldsymbol{X})]_+|\boldsymbol{X} = \boldsymbol{x}]$ *subject to* $\sum_j^k f_j(\boldsymbol{x}) = 0$ *satisfies the following:* $f_j^*(\boldsymbol{x}) = -(k-1)$ *if* $j = argmin_j P_j(\boldsymbol{x})$ *and* 1 *otherwise.*

Proof: $E[[1 - f_Y(\boldsymbol{X})]_+] = E[\sum_{l=1}^k [1 - f_l(\boldsymbol{X})]_+ P_l(\boldsymbol{X})]$. For any fixed $\boldsymbol{X} = \boldsymbol{x}$, our goal is to minimize $\sum_{l=1}^k [1 - f_l(\boldsymbol{x})]_+ P_l(\boldsymbol{x})$.

We first show the minimizer $\boldsymbol{f}^*$ satisfies $f_j^* \le 1$ for $\forall j = 1, \ldots, k$. To show this, suppose a solution $\boldsymbol{f}^1$ having $f_j^1 > 1$. Then we can construct another solution $\boldsymbol{f}^2$ with $f_j^2 = 1$ and $f_l^2 = f_l^1 + A$, where $A = (f_j^1 - 1)/(k-1) > 0$. Then $\sum_l f_l^2 = 0$ and $f_l^2 > f_l^1; \forall l \ne j$. Consequently, $\sum_{l=1}^k [1 - f_l^2]_+ P_l < \sum_{l=1}^k [1 - f_l^1]_+ P_l$. This implies that $\boldsymbol{f}^1$ cannot be the minimizer. Therefore, the minimizer $\boldsymbol{f}^*$ satisfies $f_j^* \le 1$ for $\forall j$.

Using the property of $\boldsymbol{f}^*$, we only need to consider $\boldsymbol{f}$ with $f_j \le 1$ for $\forall j$. Thus, $\sum_{l=1}^k [1 - f_l(\boldsymbol{x})]_+ P_l(\boldsymbol{x}) = \sum_{l=1}^k (1 - f_l(\boldsymbol{x})) P_l(\boldsymbol{x}) = 1 - \sum_{l=1}^k f_l(\boldsymbol{x}) P_l(\boldsymbol{x})$. Then the problem reduces to

$$\max_{\boldsymbol{f}} \quad \sum_{l=1}^k P_l(\boldsymbol{x}) f_l(\boldsymbol{x}) \tag{5}$$

$$\text{subject to} \quad \sum_{l=1}^k f_l(\boldsymbol{x}) = 0; f_l(\boldsymbol{x}) \le 1, \forall l. \tag{6}$$

It is easy to see that the solution satisfies $f_j^*(\boldsymbol{x}) = -(k-1)$ if $j = \text{argmin}_j P_j(\boldsymbol{x})$ and 1 otherwise. □

From Lemma 1, we can see that Loss (a) is not Fisher consistent since except the smallest element, all remaining elements of its minimizer are 1. Consequently, the argmax rule cannot be uniquely determined and thus the loss is not Fisher consistent when $k \ge 3$ no matter how the P_j's are distributed.

Consistency of Loss (b):

Lemma 2. *The minimizer* $\boldsymbol{f}^*$ *of* $E[\sum_{j \ne Y}[1 + f_j(\boldsymbol{X})]_+|\boldsymbol{X} = \boldsymbol{x}]$ *subject to* $\sum_j^k f_j(\boldsymbol{x}) = 0$ *satisfies the following:* $f_j^*(\boldsymbol{x}) = k - 1$ *if* $j = argmax_j P_j(\boldsymbol{x})$ *and* -1 *otherwise.*

Proof: Note that $E[\sum_{j \ne Y}[1 + f_j(\boldsymbol{X})]_+] = E[E(\sum_{j \ne Y}[1 + f_j(\boldsymbol{x})]_+|\boldsymbol{X} = \boldsymbol{x})]$. Thus, it is sufficient to consider the minimizer for a given $\boldsymbol{x}$ and $E(\sum_{j \ne Y}[1 + f_j(\boldsymbol{x})]_+|\boldsymbol{X} = \boldsymbol{x}) = \sum_{l=1}^k \sum_{j \ne l}[1 + f_j(\boldsymbol{x})]_+ P_l(\boldsymbol{x})$.

Next, we show the minimizer $\boldsymbol{f}^*$ satisfies $f_j^* \ge -1$ for $\forall j = 1, \ldots, k$. To show this, suppose a solution $\boldsymbol{f}^1$ having $f_j^1 < -1$. Then we can construct another solution $\boldsymbol{f}^2$ with $f_j^2 = -1$ and $f_l^2 = f_l^1 - A$, where $A = (-1 - f_j^1)/(k-1) > 0$. Then $\sum_l f_l^2 = 0$ and $f_l^2 < f_l^1; \forall l \ne j$. Consequently, $\sum_{l=1}^k \sum_{j \ne l}[1 + f_j^2]_+ P_l < \sum_{l=1}^k \sum_{j \ne l}[1 + f_j^1]_+ P_l$. This implies that $\boldsymbol{f}^1$ cannot be the minimizer. Therefore, the minimizer $\boldsymbol{f}^*$ satisfies $f_j^* \ge -1$ for $\forall j$.

Using the property of $\boldsymbol{f}^*$, we only need to consider $\boldsymbol{f}$ with $f_j \ge -1$ for $\forall j$. Thus, $\sum_{l=1}^k \sum_{j \ne l}[1 + f_j]_+ P_l = \sum_{l=1}^k P_l \sum_{j \ne l}(1 + f_j) = \sum_{l=1}^k P_l (k - 1 + \sum_{j \ne l} f_j) = \sum_{l=1}^k P_l (k - 1 - f_l) = k - 1 - \sum_{l=1}^k P_l f_l$. Consequently, minimizing $\sum_{l=1}^k \sum_{j \ne l}[1 + f_j]_+ P_l$ is equivalent to maximizing $\sum_{l=1}^k P_l f_l$. Then the problem reduces to

$$\max_{\boldsymbol{f}} \quad \sum_{l=1}^k P_l(\boldsymbol{x}) f_l(\boldsymbol{x}) \tag{7}$$

$$\text{subject to} \quad \sum_{l=1}^k f_l(\boldsymbol{x}) = 0; f_l(\boldsymbol{x}) \ge -1, \forall l. \tag{8}$$

It is easy to see that the solution satisfies $f_j^*(\boldsymbol{x}) = k - 1$ if $j = \text{argmax}_j P_j(\boldsymbol{x})$ and -1 otherwise. □

Lemma 2 implies that Loss (b) yields the Bayes classification boundary asymptotically, consequently it is a Fisher consistent loss. A similar result was also established by Lee et al. (2004).

Inconsistency of Loss (c):

Lemma 3. *Consider* $k = 3$ *with* $1/2 > P_1 > P_2 > P_3$. *Then minimizer(s) of* $E[\sum_{j \ne Y}[1 - (f_Y(\boldsymbol{X}) - f_j(\boldsymbol{X}))]_+|\boldsymbol{X} = \boldsymbol{x}]$ *is(are) as follows:*

- $P_2 = 1/3$*: Any* $\boldsymbol{f}^* = (f_1^*, f_2^*, f_3^*)$ *satisfying* $f_1^* \ge f_2^* \ge f_3^*$ *and* $f_1^* - f_3^* = 1$ *is a minimizer.*
- $P_2 > 1/3$*:* $\boldsymbol{f}^* = (f_1^*, f_2^*, f_3^*)$ *satisfying* $f_1^* \ge f_2^* \ge f_3^*$, $f_1^* = f_2^*$, *and* $f_2^* - f_3^* = 1$.
- $P_2 < 1/3$*:* $\boldsymbol{f}^* = (f_1^*, f_2^*, f_3^*)$ *satisfying* $f_1^* \ge f_2^* \ge f_3^*$, $f_2^* = f_3^*$, *and* $f_1^* - f_2^* = 1$.

Proof: Note that $E[\sum_{j\neq Y}[1-(f_Y(\boldsymbol{X})-f_j(\boldsymbol{X}))]_+|\boldsymbol{X}=\boldsymbol{x}]$ can be rewritten as $\sum_{l=1}^k P_l \sum_{j\neq l}[1-(f_l(\boldsymbol{x})-f_j(\boldsymbol{x}))]_+$. By the non-increasing property of $[1-u]_+$, the minimizer $\boldsymbol{f}^*$ must satisfy $f_1^* \geq \ldots \geq f_k^*$ if $P_1 \geq \ldots P_k$.

Next we focus on $k=3$. Consider $f_1 \geq f_2 \geq f_3$ with $f_1 - f_2 = A \geq 0$ and $f_2 - f_3 = B \geq 0$. Then the objective function becomes

$$\begin{aligned} L(A,B) &= P_1[[1-A]_+ + [1-(A+B)]_+] \quad (9)\\ &\quad +P_2[1+A+[1-B]_+] \quad (10)\\ &\quad +P_3[1+A+B+1+B]. \quad (11)\end{aligned}$$

Before proving the lemma, we first need to prove that the minimizer must satisfy that $A+B \leq 1$. To this end, we prove it in two steps: (1) $A \leq 1$, $B \leq 1$; (2) $A+B \leq 1$.

To prove (1), suppose that the minimizer satisfies $A > 1$. Using contradiction, we can consider another solution $\boldsymbol{f}^1$ with $A^1 = A - \epsilon > 1$ with $\epsilon > 0$ and $B^1 = B$. It is easy to see that $L(A-\epsilon, B) < L(A,B)$. Thus, the minimizer must have $A \leq 1$. Similarly, we can show $B \leq 1$.

To prove (2), we know $A \leq 1$ and $B \leq 1$ by property (1). Suppose the minimizer have $A+B > 1$. Using contradiction, consider another solution $\boldsymbol{f}^1$ with $A^1 + B^1 = A + B - \epsilon > 1$ and $A^1 = A - \epsilon$ and $B^1 = B$. Then by the fact that $P_1 < P_2 + P_3$, we have

$$L(A^1, B^1) - L(A,B) = P_1\epsilon - P_2\epsilon - P_3\epsilon < 0. \quad (12)$$

This implies $A + B \leq 1$.

Using properties (1) and (2), minimizing $L(A,B)$ can be reduced as follows:

$$\begin{aligned} \min_{A,B} \quad & (-2P_1+P_2+P_3)A + (-P_1-P_2+2P_3)B \\ \text{s.t.} \quad & A+B \leq 1, A \geq 1, B \geq 1. \end{aligned} \quad (13)$$

If $P_2 = 1/3$, then $(-2P_1+P_2+P_3) = (-P_1-P_2+2P_3) < 0$. Therefore, the solution of (13) satisfies $A+B=1$, implying $f_1^* - f_3^* = 1$. If $P_2 > 1/3$, then $0 > (-2P_1+P_2+P_3) > (-P_1-P_2+2P_3)$ and consequently the solution satisfies $A=0$ and $B=1$, i.e., $f_1^* = f_2^*$ and $f_2^* - f_3^* = 1$. If $P_2 < 1/3$, then $0 > (-P_1-P_2+2P_3) > (-2P_1+P_2+P_3)$ and consequently the solution satisfies $A=1$ and $B=0$, i.e., $f_1^* - f_2^* = 1$ and $f_2^* = f_3^*$. The desired result then follows. □

Lemma 3 tells us that Loss (c) may be Fisher inconsistent. In fact, in the case of $k=3$ with $1/2 > P_1 > P_2 > P_3$, Loss (c) is Fisher consistent only when $P_2 < 1/3$.

Remark: Lee et al. (2004) considered a special case with $P_2 > 1/3$. Our Lemma 3 here is more general. Interestingly, for the case of $P_2 = 1/3$, there are infinite minimizers although the loss is convex. For example, if $(P_1, P_2, P_3) = (5/12, 1/3, 1/4)$, both $(1/2, 0, -1/2)$ and $(1/3, 1/3, -2/3)$ are minimizers. In general, one cannot claim uniqueness of the minimizer unless the objective function to be minimized is *strictly* convex.

Inconsistency of Loss (d):

Denote $\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), y)) = \{f_y(\boldsymbol{x}) - f_j(\boldsymbol{x}); j \neq y\}$ and $H_1(u) = [1-u]_+$. Then Loss (d) can written as $H_1(\min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), y))$.

Lemma 4. *The minimizer $\boldsymbol{f}^*$ of $E[H_1(\min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{X}), Y))|\boldsymbol{X}=\boldsymbol{x}]$ has the following properties:*
(1). If $\max_j P_j > 1/2$, then $argmax_j f_j^* = argmax_j P_j$ and $\min \boldsymbol{g}^*(\boldsymbol{f}(\boldsymbol{x}), argmax_j f_j^*) = 1$;
(2). If $\max_j P_j < 1/2$, then $\boldsymbol{f}^* = \boldsymbol{0}$.

Proof: Note $E[H_1(\min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{X}), Y))]$ can be written as

$$\begin{aligned} & E[E(H_1(\min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), Y))|\boldsymbol{X}=\boldsymbol{x})] \\ = \; & E[\sum_{j=1}^k P_j(\boldsymbol{X}) H_1(\min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{X}), j))] \\ = \; & E[\sum_{j=1}^k P_j(\boldsymbol{X})(1 - \min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{X}), j))_+]. \end{aligned}$$

For any given $\boldsymbol{X} = \boldsymbol{x}$, let $g_j = \min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), j)$; $j = 1, \cdots, k$. The problem reduces to minimizing $\sum_{j=1}^k P_j(1-g_j)_+$. To prove the lemma, we utilize two properties of the minimizer $\boldsymbol{f}^*$: (1) The minimizer satisfies $\max g_j^* \leq 1$; (2) Let $j_0 = \text{argmax} g_j^*$. For $\forall l \neq j_0$, g_l^* equals to $-\max g_j^* = -g_{j_0}^*$. Using property (1), we have

$$\sum_{j=1}^k P_j(1-g_j^*)_+ = 1 - \sum_{j=1}^k P_j g_j^*.$$

Therefore, minimizing $\sum_{j=1}^k P_j(1-g_j^*)_+$ is equivalent to maximizing $\sum_{j=1}^k P_j g_j^*$. Using property (2), the problem reduces to

$$\max_{\boldsymbol{f}} (2P_{j_0} - 1) g_{j_0} \quad \text{for } g_{j_0} \in [0,1].$$

Clearly, the minimizer satisfies the following conditions: If $P_{j_0} > 1/2$, $g_{j_0} = 1$; if $P_{j_0} < 1/2$, $g_{j_0} = 0$. The desired result of the lemma then follows.

We are now left to show the two properties of $\boldsymbol{f}^*$.
Property (1): The minimizer $\boldsymbol{f}^*$ satisfies $\max g_j^* \leq 1$.
To show this property, we note $(1 - g_{j_0}^*)_+ = 0$ for $g_{j_0}^* \geq 1$. However, for $l \neq j_0$,

$$\begin{aligned} & \min_{j\neq l}\{f_l^* - f_j^*\} = f_l^* - f_{j_0}^* \\ & \leq -\min_{l\neq j_0}\{f_{j_0}^* - f_l^*\} \\ & = -g_{j_0}^* = -\max g_j^*, \end{aligned} \quad (14)$$

which is less than -1 if $\max g_j^* > 1$. Therefore, $\boldsymbol{f}^*$ cannot be a minimizer if $\max g_j^* > 1$. Property (1) then follows.

Property (2): For $\forall l \neq j_0$, g_l^* equals to $-\max g_j^* = -g_{j_0}^*$.
Since $g_l^* \leq -g_{j_0}^*$ for $\forall l \neq j_0$ as shown in (14), to maximize $\sum_{j=1}^k P_j g_j^*$, property (2) holds. □

Lemma 4 suggests that $H_1(\min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), y))$ is Fisher consistent when $\max_j P_j > 1/2$, i.e., when there is a dominating class. Except for the Bayes decision boundary, this condition always holds for a binary problem. For a problem with $k > 2$, however, existence of a dominating class may not be guaranteed. If $\max_j P_j(\boldsymbol{x}) < 1/2$ for a given $\boldsymbol{x}$, then $\boldsymbol{f}^*(\boldsymbol{x}) = \mathbf{0}$ and consequently the argmax of $\boldsymbol{f}^*(\boldsymbol{x})$ cannot be uniquely determined. The Fisher inconsistency of Loss (d) was also noted by Zhang (2004) and Tewari and Bartlett (2005).

4 A Consistent Hinge Loss with Bounded Constraints

In Section 3, we show that Losses (a), (c), and (d) may not always be Fisher consistent while in contrast, Loss (b) is always Fisher consistent. An interesting observation we have is that although Loss (a) is inconsistent while Loss (b) is consistent, the reduced problems (5) and (7) in the derivation of their asymptotic properties are very similar. In fact, the only difference is that Loss (b) has the constraint $f_l(\boldsymbol{x}) \geq -1$ for $\forall l$, and in contrast, Loss (a) has the constraint $f_l(\boldsymbol{x}) \leq 1$ for $\forall l$. Our observation is that one can add additional constraints on $\boldsymbol{f}$ to force Loss (a) to be consistent. In fact, if additional constraints $f_j \geq -1/(k-1)$ for $\forall j$ are imposed on (5), then the minimizer becomes $f_j^*(\boldsymbol{x}) = 1$ if $j = \text{argmax}_j P_j(\boldsymbol{x})$ and $-1/(k-1)$ otherwise. Consequently, the corresponding loss is Fisher consistent.

To be specific, we can modify the scale of Loss (a) and propose the following new loss:
$[k - 1 - f_y(\boldsymbol{x})]_+$ subject to $\sum_j^k f_j(\boldsymbol{x}) = 0$ and $f_l(\boldsymbol{x}) \geq -1$ for $\forall l$.
Clearly, this loss can be reduced to

(e). $-f_y(\boldsymbol{x})$ subject to $\sum_j^k f_j(\boldsymbol{x}) = 0$ and $-1 \leq f_l(\boldsymbol{x}) \leq k-1$ for $\forall l$.

Consistency of Loss (e):

Lemma 5. *The minimizer $\boldsymbol{f}^*$ of $E[-f_Y(\boldsymbol{X})]$, subject to $\sum_j^k f_j(\boldsymbol{x}) = 0$ and $-1 \leq f_l(\boldsymbol{x}) \leq k-1$ for $\forall l$, satisfies the following: $f_j^*(\boldsymbol{x}) = k-1$ if $j = argmax_j P_j(\boldsymbol{x})$ and -1 otherwise.*

Proof: The proof is analogous to that of Lemma 1. It is easy to see that the problem can be reduced to

$$\max_{\boldsymbol{f}} \sum_{l=1}^k P_l(\boldsymbol{x}) f_l(\boldsymbol{x}) \tag{15}$$

$$\text{s.t.} \quad \sum_{l=1}^k f_l(\boldsymbol{x}) = 0; -1 \leq f_l(\boldsymbol{x}) \leq k-1, \forall l.$$

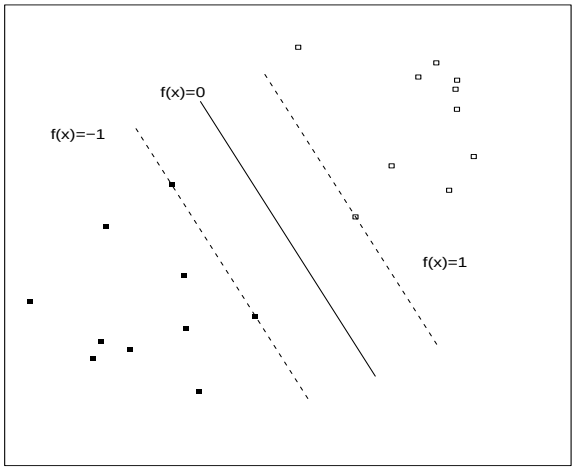

Figure 2: Illustration of the binary SVM classifier using the hinge loss (a).

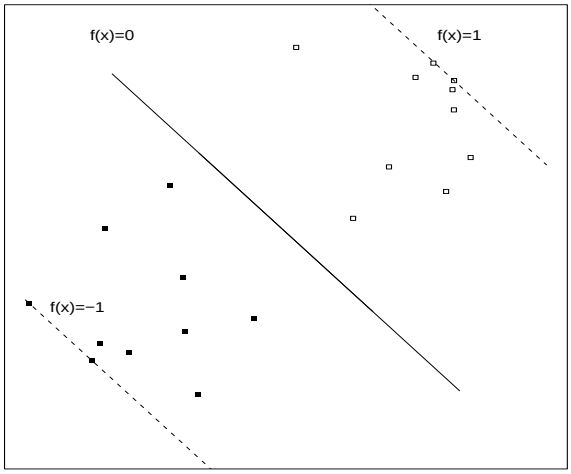

Figure 3: Illustration of the binary SVM classifier using the bounded hinge loss (e).

Thus, the solution satisfies $f_j^*(\boldsymbol{x}) = k-1$ if $j = \text{argmax}_j P_j(\boldsymbol{x})$ and -1 otherwise. □

It is worthwhile to point out that Losses (b) and (c) can be reduced to Loss (e) using bounded constraints. Specifically, for Loss (b), if $-1 \leq f_l(\boldsymbol{x}) \leq k-1$ for $\forall l$ and $\sum_{l=1}^k f_l(\boldsymbol{x}) = 0$, then $\sum_{y \neq j}[1 + f_j]_+ = \sum_{y \neq j}(1 + f_j) = k - 1 - f_y$. Thus, it is equivalent to Loss (e). For Loss (c), we can rewrite it in an equivalent form as $\sum_{j \neq y}[k - (f_y - f_j)]_+$. Then if $-1 \leq f_l(\boldsymbol{x}) \leq k-1, \forall l$ and $\sum_{l=1}^k f_l(\boldsymbol{x}) = 0$, $\sum_{j \neq y}[k - (f_y - f_j)]_+ = \sum_{j \neq y}(k - (f_y - f_j)) = k(k-1) - kf_y$. Thus, it is equivalent to Loss (e) as well.

As a remark, we want to point out that the constraint $-1 \leq f_l(\boldsymbol{x}) \leq k-1$ for $\forall l$ can be difficult to implement for all $\boldsymbol{x} \in \mathcal{S}$. For simplicity of learning, we suggest to relax such constraints on all training points only, that is $-1 \leq f_l(\boldsymbol{x}_i) \leq k-1$ for $i = 1, \ldots, n$ and $l = 1, \ldots, k$. Then we have the

following multicategory learning method:

$$\min_{\boldsymbol{f}} \sum_{j=1}^{k} \|f_j\|_2^2 - C \sum_{i=1}^{n} f_{y_i}(\boldsymbol{x}_i) \quad (16)$$

$$\text{s.t.} \quad \sum_{j}^{k} f_j(\boldsymbol{x}_i) = 0; f_l(\boldsymbol{x}_i) \geq -1;$$
$$l = 1, \ldots, k, i = 1, \ldots, n.$$

To further understand learning using the proposed Loss (e), we discuss the binary case with $y \in \{\pm 1\}$. Recall in the separable case, the standard SVM tries to find a hyperplane with maximum separation, i.e., the distance between $f(\boldsymbol{x}) = \pm 1$ is maximized. As we can see Figure 2, only a small subset of the training data, the so called support vectors, determines the solution. In contrast, using the relaxed bounded constraints in Loss (e) amounts to forcing $f(\boldsymbol{x}_i) \in [-1, 1]$ for all training points. As a result, in the separable case, this new classifier tries to find a hyperplane so that the total "distance" of all training points to the classification boundary, $\sum_{i=1}^{n} y_i f(\boldsymbol{x}_i)$, subject to $f(\boldsymbol{x}_i) \in [-1, 1]; i = 1, \ldots, n$, is maximized as shown in Figure 3. Thus, all points play a role in determining the final solution. In summary, we can conclude that Loss (e) aims for a different classifier from that of Loss (a) due to the bounded constraints.

Note that in contrast to the notion of support vectors of the SVM, the new classifier in (16) utilizes all training points to determine its solution, as illustrated in Figure 3. In order to make the classifier sparse in training points, one can select a fraction of training points to approximate the classifier using the whole training set without jeopardizing the performance in classification. Using such a sparse classifier may help to reduce the computational cost, especially for large training datasets. The idea of Import Vector Machine (IVM, Zhu and Hastie, 2005) may be proven useful here.

Although Losses (a), (b) and (c) can all be reduced to Loss (e) using bounded constraints, Loss (d) appears to behave differently. Currently, we are not able to make Loss (d) Fisher consistent using bounded constraints due to special properties of the min function. In Section 5, we propose to use truncation to make Loss (d) Fisher consistent.

5 Fisher Consistent Truncated Hinge Loss

In many applications, outliers may exist in the training sample and unbounded losses can be sensitive to such points (Shen et al. 2003, Liu et al, 2005, Liu and Shen, 2006). The hinge loss is unbounded and truncating unbounded hinge losses may help to improve robustness of the corresponding classifiers. In this section, we explore the idea of truncation on Loss (d). We show that the truncated version of Loss (d) can be Fisher consistent for certain truncating locations.

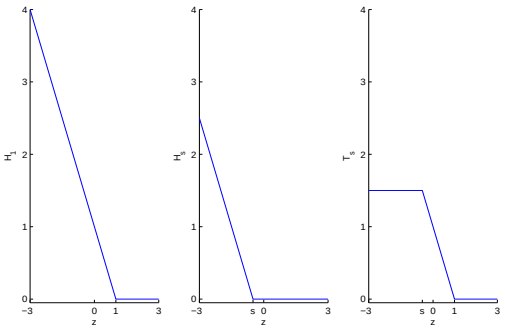

Figure 4: The left, middle, and right panels display functions $H_1(u)$, $H_s(u)$, and $T_s(u)$ respectively.

Define $T_s(u) = H_1(u) - H_s(u)$. Then $T_s(\min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), y))$ with $s \leq 0$ becomes a truncated version of Loss (d). Figure 4 shows functions $H_1(u)$, $H_s(u)$, and $T_s(u)$. We first show in Lemma 6 that for a binary problem, T_s is Fisher consistent for any $s \leq 0$. For multicategory problems, truncating $H_1(\min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), y))$ can make it Fisher consistent even in the situation of no dominating class as shown in Lemma 7.

The following lemma establishes Fisher consistency of the truncated hinge loss T_s for the binary case:

Lemma 6. *The minimizer of* $E[T_s(Yf(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}]$ *has the same sign as* $P(\boldsymbol{x}) - 1/2$ *for any* $s \leq 0$.

Proof: Notice $E[T_s(Yf(\boldsymbol{X}))] = E[E(T_s(Yf(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x})]$. We can minimize $E[T_s(Yf(\boldsymbol{X}))]$ by minimizing $E(T_s(Yf(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x})$ for every $\boldsymbol{x}$.

For any fixed $\boldsymbol{x}$, $E(T_s(Yf(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x})$ can be written as $P(\boldsymbol{x})T_s(f(\boldsymbol{x})) + (1 - P(\boldsymbol{x}))T_s(-f(\boldsymbol{x}))$. Since T_s is a non-increasing function, the minimizer f^* must satisfy that $f^*(\boldsymbol{x}) \geq 0$ if $P(\boldsymbol{x}) > 1/2$ and $f^*(\boldsymbol{x}) \leq 0$ otherwise. Thus, it is sufficient to show that $f = 0$ is not a minimizer. Without loss of generality, assume $P(\boldsymbol{x}) > 1/2$, then $E(T_s(0)|\boldsymbol{X} = \boldsymbol{x}) = 1$. Consider another solution $f(\boldsymbol{x}) = 1$. Then $E(T_s(Yf(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}) = (1 - P)T_s(-1) \leq 2(1 - P) < 1$ for any $s \leq 0$. Therefore, $f(\boldsymbol{x}) = 1$ gives a smaller value of $E(T_s(Yf(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x})$ than $f(\boldsymbol{x}) = 0$, which implies that $f = 0$ is not a minimizer. We can then conclude that $f^*(\boldsymbol{x})$ has the same sign as $P(\boldsymbol{x}) - 1/2$. □

Truncating Loss (d):

Lemma 7. *The minimizer* $\boldsymbol{f}^*$ *of* $E[T_s(\min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{X}), Y))|\boldsymbol{X} = \boldsymbol{x}]$ *satisfies that* $argmax_j f_j^* = argmax_j P_j$, *for any* $s \in [-\frac{1}{k-1}, 0]$.

Proof: Note that $E[T_s(\min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{X}), Y))] =$

$E[\sum_{j=1}^k T_s(\min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{X}), j) P_j(\boldsymbol{X}))]$. For any given $\boldsymbol{x}$, we need to minimize $\sum_{j=1}^k T_s(g_j)P_j$ where $g_j = \min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), j)$. By definition and the fact that $\sum_{j=1}^k f_j = 0$, we can conclude that $\max_j g_j \geq 0$ and at most one of g_j's is positive. Let j_p satisfy that $P_{j_p} = \max_j P_j$. Then using the non-increasing property of T_s, the minimizer $\boldsymbol{f}^*$ satisfies that $g^*_{j_p} \geq 0$.

We are now left to show that $g^*_{j_p} \neq 0$, equivalently that $\mathbf{0}$ cannot be a minimizer. To this end, assume $P_{j_p} > 1/k$ and consider another solution $\boldsymbol{f}^0$ with $f^0_{j_p} = (k-1)/k$ and $f^0_j = -1/k$ for $\forall j \neq j_p$. Then $g^0_{j_p} = 1$ and $g^0_j = -1$ for $\forall j \neq j_p$. Clearly, $\sum_{j=1}^k T_s(g^0_j)P_j \leq (1 + 1/(k-1))(1 - P_{j_p}) < 1$. Therefore, $\boldsymbol{f}^0$ gives a smaller value of $\sum_{j=1}^k T_s(g_j)P_j$ than $\mathbf{0}$ and consequently $g^*_{j_p} > 0$. The desired result then follows. □

Remark: The truncation operation can be applied to many other losses such as Loss (b). Denote $H^*_s(u) = [u - s]_+$. Then Loss (b) can be expressed as $\sum_{j=1}^k I(y \neq j) H^*_{-1}(f_j(\boldsymbol{x}))$. Define $T^*_s(u) = H^*_{-1}(u) - H^*_s(u)$ for $s \geq 0$. Then the truncated version of Loss (b) becomes $\sum_{j=1}^k I(y \neq j) T^*_s(f_j(\boldsymbol{x}))$. It can be shown that the truncated loss (b) is Fisher consistent for any $s \leq 0$ (Wu and Liu, 2006a).

6 Discussion

Fisher consistency is a desirable property for a loss function in classification. It ensures the corresponding classifier delivers the Bayes classification boundary asymptotically. Consequently, we view Fisher consistency a necessary property for any "good" loss to have.

In this paper, we focus on the method of the SVM and discuss Fisher consistency of several commonly used multicategory hinge losses. We study four different losses and three out of four are Fisher inconsistent. For Fisher inconsistent losses, we propose two methods to make them Fisher consistent.

Our first approach is to add bounded constraints to force the corresponding hinge loss to be always Fisher consistent. This results in a very interesting new loss (Loss (e)) and a new learning algorithm. In contrast to the notion of support vectors in the standard SVM, the new classifier utilizes all points to determine the classification boundary. Implementation of the new loss involves convex minimization and solves a similar quadratic programming problem as that of the standard SVM. For our future research, we will explore performance of the new classifier and compare it with the standard SVM both theoretically and numerically. With Fisher consistency of the new loss for both binary and multicategory problems, we believe this new classifier is promising.

Our second approach is to use truncation to make Loss (d) Fisher consistent. Interestingly, truncation not only helps to remedy Fisher inconsistency, it also improves robustness of the resulting classifier as discussed in Wu and Liu (2006b). One drawback of the truncated losses, however, is that the corresponding optimization is nonconvex. For a truncated hinge loss, one can decompose it into a difference of two convex functions and then apply the d.c. algorithm (Liu et al., 2005).

Acknowledgements

This work is partially supported by Grant DMS-0606577 from the National Science Foundation and the UNC Junior Faculty Development Award.

References

Bartlett, P. and Jordan, M., and McAuliffe, J. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, **101**,138-156.

Bredensteiner, E. and Bennett, K. (1999). Multicategory classification by support vector machines. *Computational Optimizations and Applications*, 12, 53-79.

Crammer, K., and Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, **2**, 265-292.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and other Kernel-Based Learning Methods*. Cambridge University Press.

Guermeur, Y. (2002). Combining discriminant models with new multiclass SVMS. *Pattern Analysis and Applications (PAA)*, **5**, 168-179.

Hastie, T., Tibshirani, R., and Friedman, J. (2000). *The Elements of Statistical Learning*. Springer-Verlag, New York.

Hsu C. W. and Lin C. J. (2002). A comparison of methods for multiclass support vector machines. IEEE transations on neural networks, 13, 2, 415-425.

Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory Support Vector Machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, **99**, 465, 67-81.

Lin, Y. (2002). Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery*, 6, 259-275.

Liu, Y. and Shen, X. (2006). Multicategory ψ-learning. *Journal of the American Statistical Association*, **101**, 474, 500-509.

Liu, Y., Shen, X., and Doss, H. (2005). Multicategory ψ-learning and support vector machine: computational tools. *Journal of Computational and Graphical Statistics*, **14**, 1, 219-

236.

Rifkin R. and Klautau A. (2004). In defense of one-vs-all classification, *Journal of Machine Learning Research*, **5**, 101-141.

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

Tewari, A. and Bartlett, P. L. (2005). On the consistency of multiclass classification methods. In Proceedings of the 18th Annual Conference on Learning Theory, volume 3559, pages 143-157. Springer.

Vapnik, V. (1998). *Statistical learning theory.* Wiley, New York.

Wahba, G. (1998). Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In: B. Schölkopf, C. J. C. Burges and A. J. Smola (eds), *Advances in Kernel Methods: Support Vector Learning*, MIT Press, 125-143.

Weston, J., and Watkins, C. (1999). Support vector machines for multi-class pattern recognition. *Proceedings of the Seventh European Symposium On Artificial Neural Networks.*

Wu, Y. and Liu, Y. (2006a). On multicategory truncated-hinge-loss support vector machines. To appear in Proceedings of Joint Summer Research Conference on Machine and Statistical Learning: Prediction and Discovery.

Wu, Y. and Liu, Y. (2006b). Robust truncated-hinge-loss support vector machines. Submitted.

Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, **5**, 1225-1251.

Zhu, J. and Hastie, T. (2005). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, **14**, 1, 185-205.

Zou, H., Zhu, J. and Hastie, T. (2006). The margin vector, admissible loss, and multiclass margin-based classifers. Technical report, Department of Statistics, University of Minnesota.