Supplementary materials for this article are available online.
Please click the JCGS link at http://pubs.amstat.org.

Reinforced Multicategory Support Vector Machines

Yufeng LIU and Ming YUAN

Support vector machines are one of the most popular machine learning methods for classification. Despite its great success, the SVM was originally designed for binary classification. Extensions to the multicategory case are important for general classification problems. In this article, we propose a new class of multicategory hinge loss functions, namely reinforced hinge loss functions. Both theoretical and numerical properties of the reinforced multicategory SVMs (MSVMs) are explored. The results indicate that the proposed reinforced MSVMs (RMSVMs) give competitive and stable performance when compared with existing approaches. R implementation of the proposed methods is also available online as supplemental materials.

Key Words: Fisher consistency; Multicategory classification; Regularization; SVM.

1. INTRODUCTION

Classification is a very important statistical task for information extraction from data. Among numerous classification techniques, the Support Vector Machine (SVM) is one of the most well-known large-margin classifiers and has achieved great success in many applications (Boser, Guyon, and Vapnik 1992; Cortes and Vapnik 1995). The basic concept behind the binary SVM is to find a separating hyperplane with maximum separation between the two classes. Because of its flexibility in estimating the decision boundary using kernel learning as well as its ability in handling high-dimensional data, the SVM has become a very popular classifier and has been widely applied in many different fields. More details about the SVM can be found, for example, in the works of Cristianini and Shawe-Taylor (2000), Hastie, Tibshirani, and Friedman (2001), Schölkopf and Smola (2002).

Recent theoretical developments provide us more insight on the success of the SVM. Lin (2004) showed Fisher consistency of binary SVMs in the sense that the theoretical minimizer of the hinge loss yields the Bayes classification boundary. As a result, the SVM

Yufeng Liu is Associate Professor, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (E-mail: *yfliu@email.unc.edu*). Ming Yuan is Associate Professor, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205 (E-mail: *myuan@isye.gatech.edu*).

Journal of Computational and Graphical Statistics, Volume 20, Number 4, Pages 901–919
DOI: 10.1198/jcgs.2010.09206

targets on the decision boundary directly without estimating the conditional class probability. More theoretical characterization of general binary large margin losses can be found in the articles by Zhang (2004b), Bartlett, Jordan, and McAuliffe (2006).

The standard SVM only solves binary problems. However, one often encounters multicategory problems in practice. To solve a multicategory problem using the SVM, typically there are two possible approaches. The first approach is to solve the multicategory problem via a sequence of binary problems, for example, one-versus-rest and one-versus-one (Dietterich and Bakiri 1995; Allwein et al. 2000). The second approach is to generalize the binary SVM to a simultaneous multicategory formulation which deals with all classes at once (Vapnik 1998; Weston and Watkins 1999; Crammer and Singer 2001; Lee, Lin, and Wahba 2004; Liu and Shen 2006). The first approach is conceptually simple to implement since one can use the existing binary techniques directly to solve multicategory problems. Despite its simplicity, the one-versus-rest approach may be inconsistent when there is no dominating class (Liu 2007). On the contrary, Rifkin and Klautau (2004) showed that the one-versus-rest approach can work as accurately as other simultaneous classification methods using a substantial collection of numerical comparisons.

In this article, we reformulate the one-versus-rest approach as an instance of the simultaneous multicategory formulation and focus on various simultaneous extensions. In particular, we propose a convex combination of an existing consistent multicategory hinge loss and another direct generalized hinge loss. Since the two components of the combination intend to enforce correct classification in a complementary fashion, we call this family of loss functions the *reinforced* multicategory hinge loss. We show that the proposed family of loss functions gives rise to a continuum of loss functions that are Fisher consistent. Moreover, the proposed reinforced multicategory SVM (RMSVM) appears to deliver more accurate classification results than the uncombined ones.

The rest of this article is organized as follows. In Section 2.1, we introduce the new reinforced hinge loss functions. Section 2.2 studies Fisher consistency of the new class of loss functions. A computational algorithm of the RMSVM is given in Section 3. In Section 4, we use both simulated examples and an application to lung cancer Microarray data to illustrate performance of the proposed RMSVMs with different choices of the combining weight parameter. Some discussion and remarks are given in Section 5, followed by proofs of the theoretical results in the Appendix.

2. METHODOLOGY

2.1 REINFORCED MULTICATEGORY HINGE LOSSES

Suppose we are given a training dataset containing n training pairs $\{\mathbf{x}_i, y_i\}_{i=1}^n$, iid realizations from probability distribution $P(\mathbf{x}, y)$, where $\mathbf{x}$ is the d-dimensional input and y is the corresponding class label. For simplicity, we consider $\mathbf{x} \in \Re^d$. Our method, however, may be easily extended to include discrete and categorical input variables. For simplicity, in the rest of the article, we shall focus only on the standard learning where all types of misclassification are treated equally. The discussion, however, can be extended straightforwardly to more general settings with unequal losses.

In the binary case, the goal is to search for a function $f(\mathbf{x})$ so that $\text{sign}(f(\mathbf{x}))$ can be used for prediction of class labels for new inputs. The standard binary SVM can be viewed as an example of the regularization framework (Wahba 1999) as follows:

$$\min_f \left[\lambda J(f) + \frac{1}{n} \sum_{i=1}^{n} V(y_i f(\mathbf{x}_i)) \right],$$

where $J(f)$ is the roughness penalty of f, V is the hinge loss with $V(u) = [1-u]_+ = 1-u$ if $u \le 1$ and 0 otherwise, and $\lambda \ge 0$ is a tuning parameter. Denote $P(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$. Lin (2004) showed that the minimizer of $E[V(Yf(\mathbf{X}))|\mathbf{X} = \mathbf{x}]$ has the same sign as $P(\mathbf{x}) - 1/2$ and consequently the hinge loss of SVM targets on the Bayes decision boundary asymptotically. This property is known as Fisher consistency and it is a desirable condition for a loss function in classification.

Extension of the SVM from the binary to multicategory case is nontrivial and the key is the generalization of the binary hinge loss to the multicategory case. Consider a k-class classification problem with $k \ge 2$. Let $\mathbf{f} = (f_1, f_2, \ldots, f_k)$ be the decision function vector, where each component represents one class and maps from $\Re^d$ to $\Re$. For any new input vector $\mathbf{x}$, its label is estimated via a decision rule $\hat{y} = \text{argmax}_{j=1,2,\ldots,k} f_j(\mathbf{x})$. Clearly, the argmax rule is equivalent to the sign function used in the binary case if a sum-to-zero constraint $\sum_{j=1}^{k} f_j = 0$ is employed.

Similarly to the binary case, we consider solving the following problem in order to learn $\mathbf{f}$:

$$\min_{\mathbf{f}} \left[\lambda \sum_{j=1}^{k} J(f_j) + \frac{1}{n} \sum_{i=1}^{n} V(\mathbf{f}(\mathbf{x}_i), y_i) \right], \tag{2.1}$$

subject to $\sum_{j=1}^{k} f_j(\mathbf{x}) = 0$. Here, a sum-to-zero constraint is used to remove redundancy and reduce the dimension of the problem. Note that a point $(\mathbf{x}, y)$ is misclassified by $\mathbf{f}$ if $y \ne \text{argmax}_j f_j(\mathbf{x})$. Thus a sensible loss V should try to encourage f_y to be the maximum.

In the literature, a number of extensions of the binary hinge loss to the multicategory case have been proposed. See, for example, the works by Vapnik (1998), Weston and Watkins (1999), Bredensteiner and Bennett (1999), Crammer and Singer (2001), Lee, Lin, and Wahba (2004), Liu and Shen (2006). In this article, we consider a new class of multicategory hinge loss functions as follows:

$$V(\mathbf{f}(\mathbf{x}), y) = \gamma[(k-1) - f_y(\mathbf{x})]_+ + (1-\gamma) \sum_{j \ne y} [1 + f_j(\mathbf{x})]_+ \tag{2.2}$$

subject to $\sum_{j=1}^{k} f_j(\mathbf{x}) = 0$, where $\gamma \in [0, 1]$. We call the loss function (2.2) the reinforced hinge loss function since there are two terms in the loss and both terms try to force f_y to be the maximum. We choose the constant to be $k-1$ for the first part of the loss since if $f_j = -1$ for $\forall j \ne y$, $f_y = k-1$ using the sum-to-zero constraint. Thus $k-1$ is a natural choice to use for the reinforced loss (2.2). The main motivation for this new loss function is based on the consideration of the argmax rule for multicategory problems. In order to get a correct classification result on a data point, we need to have the corresponding $f_y(\mathbf{x})$ to be the maximum among k different $f_j(\mathbf{x})$; $j = 1, \ldots, k$. To that end, the first term

encourages f_y to be big while the second term encourages other f_j's to be small. As we will discuss later, each separate term of the loss function has certain drawbacks in view of consistency and empirical performance. The proposed combined loss, however, yields better classification performance.

With different choices of γ, (2.2) constitutes a large class of loss functions. When $\gamma = 0$, (2.2) reduces to $\sum_{j\neq y}[1 + f_j(\mathbf{x})]_+$ subject to $\sum_{j=1}^k f_j(\mathbf{x}) = 0$ and it is the same loss as the one used by Lee, Lin, and Wahba (2004). When $\gamma = 1/2$, if we replace $k-1$ in (2.2) by 1, it reduces to $\sum_{j=1}^k [1 - c_j^y f_j(\mathbf{x})]_+$, where $c_j^y = 1$ if $j = y$ and -1 otherwise. This is the loss employed by the one-versus-rest approach (Weston 1999) except that the latter generally does not enforce the sum-to-zero constraint so that the minimization can be decoupled. Because of these connections, the new loss (2.2) can be viewed as a combination of the one-versus-rest approach and the simultaneous classification approach. Our emphasis is the effect of different choices of γ on the resulting classifiers.

To better comprehend the reinforced loss functions in comparison with the 0–1 loss, we rewrite the loss using the multiple comparison vector representation proposed by Liu and Shen (2006). Specifically, Liu and Shen (2006) defined the comparison vector $\mathbf{g}(\mathbf{f}(\mathbf{x}), y) = (f_y(\mathbf{x}) - f_1(\mathbf{x}), \ldots, f_y(\mathbf{x}) - f_{y-1}(\mathbf{x}), f_y(\mathbf{x}) - f_{y+1}(\mathbf{x}), \ldots, f_y(\mathbf{x}) - f_k(\mathbf{x}))$. Then by the argmax classification rule, an instance $(\mathbf{x}, y)$ is misclassified if and only if $\min(\mathbf{g}(\mathbf{f}(\mathbf{x}), y)) \leq 0$. For simplicity, denote $\mathbf{u} = \mathbf{g}(\mathbf{f}(\mathbf{x}), y)$. Then the 0–1 loss can be written as $I(\min_j u_j \leq 0)$. The reinforced loss (2.2) can then be expressed as $\gamma[(k-1) - \sum_{l=1}^{k-1} u_l/k]_+ + (1-\gamma)\sum_{j=1}^{k-1}[1 + \sum_{l=1}^{k-1} u_l/k - u_j]_+$. Figure 1 shows the 0–1 loss and the reinforced hinge loss functions with $\gamma = 1, 0, 0.5$ using the notation $\mathbf{u}$ for $k = 3$. Clearly, various reinforced hinge loss functions are convex upper envelopes of the 0–1 loss function. Furthermore, the shape of the reinforced loss varies dramatically as γ changes. For the reinforced hinge loss functions with $\gamma = 1, 0, 0.5$, the corresponding plots have 2, 4, 6 jointing planes, respectively. As shown later in this article, reinforced SVMs indeed behave very differently when γ varies.

2.2 FISHER CONSISTENCY

Fisher consistency is also known as "classification calibrated" (Bartlett, Jordan, and McAuliffe 2006). Write $P_j(\mathbf{x}) = P(Y = j|\mathbf{x})$. A loss function V is Fisher consistent if and only if $\operatorname{argmax}_j f_j^* = \operatorname{argmax}_j P_j$, where $\mathbf{f}^*(\mathbf{x}) = (f_1^*(\mathbf{x}), \ldots, f_k^*(\mathbf{x}))$ denotes the minimizer of $E[V(\mathbf{f}(\mathbf{X}), Y)|\mathbf{X} = \mathbf{x}]$. For multicategory classification, Zhang (2004a), Tewari and Bartlett (2007) explored Fisher consistency for several convex margin-based multicategory losses. Hill and Doucet (2007) provided geometric illustrations of Fisher consistency for several existing multicategory hinge loss functions. In this section, we investigate Fisher consistency of the reinforced multicategory hinge losses in (2.2) with different choices of γ. Interestingly, there exists a dichotomy: the reinforced hinge loss function is Fisher consistent if and only if $\gamma \leq 1/2$. Denote by $f_1^*, \ldots, f_k^*$ the minimizer of $E(V(\mathbf{f}(\mathbf{X}), Y)|\mathbf{x})$. We have

Theorem 1. *If $\gamma \leq 1/2$, then the reinforced hinge loss function (2.2) under the constraint that $\sum_j f_j(\mathbf{x}) = 0$ is Fisher consistent.*

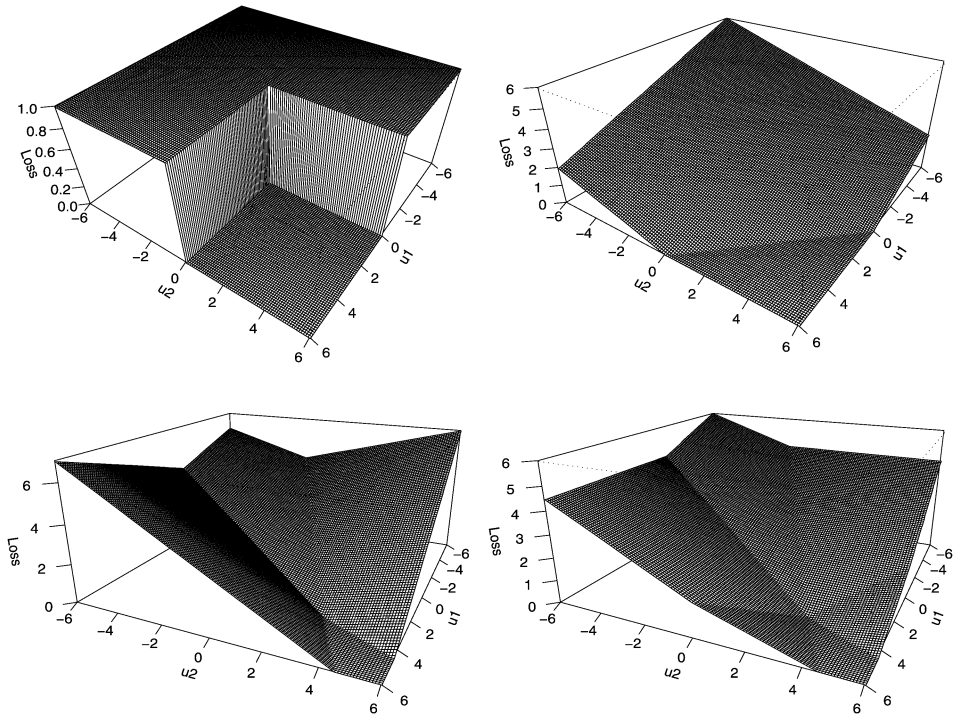

Figure 1. Plots of various loss functions using $\mathbf{u}$ as the argument for $k = 3$: the 0–1 loss on the top-left panel, the reinforced hinge loss functions with $\gamma = 1, 0, 0.5$ on the top-right, bottom-left, bottom-right panels.

Theorem 1 establishes Fisher consistency of the reinforced hinge loss with $\gamma \leq 1/2$. Our next theorem explores the case of $\gamma > 1/2$.

Theorem 2. *If $k > 2$, then for any $\gamma > 1/2$, there exists a set of $P_1(\mathbf{x}) > P_2(\mathbf{x}) \geq \cdots \geq P_k(\mathbf{x})$ such that $f_1^*(\mathbf{x}) = f_2^*(\mathbf{x})$ and therefore $V(\mathbf{f}(\mathbf{x}), y)$ in (2.2) under the constraint that $\sum_j f_j(\mathbf{x}) = 0$ is not always Fisher consistent.*

The proofs are provided in the Appendix. From Theorems 1 and 2, we can conclude that the proposed reinforced hinge loss is Fisher consistent if and only if $0 \leq \gamma \leq 1/2$. This provides a large class of consistent multicategory hinge loss functions. When $\gamma > 1/2$, Fisher consistency cannot be guaranteed when there is no dominating class, that is, $\max_j P_j(\mathbf{x}) < 1/2$. Modifications such as additional constraints as in the article by Liu (2007) may be applied to make the loss consistent. However, such modifications will result in loss functions that are no longer hinge losses and are thus not pursued here.

Interestingly, as indicated by the numerical examples in Section 4, RMSVMs with the values of γ in the middle range of [0,1] such as $\gamma = 0.5$ work better than those of $\gamma = 0$ or 1. This reflects the advantages of proposed combined loss functions which encourage f_y to be maximum both explicitly and implicitly through the two components in (2.2).

3. COMPUTATIONAL ALGORITHM

We now derive a computational algorithm for the RMSVM within the kernel learning framework. Using the representer theorem (Kimeldorf and Wahba 1971; Wahba 1999), $f_j(\mathbf{x})$ can be represented as $b_j + \sum_{i'=1}^{n} K(\mathbf{x}, \mathbf{x}_{i'}) v_{i'j}$, where $K(\cdot,\cdot)$ is the kernel function, and $b_i, v_{i'j}; i' = 1, \ldots, n$ are coefficients for f_j. Then we have

$$f_j(\mathbf{x}_i) = b_j + \sum_{i'=1}^{n} K(\mathbf{x}_i, \mathbf{x}_{i'}) v_{i'j} = b_j + \mathbf{K}_i^T \mathbf{v}_{\cdot j}, \tag{3.1}$$

where $\mathbf{K}_i = (K(\mathbf{x}_i, \mathbf{x}_1), K(\mathbf{x}_i, \mathbf{x}_2), \ldots, K(\mathbf{x}_i, \mathbf{x}_n))^T$ and $\mathbf{v}_{\cdot j} = (v_{1j}, \ldots, v_{nj})^T$. Moreover, $J(f_j) = \frac{1}{2}\mathbf{v}_{\cdot j}^T \mathbf{K} \mathbf{v}_{\cdot j}$. Thus the RMSVM can be reduced to

$$\begin{aligned} \min_{\Theta} \frac{\lambda}{2} \sum_{j=1}^{k} \mathbf{v}_{\cdot j}^T \mathbf{K} \mathbf{v}_{\cdot j} + \frac{1}{n} \sum_{i=1}^{n} \Big(& \gamma \big[(k-1) - b_{y_i} - \mathbf{K}_i^T \mathbf{v}_{\cdot y_i} \big]_+ \\ & + (1-\gamma) \sum_{j \neq y_i} [1 + b_j + \mathbf{K}_i^T \mathbf{v}_{\cdot j}]_+ \Big), \\ \text{s.t.} \quad & \mathbf{e} \sum_{j=1}^{k} b_j + \mathbf{K} \sum_{j=1}^{k} \mathbf{v}_{\cdot j} = \mathbf{0}, \end{aligned} \tag{3.2}$$

where Θ denotes $\{\mathbf{v}_{\cdot j}, b_j\}_{j=1}^{k}$, $\mathbf{K}$ denotes the kernel matrix with the (i, i') element being $K(\mathbf{x}_i, \mathbf{x}_{i'})$, and $\mathbf{e} = (1, 1, \ldots, 1)^T$ is a vector of length n.

To solve (3.2), we introduce nonnegative slack variables $\xi_{ij}; i = 1, \ldots, n, j = 1, \ldots, k$, and then the primal problem of our RMSVM can be written as

$$\begin{aligned} \min_{\Theta, \boldsymbol{\xi}} \frac{n\lambda}{2} \sum_{j=1}^{k} \mathbf{v}_{\cdot j}^T \mathbf{K} \mathbf{v}_{\cdot j} + \sum_{i=1}^{n} \Big(\gamma \xi_{iy_i} + (1-\gamma) \sum_{j \neq y_i} \xi_{ij} \Big), \\ \text{s.t.} \quad \xi_{ij} \geq 0; \qquad i = 1, \ldots, n, \, j = 1, \ldots, k, \\ \xi_{iy_i} + \big(b_{y_i} + \mathbf{K}_i^T \mathbf{v}_{\cdot y_i} - (k-1) \big) \geq 0; \qquad i = 1, \ldots, n, \\ \xi_{ij} - \big(b_j + \mathbf{K}_i^T \mathbf{v}_{\cdot j} + 1 \big) \geq 0; \qquad i = 1, \ldots, n, \, j \neq y_i, \\ \Big(\sum_{j=1}^{k} b_j \Big) \mathbf{e} + \mathbf{K} \Big(\sum_{j=1}^{k} \mathbf{v}_{\cdot j} \Big) = \mathbf{0}. \end{aligned}$$

The corresponding Lagrangian function is

$$\begin{aligned} L_D = & \frac{n\lambda}{2} \sum_{j=1}^{k} \mathbf{v}_{\cdot j}^T \mathbf{K} \mathbf{v}_{\cdot j} + \sum_{i=1}^{n} \Big(\gamma \xi_{iy_i} + (1-\gamma) \sum_{j \neq y_i} \xi_{ij} \Big) \\ & - \sum_{i=1}^{n} \sum_{j=1}^{k} \tau_{ij} \xi_{ij} + \boldsymbol{\delta}^T \left(\mathbf{K} \left(\sum_{j=1}^{k} \mathbf{v}_{\cdot j} \right) + \left(\sum_{j=1}^{k} b_j \right) \mathbf{e} \right) \end{aligned}$$

$$
\begin{aligned}
&- \sum_{i=1}^{n} \alpha_{iy_i} \big(\xi_{iy_i} + b_{y_i} + \mathbf{K}_i^T \mathbf{v}_{\cdot y_i} - (k-1)\big) - \sum_{i=1}^{n} \sum_{j \neq y_i} \alpha_{ij} (\xi_{ij} - b_j - \mathbf{K}_i^T \mathbf{v}_{\cdot j} - 1) \\
&= \frac{n\lambda}{2} \sum_{j=1}^{k} \mathbf{v}_{\cdot j}^T \mathbf{K} \mathbf{v}_{\cdot j} + \sum_{i=1}^{n} \sum_{j=1}^{k} (A_{ij} - \tau_{ij} - \alpha_{ij}) \xi_{ij} + \sum_{i=1}^{n} (k-1) \alpha_{iy_i} + \sum_{i=1}^{n} \sum_{j \neq y_i} \alpha_{ij} \\
&\quad + \sum_{j=1}^{k} b_j \big(-(\boldsymbol{\alpha}_{\cdot j} \cdot (\mathbf{e} - \mathbf{L}_{\cdot j}))^T \mathbf{e} + (\boldsymbol{\alpha}_{\cdot j} \cdot \mathbf{L}_{\cdot j})^T \mathbf{e} + \boldsymbol{\delta}^T \mathbf{e} \big) \\
&\quad + \sum_{j=1}^{k} \big\langle \mathbf{K}(\boldsymbol{\alpha}_{\cdot j} \cdot \mathbf{L}_{\cdot j}) - \mathbf{K}(\boldsymbol{\alpha}_{\cdot j} \cdot (\mathbf{e} - \mathbf{L}_{\cdot j})) + \mathbf{K}\boldsymbol{\delta}, \mathbf{v}_{\cdot j} \big\rangle,
\end{aligned}
$$

where $\alpha_{ij} \geq 0$ and $\tau_{ij} \geq 0$, $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)^T$ are Lagrangian multipliers, $\mathbf{L}_{\cdot j}$ is a vector of length n with its ith element being 0 if $y_i = j$ and 1 otherwise, $\boldsymbol{\alpha}_{\cdot j} \cdot \mathbf{L}_{\cdot j}$ denotes componentwise product between $\boldsymbol{\alpha}_{\cdot j}$ and $\mathbf{L}_{\cdot j}$, and $A_{ij} = [\gamma I(y_i = j) + (1-\gamma) I(y_i \neq j)]$. Setting $\frac{\partial L_D}{\partial \xi_{ij}} = 0$, $\frac{\partial L_D}{\partial b_j} = 0$, and $\frac{\partial L_D}{\partial \mathbf{v}_{\cdot j}} = 0$, we have

$$
\frac{\partial L_D}{\partial \xi_{ij}} = A_{ij} - \tau_{ij} - \alpha_{ij} = 0, \tag{3.3}
$$

$$
\frac{\partial L_D}{\partial b_j} = -(\boldsymbol{\alpha}_{\cdot j} \cdot (\mathbf{e} - \mathbf{L}_{\cdot j}))^T \mathbf{e} + (\boldsymbol{\alpha}_{\cdot j} \cdot \mathbf{L}_{\cdot j})^T \mathbf{e} + \boldsymbol{\delta}^T \mathbf{e} = 0, \tag{3.4}
$$

$$
\frac{\partial L_D}{\partial \mathbf{v}_{\cdot j}} = n\lambda \mathbf{K} \mathbf{v}_{\cdot j} + \mathbf{K}(\boldsymbol{\alpha}_{\cdot j} \cdot \mathbf{L}_{\cdot j}) + \mathbf{K}\boldsymbol{\delta} - \mathbf{K}(\boldsymbol{\alpha}_{\cdot j} \cdot (\mathbf{e} - \mathbf{L}_{\cdot j})) = \mathbf{0}. \tag{3.5}
$$

Due to the positive definitive kernel $K(\cdot, \cdot)$, (3.5) implies that $\mathbf{v}_{\cdot j} = \frac{1}{n\lambda}(\boldsymbol{\alpha}_{\cdot j} \cdot (\mathbf{e} - \mathbf{L}_{\cdot j}) - \boldsymbol{\alpha}_{\cdot j} \cdot \mathbf{L}_{\cdot j} - \boldsymbol{\delta})$. Let $\bar{\boldsymbol{\alpha}} = \frac{1}{k} \sum_{j=1}^{k} (\boldsymbol{\alpha}_{\cdot j} \cdot \mathbf{L}_{\cdot j})$ and $\bar{\tilde{\boldsymbol{\alpha}}} = \frac{1}{k} \sum_{j=1}^{k} (\boldsymbol{\alpha}_{\cdot j} \cdot (\mathbf{e} - \mathbf{L}_{\cdot j}))$. Then from (3.4) and (3.5), we have $\boldsymbol{\delta} = \bar{\tilde{\boldsymbol{\alpha}}} - \bar{\boldsymbol{\alpha}}$ and

$$
\mathbf{v}_{\cdot j} = \frac{1}{n\lambda} [(\boldsymbol{\alpha}_{\cdot j} \cdot (\mathbf{e} - \mathbf{L}_{\cdot j}) - \boldsymbol{\alpha}_{\cdot j} \cdot \mathbf{L}_{\cdot j}) - (\bar{\tilde{\boldsymbol{\alpha}}} - \bar{\boldsymbol{\alpha}})]. \tag{3.6}
$$

After plugging (3.3)–(3.6) into L_D, we can derive the corresponding dual problem as follows:

$$
\begin{aligned}
\min_{\boldsymbol{\alpha}} \; & \frac{1}{2} \sum_{j=1}^{k} \big\langle \big[(\boldsymbol{\alpha}_{\cdot j} \cdot (\mathbf{e} - \mathbf{L}_{\cdot j}) - \boldsymbol{\alpha}_{\cdot j} \cdot \mathbf{L}_{\cdot j}) - (\bar{\tilde{\boldsymbol{\alpha}}} - \bar{\boldsymbol{\alpha}}) \big], \\
& \mathbf{K} \big[(\boldsymbol{\alpha}_{\cdot j} \cdot (\mathbf{e} - \mathbf{L}_{\cdot j}) - \boldsymbol{\alpha}_{\cdot j} \cdot \mathbf{L}_{\cdot j}) - (\bar{\tilde{\boldsymbol{\alpha}}} - \bar{\boldsymbol{\alpha}}) \big] \big\rangle \\
& - n\lambda \sum_{i=1}^{n} (k-1) \alpha_{iy_i} - n\lambda \sum_{i=1}^{n} \sum_{j \neq y_i} \alpha_{ij}, \qquad (3.7) \\
\text{s.t.} \quad & 0 \leq \alpha_{ij} \leq A_{ij}; \qquad i = 1, \ldots, n, \; j = 1, \ldots, k, \\
& [(\boldsymbol{\alpha}_{\cdot j} \cdot (\mathbf{e} - \mathbf{L}_{\cdot j}) - \boldsymbol{\alpha}_{\cdot j} \cdot \mathbf{L}_{\cdot j}) - (\bar{\tilde{\boldsymbol{\alpha}}} - \bar{\boldsymbol{\alpha}})]^T \mathbf{e} = 0; \qquad j = 1, \ldots, k.
\end{aligned}
$$

To further simplify (3.7), define $\boldsymbol{\beta} = (\boldsymbol{\alpha}_{\cdot 1}^T, \ldots, \boldsymbol{\alpha}_{\cdot k}^T)^T$, $\mathbf{e}_j$ as a vector of length k with its jth element being 1 and the remaining ones being 0, $U_j = \mathbf{e}_j^T \otimes I_n$, and V_j as a diagonal matrix with $\mathbf{L}_{\cdot j}$ as the diagonal elements, where $\otimes$ denotes the Kronecker product and I_n denotes the $n \times n$ identity matrix. Then $\boldsymbol{\alpha}_{\cdot j} \cdot \mathbf{L}_{\cdot j} = V_j U_j \boldsymbol{\beta}$ and $\boldsymbol{\alpha}_{\cdot j} \cdot (\mathbf{e} - \mathbf{L}_{\cdot j}) = (I_n - V_j) U_j \boldsymbol{\beta}$. Furthermore, (3.7) can be simplified as the following quadratic programming (QP) problem:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \frac{1}{2} \boldsymbol{\beta}^T \sum_{j=1}^{k} H_j^T \mathbf{K} H_j \boldsymbol{\beta} + \mathbf{g}^T \boldsymbol{\beta}, \\ &\text{s.t.} \quad 0 \le \alpha_{ij} \le A_{ij}; \qquad i = 1, \ldots, n, \, j = 1, \ldots, k, \\ &\qquad \mathbf{e}^T H_j \boldsymbol{\beta} = 0; \qquad j = 1, \ldots, k, \end{aligned} \tag{3.8}$$

where $H_j = (I_n - V_j) U_j - V_j U_j - \frac{1}{k} \sum_{m=1}^{k} (I_n - V_m) U_m + \frac{1}{k} \sum_{m=1}^{k} V_m U_m$ and $\mathbf{g}$ is a vector of length nk with its $(j-1)n + i$th elements being $-n\lambda(k-1)$ if $j = y_i$ and $-n\lambda$ otherwise.

Once $\{\mathbf{v}_j; \, j = 1, \ldots, k\}$ are obtained, we can solve $\mathbf{b}$ either by the KKT conditions or linear programming (LP). More explicitly, with $\{\mathbf{v}_j; \, j = 1, \ldots, k\}$ given, we can obtain $\mathbf{b}$ by solving

$$\begin{aligned} &\min_{\mathbf{b}, \boldsymbol{\eta}} \sum_{j=1}^{k} \left(\gamma \eta_{i y_i} + (1 - \gamma) \sum_{j \ne y_i} \eta_{ij} \right), \\ &\text{subject to} \quad \sum_{j=1}^{k} b_j = 0, \\ &\qquad \eta_{ij} \ge 0; \qquad i = 1, \ldots, n, \, j = 1, \ldots, k, \\ &\qquad \eta_{i y_i} + \left(b_{y_i} + \mathbf{K}_i^T \mathbf{v}_{\cdot y_i} - (k-1) \right) \ge 0; \qquad i = 1, \ldots, n, \\ &\qquad \eta_{ij} - (b_j + \mathbf{K}_i^T \mathbf{v}_{\cdot j} + 1) \ge 0; \qquad i = 1, \ldots, n, \, j \ne y_i. \end{aligned} \tag{3.9}$$

Our algorithm for the RMSVM with a given λ can be summarized as follows:

Step 1: Solve the QP problem (3.8) to obtain solution $\boldsymbol{\beta}$.

Step 2: With $\boldsymbol{\beta}$ given, solve (3.6) to get the solution for $\mathbf{v}_{\cdot j}$; $j = 1, \ldots, k$.

Step 3: With $\{\mathbf{v}_j; \, j = 1, \ldots, k\}$ given, $\mathbf{b}$ can be derived by solving the LP problem (3.9).

4. NUMERICAL EXAMPLES

4.1 SIMULATION

In this section, we use two simulated examples to examine the behavior of the RMSVMs and how their performance varies with γ. Since $\gamma \in [0, 1]$, we examine 11 choices with $\gamma = 0, 0.1, \ldots, 1$. As shown in Theorems 1 and 2, the RMSVMs have Fisher consistency for $\gamma \in [0, 0.5]$ and are not always Fisher consistent for $\gamma > 0.5$. Thus, these values of γ should provide a broad range of behaviors of the corresponding RMSVMs. The case of $\gamma = 0$ corresponds to the version by Lee, Lin, and Wahba (2004).

4.1.1 Example With a Piecewise Linear Bayes Decision Boundary

In this three-class example, $P(Y=1)=P(Y=2)=P(Y=3)=1/3$, $P(X|Y=1)\sim N(\boldsymbol{\mu}=(0,2)^T, 1.5^2\mathbf{I}_2)$, $P(X|Y=2)\sim N(\boldsymbol{\mu}=(-\sqrt{3},-1)^T, 1.5^2\mathbf{I}_2)$, and $P(X|Y=3)\sim N(\boldsymbol{\mu}=(\sqrt{3},-1)^T, 1.5^2\mathbf{I}_2)$. Due to the design of this example, linear learning can be sufficient and the corresponding Bayes boundary is piecewise linear as displayed in the left panel of Figure 3 below.

We simulate n observations for training, n observations for tuning, and a large set for testing. We use the training set to build RMSVM classifiers and then use the separate tuning set to choose the tuning parameter λ among the set $\{2^{-16}, 2^{-15}, \ldots, 2^{15}\}$. After the tuning parameter gets selected, we use the test set to evaluate the corresponding test error of the tuned RSVMs. To examine the effect of different choices of the function class, we use both linear kernel, $K(\mathbf{u},\mathbf{v})=\langle u, v\rangle$, and the polynomial kernel of order 2, $K(\mathbf{u},\mathbf{v})=(1+\langle u, v\rangle)^2$.

Table 1 reports the estimated test errors based on a test set of size 10^5 for $n=50$ and 100. The results show very interesting behaviors of RMSVMs with different γ's using different kernels. When we use the linear kernel, $\gamma=0$ gives the worst performance. Although the reinforced multicategory hinge loss is consistent when $\gamma=0$ and not consistent when $\gamma=1$, the corresponding linear RMSVM with $\gamma=1$ gives better accuracy for this example. When we change the linear kernel to a polynomial kernel, the RMSVMs with $\gamma\in[0,0.5]$ work better than $\gamma>0.5$. This seemingly surprising result reflects that Fisher consistency is only a pointwise consistency result. When the function class is relatively small, forcing f_y to be the maximum directly as in the first part of reinforced loss may work better than the second part which encourages only large f_y implicitly. As the function class becomes large, Fisher consistency becomes more relevant and the consistent RMSVMs with $\gamma\in[0,0.5]$ using the polynomial kernel perform better in this example. Figure 2 gives a clear visualization of the effect of γ. Furthermore, the left panel

Table 1. Estimated test errors and the corresponding estimated standard errors based on 100 replications for Example 4.1.1 based on the RSVMs with different γ, the MSVM by Weston and Watkins (1999) (WW-SVM), and the one-versus-rest approach (OVR). The estimated Bayes error is 0.2039.

γ	Lin $n=50$	Lin $n=100$	poly2 $n=50$	poly2 $n=100$
0	0.2948 (0.0075)	0.2428 (0.0041)	0.2359 (0.0025)	0.2214 (0.0013)
0.1	0.2760 (0.0063)	0.2342 (0.0031)	0.2357 (0.0024)	0.2207 (0.0011)
0.2	0.2618 (0.0053)	0.2263 (0.0022)	0.2368 (0.0026)	0.2211 (0.0012)
0.3	0.2469 (0.0040)	0.2214 (0.0017)	0.2366 (0.0023)	0.2219 (0.0013)
0.4	0.2415 (0.0035)	0.2186 (0.0014)	0.2371 (0.0025)	0.2216 (0.0011)
0.5	0.2359 (0.0027)	0.2161 (0.0011)	0.2381 (0.0024)	0.2227 (0.0011)
0.6	0.2315 (0.0023)	0.2153 (0.0009)	0.2406 (0.0024)	0.2230 (0.0011)
0.7	0.2298 (0.0023)	0.2154 (0.0011)	0.2399 (0.0026)	0.2236 (0.0011)
0.8	0.2307 (0.0031)	0.2151 (0.0010)	0.2448 (0.0026)	0.2249 (0.0012)
0.9	0.2325 (0.0030)	0.2168 (0.0012)	0.2557 (0.0036)	0.2325 (0.0019)
1	0.2419 (0.0033)	0.2242 (0.0016)	0.3051 (0.0056)	0.2639 (0.0032)
WW-SVM	0.2359 (0.0029)	0.2176 (0.0012)	0.2505 (0.0032)	0.2267 (0.0019)
OVR	0.2506 (0.0049)	0.2197 (0.0015)	0.2550 (0.0034)	0.2256 (0.0013)

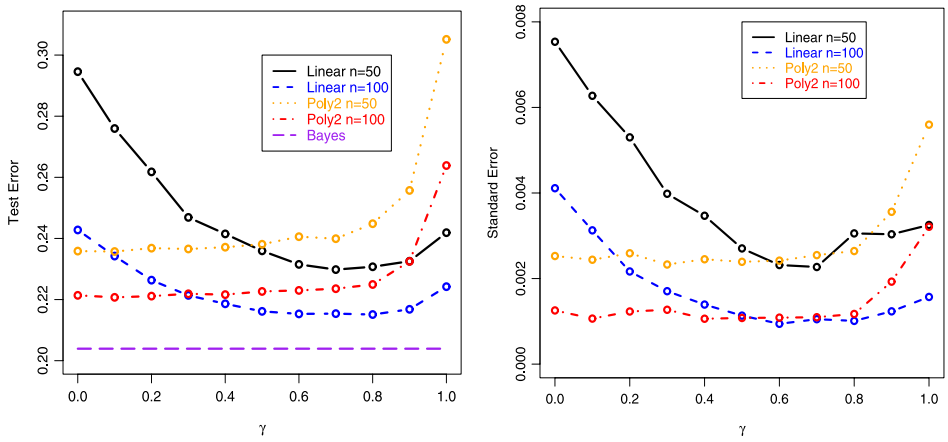

Figure 2. Left panel: Plot of the average estimated test errors of RMSVMs based on 100 replications with $\gamma = 0, 0.1, \ldots, 1$ for Example 4.1.1. Right panel: Plot of the corresponding standard errors of the estimated test errors of RMSVMs. The online version of this figure is in color.

of Figure 3 plots the classification boundaries of the RMSVMs with $\gamma = 0, 0.5, 1$ using the linear kernel. As shown in the plot, $\gamma = 0.5$ works remarkably well. Overall, the middle range values of γ such as 0.5 give the most accurate and stable classification performance.

For comparison, we also include the results by the MSVM approach proposed by Weston and Watkins (1999) (WW-SVM), and the one-versus-rest approach (OVR). Overall, the proposed RMSVM with $\gamma = 0.5$ delivers very competitive performance.

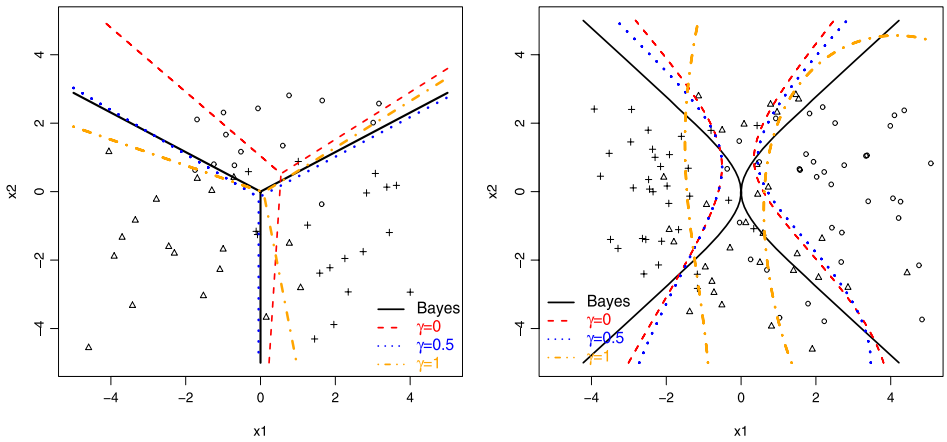

Figure 3. Plots of the typical classification boundaries of the RMSVMs with $\gamma = 0, 0.5, 1$. Left panel: Classification boundaries using linear kernel for Example 4.1.1 with $n = 50$. Right panel: Classification boundaries using Gaussian kernel for Example 4.1.2 with $n = 100$. The plots show that the RSVM with $\gamma = 0.5$ yields very accurate classification boundaries. The online version of this figure is in color.

4.1.2 Example With a Nonlinear Bayes Decision Boundary

In this three-class example, $P(Y=1)=P(Y=2)=P(Y=3)=1/3$, $P(X|Y=1)\sim N(\boldsymbol{\mu}=(2,0)^T, 1.5^2\mathbf{I}_2)$, $P(X|Y=2)\sim 0.5N(\boldsymbol{\mu}=(0,2)^T, 1.5^2\mathbf{I}_2)+0.5N(\boldsymbol{\mu}=(0,-2)^T, 1.5^2\mathbf{I}_2)$, and $P(X|Y=3)\sim N(\boldsymbol{\mu}=(-2,0)^T, 1.5^2\mathbf{I}_2)$. Due to the design of the underlying distribution, linear $\mathbf{f}$'s will not work well for this example. The nonlinear Bayes decision boundary is shown in the right panel of Figure 3.

We simulate data and build RMSVMs in the same manner as in Example 4.1.1. Since linear learning does not work for this example, to examine the effect of different choices of the function class, we use both the polynomial kernel of order 2 and the Gaussian kernel. The parameter λ is tuned in the same way as in Example 4.1.1. As to the second parameter σ of the Gaussian kernel, we use the median of the between-class pairwise Euclidean distances of training inputs to avoid extensive grid search (Brown et al. 2000; Wu and Liu 2007).

Table 2 reports the estimated test errors based on a test set of size 10^5 for $n=50$ and 100. The results show that RMSVMs with large γ's, that is, close to 1, give worse accuracy than those of smaller γ's. Overall, RMSVMs with $\gamma=0.5$ give the best or close to the best performance. In contrast to Example 4.1.1, RMSVMs with $\gamma=0$ give reasonable performance. This may be because the kernel space here is relatively more flexible than the linear kernel space. Figure 4 illustrates the effect of γ. From the plot, we can see that the values of $\gamma=0.5, 0.6, 0.7$ generally yield the best accuracy. Furthermore, the Gaussian kernel works consistently better than the polynomial kernel of order 2 in this example. Lastly, the right panel of Figure 3 indicates that $\gamma=0, 0.5$ give similar classification boundaries, much better than that of $\gamma=1$ for this example. Similarly to Example 4.1.1, we also include the results by the WW-SVM and OVR. Again, the proposed RMSVM with $\gamma=0.5$ is very competitive.

In summary, both the linear Example 4.1.1 and the nonlinear Example 4.1.2 indicate that γ around 0.5 works the best in terms of accuracy and stability. Combining the Fisher

Table 2. Estimated test errors based on 100 replications for Example 4.1.2 based on the RSVMs, WW-SVM, and OVR. The estimated Bayes error is 0.2883.

γ	poly2 $n=50$	poly2 $n=100$	Gauss $n=50$	Gauss $n=100$
0	0.3473 (0.0038)	0.3209 (0.0019)	0.3420 (0.0029)	0.3199 (0.0017)
0.1	0.3444 (0.0038)	0.3204 (0.0020)	0.3392 (0.0027)	0.3196 (0.0017)
0.2	0.3435 (0.0036)	0.3201 (0.0021)	0.3407 (0.0028)	0.3196 (0.0017)
0.3	0.3417 (0.0034)	0.3191 (0.0020)	0.3398 (0.0028)	0.3194 (0.0017)
0.4	0.3416 (0.0033)	0.3168 (0.0018)	0.3405 (0.0028)	0.3193 (0.0018)
0.5	0.3390 (0.0031)	0.3176 (0.0021)	0.3414 (0.0027)	0.3186 (0.0018)
0.6	0.3391 (0.0036)	0.3189 (0.0019)	0.3395 (0.0027)	0.3184 (0.0016)
0.7	0.3389 (0.0036)	0.3188 (0.0019)	0.3402 (0.0027)	0.3214 (0.0019)
0.8	0.3403 (0.0031)	0.3228 (0.0020)	0.3446 (0.0030)	0.3256 (0.0022)
0.9	0.3529 (0.0035)	0.3316 (0.0024)	0.3503 (0.0031)	0.3318 (0.0019)
1	0.3909 (0.0050)	0.3647 (0.0033)	0.3708 (0.0042)	0.3518 (0.0027)
WW-SVM	0.3429 (0.0040)	0.3223 (0.0026)	0.3525 (0.0037)	0.3283 (0.0024)
OVR	0.3500 (0.0042)	0.3306 (0.0027)	0.3488 (0.0037)	0.3235 (0.0023)

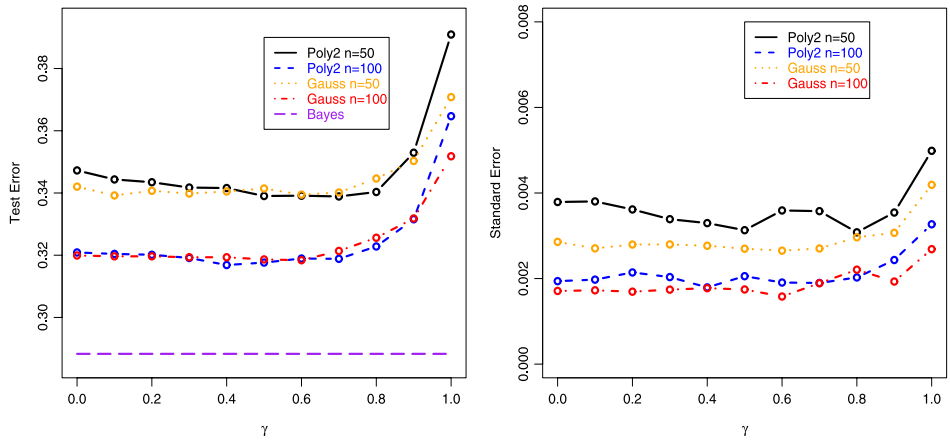

Figure 4. Left panel: Plot of the average estimated test errors of RMSVMs based on 100 replications with $\gamma = 0, 0.1, \ldots, 1$ for Example 4.1.2. Right panel: Plot of the corresponding standard errors of the estimated test errors of RMSVMs. The online version of this figure is in color.

consistency and the numerical results, we recommend the RMSVM with $\gamma = 0.5$ as the best choice.

4.2 AN APPLICATION TO LUNG CANCER MICROARRAY DATA

We use a real cancer dataset (available from *http://www.broad.mit.edu/mpr/lung/*) to demonstrate the effectiveness of the proposed RMSVMs. The dataset has been previously studied by Liu et al. (2008) and contains 2530 genes. Each gene is standardized to have sample mean 0 and standard deviation 1. There are four histological types, adenocarcinoma, pulmonary carcinoid tumors, squamous cell, and normal lung. Among the four, the first three are lung cancer subtypes.

The dataset contains 186 subjects including 128 adenocarcinoma, 20 carcinoid, 21 squamous, and 17 normal tissues. We apply the proposed RMSVMs with different values of γ on this cancer dataset. To that end, we split the sample of each class into two parts, one for model building and the other for testing. In particular, we randomly select 64, 5, 5, and 4 observations from the groups of adenocarcinoma, carcinoid, squamous, and normal tissues for testing, and the remaining for model building. As a result, for the purpose of testing, we have around 50% observations from the adenocarcinoma class as it is the majority class, and 25% observations from each of the other three classes. We use the model building part to build RMSVM classifiers. Since we have high-dimension low-sample-size data, we apply linear learning here, which appears to be sufficient. The 10-fold cross-validation is used to select the parameter λ. The testing data are used to evaluate and compare different methods. The process is repeated for 10 times.

We first carry out the four-class classification. The results are summarized in Table 3. It appears that RMSVMs with $\gamma \in [0, 0.5]$ work better than those with $\gamma > 0.5$. RMSVMs with $\gamma \leq 0.5$ give same results. To further compare the methods, we split the four-class problem into four three-class classification problems. The results are included in Table 3

Table 3. Averages and standard deviations of the test errors based on 10 replications for the lung cancer dataset. Classes adenocarcinoma, carcinoid, squamous, and normal are represented by A, C, S, and N, respectively.

Cases	$\gamma = 0$	$\gamma = 0.2$	$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 1$
A–C–S–N	0.0474 (0.012)	0.0474 (0.012)	0.0474 (0.012)	0.0500 (0.004)	0.0731 (0.012)
A–C–S	0.0297 (0.006)	0.0297 (0.006)	0.0270 (0.010)	0.0270 (0.010)	0.0324 (0.015)
A–C–N	0.0521 (0.013)	0.0521 (0.013)	0.0521 (0.013)	0.0616 (0.020)	0.0699 (0.031)
A–S–N	0.0521 (0.013)	0.0521 (0.013)	0.0507 (0.013)	0.0534 (0.019)	0.0493 (0.025)
C–S–N	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

as well. Overall, the testing errors are relatively low. Although the differences are not significantly large and the dataset is small to claim, the RMSVM with $\gamma = 0.5$ appears to work well in all cases.

One interesting point one can see from Table 3 is that we have perfect classification on the three-class case: carcinoid, squamous, and normal subtypes. We do not have perfect classification when the class adenocarcinoma is involved. To understand this further, we project the data onto the first two principal component directions for visualization to get a better idea of the classification problem. From Figure 5, we can see that except for the group of adenocarcinoma, the other three groups are well separated. The adenocarcinoma group overlaps with squamous and normal groups. This is expected in view of the previous knowledge that adenocarcinoma is a relatively heterogeneous lung cancer subtype (Bhattacharjee et al. 2001). Furthermore, although this dataset is very unbalanced with adenocarcinoma as the majority class, it does not suffer the typical difficulty of unbalanced classification. In particular, as pointed out by Qiao and Liu (2009), if the overall classification accuracy is used as the classification criterion, minority classes can be ignored since the classifiers tend to focus on the majority classes. Interestingly, this is not an issue for

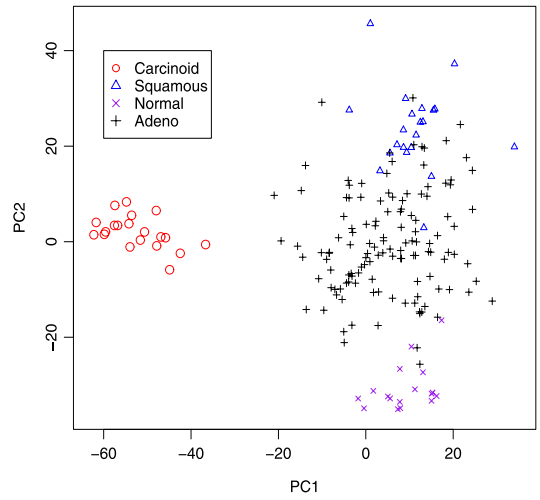

Figure 5. PCA projection plot of the lung cancer Microarray data. The online version of this figure is in color.

our example since the class of adenocarcinoma is the most difficult one to classify. The reported classification errors are mostly due to adenocarcinoma.

5. DISCUSSION

In the literature, there exist a number of different multicategory SVMs. To our knowledge, most of them are not Fisher consistent. In this article, we propose the new family of reinforced multicategory hinge loss functions. Our proposed RMSVMs include the MSVM by Lee, Lin, and Wahba (2004) as a special case, and also cover many new Fisher consistent multicategory hinge loss functions. Furthermore, the new family has some interesting connections with the one-versus-rest approach.

Our theoretical investigation and numerical studies indicate that $\gamma = 0.5$ appears to work very well. Although we do not expect the RMSVM to always outperform other existing MSVMs in real applications in view of the previous numerical study by Rifkin and Klautau (2004), Hill and Doucet (2007), we believe that the RMSVM provides a promising and useful addition to the SVM toolkit.

In comparison with the one-versus-rest method, all-at-once methods such as the RMSVM can be more expensive to compute. For binary SVMs, Platt (1999) proposed Sequential Minimal Optimization (SMO) to simplify the computation. Hill and Doucet (2007) extended the use of SMO for multicategory SVMs. One possible approach to reduce the computational cost of the RMSVM is to adopt SMO. Another approach to improve computation is to develop efficient solution-path algorithms for the RMSVM (Hastie et al. 2004; Wang and Shen 2006).

To implement the reinforced SVM in practice, one needs to choose the tuning parameter λ in (2.1). Similarly to many other regularization methods, the tuning parameter λ is important for the effectiveness of the proposed technique. Although one can use certain cross-validation procedures in practice when a separate tuning dataset is not available, the computational cost can be high. Therefore, an easy-to-compute data-dependent tuning criterion is desirable. Wahba, Lin, and Zhang (2000) developed the generalized approximate cross-validation (GACV) procedure for efficient tuning parameter selection of the SVM. It will be interesting to generalize the GACV procedures for our RMSVMs.

Another research direction of RMSVMs is the convergence properties. A number of articles on the convergence of large-margin classifiers have appeared in the literature. To list a few, Shen et al. (2003) provided learning theory for ψ-learning. Tarigan and van de Geer (2004), Wang and Shen (2007) derived rates of convergence for the L_1 SVMs. Steinwart and Scovel (2006) studied the convergence rate of the SVM using Gaussian kernels. Recently, Shen and Wang (2007) studied rates of convergence of the generalization error of a class of multicategory margin classifiers. It will be interesting to explore the effect of γ on the RMSVM in terms of its asymptotic convergence behavior.

APPENDIX: PROOFS OF THE FISHER CONSISTENCY RESULTS

In this section, we give the proofs of Theorems 1 and 2. We begin by discussing Fisher consistency of the two extremes: (I) $\gamma = 1$ and (II) $\gamma = 0$.

Lemma A.1. *Assume that* $\operatorname{argmin}_j P_j(\mathbf{x})$ *is uniquely determined. The minimizer* $\mathbf{f}^*$ *of* $E[[(k-1) - f_Y(\mathbf{X})]_+]$ *subject to* $\sum_j^k f_j(\mathbf{x}) = 0$ *satisfies the following*: $f_j^*(\mathbf{x}) = -(k-1)^2$ *if* $j = \operatorname{argmin}_j P_j(\mathbf{x})$ *and* $k-1$ *otherwise.*

From Lemma A.1, we can see that the loss (I), $[(k-1) - f_y(\mathbf{x})]_+$ subject to $\sum_j^k f_j(\mathbf{x}) = 0$, is not consistent since except the smallest element, all the remaining elements of its minimizer are $k-1$. Consequently, the argmax rule cannot be uniquely determined and thus the loss is not consistent.

Lemma A.2. *Assume that* $\operatorname{argmax}_j P_j(\mathbf{x})$ *is uniquely determined. The minimizer* $\mathbf{f}^*$ *of* $E[\sum_{j\neq Y}[1 + f_j(\mathbf{X})]_+]$ *subject to* $\sum_j^k f_j(\mathbf{x}) = 0$ *satisfies the following*: $f_j^*(\mathbf{x}) = k-1$ *if* $j = \operatorname{argmax}_j P_j(\mathbf{x})$ *and* -1 *otherwise.*

Lemma A.2 implies that the loss (II), $\sum_{j\neq y}[1 + f_j(\mathbf{x})]_+$ subject to $\sum_j^k f_j(\mathbf{x}) = 0$, is a consistent loss since its minimizer yields the Bayes decision boundary. A similar result was also established by Lee, Lin, and Wahba (2004).

We now prove Lemmas A.1 and A.2 and then show several additional lemmas.

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

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

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

$$\max_{\mathbf{f}} \sum_{l=1}^k P_l(\mathbf{x}) f_l(\mathbf{x}),$$

$$\text{subject to} \quad \sum_{l=1}^k f_l(\mathbf{x}) = 0; \qquad f_l(\mathbf{x}) \leq (k-1)\ \forall l.$$

It is easy to see that the solution satisfies $f_j^*(\mathbf{x}) = -(k-1)^2$ if $j = \operatorname{argmin}_j P_j(\mathbf{x})$ and $k-1$ otherwise. □

Proof of Lemma A.2: Note that $E[\sum_{j\neq Y}[1 + f_j(\mathbf{X})]_+] = E[E(\sum_{j\neq Y}[1 + f_j(\mathbf{x})]_+ | \mathbf{X} = \mathbf{x})]$. Thus, it is sufficient to consider the minimizer for a given $\mathbf{x}$ and $E(\sum_{j\neq Y}[1 + f_j(\mathbf{x})]_+ | \mathbf{X} = \mathbf{x}) = \sum_{l=1}^k \sum_{j\neq l}[1 + f_j(\mathbf{x})]_+ P_l(\mathbf{x})$.

Next, we show the minimizer $\mathbf{f}^*$ satisfies $f_j^* \geq -1$ for $\forall j = 1, \ldots, k$. To show this, suppose a solution $\mathbf{f}^1$ having $f_j^1 < -1$. Then we can construct another solution $\mathbf{f}^2$ 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 $f_l^2 < f_l^1; \forall l \neq j$. Consequently, $\sum_{l=1}^k \sum_{j\neq l}[1 + f_j^2]_+ P_l < \sum_{l=1}^k \sum_{j\neq l}[1 + f_j^1]_+ P_l$. This implies that $\mathbf{f}^1$ cannot be the minimizer. Therefore, the minimizer $\mathbf{f}^*$ satisfies $f_j^* \geq -1$ for $\forall j$.

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

$$\max_{\mathbf{f}} \sum_{l=1}^k P_l(\mathbf{x}) f_l(\mathbf{x}),$$

$$\text{subject to} \quad \sum_{l=1}^k f_l(\mathbf{x}) = 0; \qquad f_l(\mathbf{x}) \geq -1 \ \forall l.$$

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

Without loss of generality, assume that $P_1 > P_2 \geq \cdots \geq P_k$. The proof of Theorem 1 can be decomposed into the following steps.

Lemma A.3. *If $P_1 \geq P_2 \geq \cdots \geq P_k$, then $f_1^* \geq \cdots \geq f_k^*$.*

Proof: Note that

$$E\big(V(\mathbf{f}(\mathbf{X}), Y)|\mathbf{x}\big) = \sum_l P_l\{\gamma[k - 1 - f_l]_+ - (1-\gamma)[1 + f_l]_+\} + (1-\gamma)\sum_l [1 + f_l]_+. \tag{A.1}$$

Denote

$$h(u) \equiv \gamma[k - 1 - u]_+ - (1-\gamma)[1 + u]_+.$$

Minimizing $E(V(\mathbf{f}(\mathbf{X}), Y)|\mathbf{x})$ would ensure that $h(f_1) \leq \cdots \leq h(f_k)$. Otherwise, assume that $h(f_j) > h(f_i)$ but $j < i$. Define $f_l^1 = f_l$ for $l \neq i, j$ and $f_j^1 = f_i$, $f_i^1 = f_j$. Clearly, $E(V(\mathbf{f}(\mathbf{X}), Y)|\mathbf{x}) > E(V(\mathbf{f}^1(\mathbf{X}), Y)|\mathbf{x})$, which is contradictory. Note that $h(\cdot)$ is a monotonically decreasing function. This implies that $f_1^* \geq \cdots \geq f_k^*$. □

Lemma A.4. *If $P_1 \geq P_2 \geq \cdots \geq P_k$, then $f_1^* \leq k - 1$.*

Proof: From Lemma A.3, we know that $f_1^* \geq \cdots \geq f_k^*$. Assume the contrary that $f_1^* > k - 1$. To ensure that $\sum f_l^* = 0$, we need $f_k^* < -1$. Define $f_l^1 = f_l$ for $1 < l < k$ and

$$f_1^1 = f_1 - \epsilon, \qquad f_k^1 = f_k + \epsilon,$$

where $\epsilon > 0$ is such that $f_1^1 > k-1$ and $f_k^1 < -1$. Note that

$$E\big(V(\mathbf{f}^1(\mathbf{X}), Y)|\mathbf{x}\big) - E\big(V(\mathbf{f}^*(\mathbf{X}), Y)|\mathbf{x}\big) = -(1-P_1)(1-\gamma)\epsilon - P_k\gamma\epsilon < 0$$

which contradicts the fact that $\mathbf{f}^*$ is the minimizer of $E(V(\mathbf{f}(\mathbf{X}), Y)|\mathbf{x})$. □

Lemma A.5. *If $P_1 \geq \cdots \geq P_k$, then $f_k^* \geq -1$ if $\gamma < (1-P_2)/(1-P_k)$.*

Proof: From Lemma A.3, we know that $f_1^* \geq \cdots \geq f_k^*$. By Lemma A.4, $f_1^* \leq k-1$. Assume the contrary that $f_k^* < -1$. Then $-1 < f_2^* \leq k-1$ due to the sum-to-zero constraint. Define $f_l^1 = f_l^*$ if $l \neq 2, k$ and

$$f_2^1 = f_2^* - \epsilon, \qquad f_k^1 = f_k^* + \epsilon,$$

where $\epsilon > 0$ is such that $f_k^1 < -1$. Then

$$E\big(V(\mathbf{f}^1(\mathbf{X}), Y)|\mathbf{x}\big) - E\big(V(\mathbf{f}^*(\mathbf{X}), Y)|\mathbf{x}\big) = [P_2 - 1 + (1-P_k)\gamma]\epsilon,$$

which is negative if and only if $P_2 - 1 + (1-P_k)\gamma < 0$, or equivalently,

$$\gamma < \frac{1-P_2}{1-P_k}.$$ □

Proof of Theorem 1: From Lemma A.5, we know that for $P_1 \geq \cdots \geq P_k$, $f_k^* \geq -1$ if $\gamma < (1-P_2)/(1-P_k)$. Note that $(1-P_2)/(1-P_k)$ is a decreasing function of P_2 and an increasing function of P_k. Thus, its lower bound is $1/2$ which corresponds to $P_2 = 1/2$ and $P_k = 0$. To show Fisher consistency, we can assume $P_1 > P_2$. Thus, we can conclude that $f_k^* \geq -1$ if $\gamma \leq 1/2 < (1-P_2)/(1-P_k)$. Since $\gamma \leq 1/2$, $f_j^* \geq -1$ for $\forall j$. Using an argument similar to that of the proof of Lemma A.2, we can get $f_j^*(\mathbf{x}) = k-1$ if $j = \text{argmax}_j P_j(\mathbf{x})$ and -1 otherwise. The desired result then follows. □

Proof of Theorem 2: Consider a simple case where $k = 3$ and $P_3 = 0$. Then (A.1) becomes

$$\{\gamma P_1[2-f_1]_+ + P_2(1-\gamma)[1+f_1]_+\}$$
$$+\{\gamma P_2[2-f_2]_+ + P_1(1-\gamma)[1+f_2]_+\} + (1-\gamma)[1+f_3]_+.$$

If $1/2 < P_1 < \gamma$, then $\gamma P_1 > P_2(1-\gamma)$ and $\gamma P_2 > P_1(1-\gamma)$, consequently the minimizer is $(2, 2, -4)$. This implies that $V(\mathbf{f}(\mathbf{x}), y)$ is not Fisher consistent. □

SUPPLEMENTARY MATERIALS

R archive for the proposed reinforced multicategory support vector machines: The archive contains R code implementing the proposed methods. (code.zip)

ACKNOWLEDGMENTS

The authors are indebted to the editor, the associate editor, and two referees, whose helpful comments and suggestions led to a much improved presentation. Liu's research was supported in part by NSF grants DMS-0606577 and DMS-0747575, and NIH grant 1R01CA149569-01. Yuan's research was supported in part by NSF grant DMS-0846234 and a grant from Georgia Cancer Coalition.

[Received November 2009. Revised August 2010.]

REFERENCES

Allwein, E. L., Schapire, R. E., Singer, Y., and Kaelbling, P. (2000), "Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers," *Journal of Machine Learning Research*, 1, 113–141. [902]

Bartlett, P., Jordan, M., and McAuliffe, J. (2006), "Convexity, Classification, and Risk Bounds," *Journal of the American Statistical Association*, 101, 138–156. [902,904]

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., and Meyerson, M. (2001), "Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinoma Subclasses," *Proceedings of the National Academy Science USA*, 13790–13795. [913]

Boser, B., Guyon, I., and Vapnik, V. N. (1992), "A Training Algorithm for Optimal Margin Classifiers," in *COLT'92 Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, eds. D. Haussler, New York: ACM, pp. 144–152. [901]

Bredensteiner, E., and Bennett, K. (1999), "Multicategory Classification by Support Vector Machines," *Computational Optimizations and Applications*, 12, 53–79. [903]

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., and Haussler, D. (2000), "Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines," *The Proceeding of National Academy of Sciences*, 97, 262–267. [911]

Cortes, C., and Vapnik, V. N. (1995), "Support-Vector Networks," *Machine Learning*, 20, 273–279. [901]

Crammer, K., and Singer, Y. (2001), "On the Algorithmic Implementation of Multiclass Kernel-Based Vector Machines," *Journal of Machine Learning Research*, 2, 265–292. [902,903]

Cristianini, N., and Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge, U.K.: Cambridge University Press. [901]

Dietterich, T. G., and Bakiri, G. (1995), "Solving Multiclass Learning Problems via Error-Correcting Output Codes," *Journal of Artificial Intelligence Research*, 2, 263–286. [902]

Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004), "The Entire Regularization Path for the Support Vector Machine," *Journal of Machine Learning Research*, 5, 1391–1415. [914]

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer-Verlag. [901]

Hill, S. I., and Doucet, A. (2007), "A Framework for Kernel-Based Multi-Category Classication," *Journal of Articial Intelligence Research*, 30, 525–564. [904,914]

Kimeldorf, G., and Wahba, G. (1971), "Some Results on Tchebycheffian Spline Functions," *Journal of Mathematical Analysis and Applications*, 33, 82–95. [906]

Lee, Y., Lin, Y., and Wahba, G. (2004), "Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data," *Journal of the American Statistical Association*, 99, 67–81. [902-904,908,914,915]

Lin, Y. (2004), "A Note on Margin-Based Loss Functions in Classification," *Statistics and Probability Letters*, 68, 73–82. [901,903]

Liu, Y. (2007), "Fisher Consistency of Multicategory Support Vector Machines," in *Proceedings of the Eleventh International Workshop on Artificial Intelligence and Statistics*, Madison, WI: Omnipress, pp. 289–296. [902, 905]

Liu, Y., and Shen, X. (2006), "Multicategory ψ-Learning," *Journal of the American Statistical Association*, 101, 500–509. [902-904]

Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. S. (2008), "Statistical Significance of Clustering for High Dimension Low Sample Size Data," *Journal of the American Statistical Association*, 103, 1281–1293. [912]

Platt, J. C. (1999), "Fast Training of Support Vector Machines Using Sequential Minimal Optimization," in *Advances in Kernel Methods—Support Vector Learning*, eds. B. Schölkopf, C. J. C. Burges, and A. J. Smola, Cambridge, MA: MIT Press, pp. 185–208. [914]

Qiao, X., and Liu, Y. (2009), "Adaptive Weighted Learning for Unbalanced Multicategory Classification," *Biometrics*, 65, 159–168. [913]

Rifkin, R. M., and Klautau, A. (2004), "In Defense of One-vs-All Classification," *Journal of Machine Learning Research*, 5, 101–141. [902,914]

Schölkopf, B., and Smola, A. J. (2002), *Learning With Kernels*, Cambridge, MA: MIT Press. [901]

Shen, X., and Wang, L. (2007), "Generalization Error for Multi-Class Margin Classification," *Electronic Journal of Statistics*, 1, 307–330. [914]

Shen, X., Tseng, G. C., Zhang, X., and Wong, W. H. (2003), "On ψ-Learning," *Journal of the American Statistical Association*, 98, 724–734. [914]

Steinwart, I., and Scovel, C. (2006), "Fast Rates for Support Vector Machines Using Gaussian Kernels," *The Annals of Statistics*, 35, 575–607. [914]

Tarigan, B., and van de Geer, S. A. (2004), "Adaptivity of Support Vector Machines With L_1 Penalty," Technical Report MI 2004-14, University of Leiden. [914]

Tewari, A., and Bartlett, P. (2007), "On the Consistency of Multiclass Classification Methods," *Journal of Machine Learning Research*, 8, 1007–1025. [904]

Vapnik, V. (1998), *Statistical Learning Theory*, New York: Wiley. [902,903]

Wahba, G. (1999), "Support Vector Machines, Reproducing Kernel Hilbert Spaces, and Randomized GACV," in *Advances in Kernel Methods: Support Vector Learning*, eds. B. Schölkopf, C. J. C. Burges, and A. J. Smola, Cambridge, MA: MIT Press, pp. 125–143. [903,906]

Wahba, G., Lin, Y., and Zhang, H. H. (2000), "Generalized Approximate Cross Validation for Support Vector Machines, or, Another Way to Look at Margin-Like Quantities," in *Advances in Large Margin Classifiers*, eds. P. Bartlett, B. Scholkopf, D. Schuurmans, and A. Smola, Cambridge, MA: MIT Press, pp. 297–309. [914]

Wang, L., and Shen, X. (2006), "Multicategory Support Vector Machines, Feature Selection and Solution Path," *Statistica Sinica*, 16, 617–634. [914]

——— (2007), "On L_1-Norm Multiclass Support Vector Machines: Methodology and Theory," *Journal of the American Statistical Association*, 102, 583–594. [914]

Weston, J. (1999), "Extensions to the Support Vector Method," Ph.D. thesis, Royal Holloway University of London. [904]

Weston, J., and Watkins, C. (1999), "Support Vector Machines for Multi-Class Pattern Recognition," in *Proceedings of the 7th European Symposium on Artificial Neural Networks (ESANN-99)*, ed. M. Verleysen, Bruges, Belgium, pp. 219–224. [902,903,909,910]

Wu, Y., and Liu, Y. (2007), "Robust Truncated-Hinge-Loss Support Vector Machines," *Journal of the American Statistical Association*, 102, 974–983. [911]

Zhang, T. (2004a), "Statistical Analysis of Some Multi-Category Large Margin Classification Methods," *Journal of Machine Learning Research*, 5, 1225–1251. [904]

——— (2004b), "Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization," *The Annals of Statistics*, 32, 56–85. [902]