On L_1-norm multi-class support vector machines: methodology and theory*

Lifeng Wang and Xiaotong Shen
School of Statistics, University of Minnesota

Summary

Binary Support Vector Machines have proven to deliver high performance. In multi-class classification, however, issues remain with respect to variable selection. One challenging issue is classification and variable selection in presence of a large number of variables in the magnitude of thousands, which greatly exceeds the size of training sample. This often occurs in genomics classification. To meet the challenge, this article proposes a novel multi-class support vector machine, which performs classification and variable selection simultaneously through an L_1-norm penalized sparse representation. The proposed methodology, together with the developed regularization solution path, permits variable selection in such a situation. For the proposed methodology, a statistical learning theory is developed to quantify the generalization error to understand the basic structure of sparse learning, permitting the number of variables greatly exceeding the sample size. The operating characteristics of the methodology are examined via both simulated and benchmark data, and are compared against some competitors in terms of accuracy of prediction. The numerical results suggest that the proposed methodology is highly competitive.

* This research was supported in part by National Science Foundation Grant IIS-0328802. The authors would like to thank the editor, the associate editor and three anonymous referees for helpful comments and suggestions.
Key words: High-dimension and low sample size, Margin classification, Regularization, Sparsity, Variable selection.

1. Introduction

Support vector machines (SVMs, c.f., Vapnik 1998), as classification tools, have proven effective in achieving the state-of-the-art performance in a variety of applications. For variable selection, Bradley and Mangasarian (1998) introduced a SVM with an L_1-norm penalty, which can perform variable selection and classification simultaneously. In multi-class classification, challenges remain, particularly for high-dimension and low sample size data. In this article, we develop a new L_1-norm multi-class SVM (L1MSVM) and investigate its feasibility in classification and variable selection.

In multi-class classification, a common treatment is the "one-versus-all" approach (OVA), which performs a sequence of binary classifications to distinguish one class from the remaining classes. Through OVA, Szedmak, Shawe-Taylor, Saunders and

Hardoon (2004) generalized the result of Bradley and Mangasarian (1998). However, this generalization has several potential difficulties. First, it trains binary decision classifiers sequentially. When the number of classes is large, each binary classification becomes highly unbalanced, with a small fraction of instances in one class. As a result, the class with a smaller fraction of instances tends to be ignored in nonseparable cases, degrading generalization performance. Second, OVA treats one variable to be relevant for all classes if it is selected in one binary classification. Consequently, unremoved redundant variables in one binary classification remain in the final classification model, yielding worse generalization performance, c.f., the simulation result in Section 4.1.1. Third, OVA may break down in absence of a dominating class, c.f., Lee, Lin and Wahba (2004), especially so when the number of classes becomes large.

To overcome these difficulties, we propose L1MSVM to incorporate variable selection in the framework of classification by treating multiple classes jointly, as opposed to OVA. L1MSVM circumvents the difficulties of OVA, and generalizes the concept of margins. It is capable of performing variable selection and classification simultaneously, while retaining the geometric interpretability of the margin of its L_2-norm counterpart. Moreover, because dimension reduction is built into classification, it bypasses the requirement of an ad hoc step of dimension reduction to attack large problems beyond the capability of conventional techniques. This occurs in cancer genomics classification, where gene pre-screening is required, c.f., Dudoit, Fridlyand and Speed (2002).

To gain an insight into L1MSVM in variable selection and classification, we develop a new statistical learning theory to quantify its generalization accuracy in the number of variables p, the sample size n and its tuning parameter s. Our theory reveals two important aspects of L1MSVM with regard to sparse learning. First, the L_1-penalty enables to control a model's complexity effectively even when p greatly exceeds the sample size n. This aspect may not be shared by its L_2-norm counterpart. As a result,

the generalization error rates of L1MSVM can be obtained as long as p grows at a speed no faster than $\exp(n)$. This is in contrast to the theory of function estimation and classification, where the effective dimension of an estimation problem needs to be no greater than n. Second, the joint distribution of the input/output pair plays an important role in classification. When the distribution possesses certain properties, a surprisingly sharp rate can be realized. In fact, our illustrative example shows that the convergence speed of the generalization error of L1MSVM can be arbitrarily fast, depending on the distribution. This is in contrast to the existing theory in classification, where only the n^{-1} rate is achieved by a large margin classifier in binary classification (c.f., Shen, Tseng, Zhang and Wong 2003). In conclusion, L1MSVM effectively battles the curse of dimensionality in classification, provided that a relative sparse representation of a decision function can be realized through the L_1-penalty.

This article is organized as follows. Section 2 briefly introduces the proposed methodology. Section 3 develops our learning theory. Section 4 presents some numerical results on both simulated and real data. Section 5 contains a discussion, and Section 6 is devoted to technical proofs.

2. Methodology

In k-class classification, input $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(p)}) \in \mathbb{R}^p$ is a vector of p variables, and output Y, coded as $\{1, 2, \ldots, k\}$, indicates class labeling. A decision function vector $\boldsymbol{f} = (f_1, \ldots, f_k)^T$ is introduced, with f_c representing class c; $c = 1, \ldots, k$, together with a classification rule $\Phi_{\boldsymbol{f}}(x) = \arg\max_c f_c(x)$ that assigns a new input vector x to class c with the highest value $f_c(x)$. To avoid redundancy in $\boldsymbol{f}$, a zero-sum constraint $\sum_{c=1}^k f_c = 0$ is enforced (c.f., Lee, Lin and Wahba 2004). Our goal is to seek $\boldsymbol{f}$ that has small generalization error $\mathrm{Err}(\boldsymbol{f}) = E(I[Y \neq \Phi_{\boldsymbol{f}}(X)])$, based on a training sample $\{\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)\}_{i=1}^n$, sampled from an unknown probability distribution P.

2.1 Motivation

For motivation, we begin our discussion with the binary L_1-norm SVM (L1SVM) with $Y \in \{-1, +1\}$. In this case, SVM uses an p-dimensional hyperplane $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$ as a decision function with the corresponding decision rule $\Phi(\boldsymbol{x}) = \text{sign}(f(\boldsymbol{x}))$. Bradley et al. (1998) proposed L1SVM in the form of

$$\min_{w,b} V(y_i f(\boldsymbol{x}_i)) + \lambda\|\boldsymbol{w}\|_1, \tag{1}$$

where $V(z) = [1-z]_+$ is the hinge loss (c.f., Wahba 1999), and $\|\boldsymbol{w}\|_1 = \sum_{j=1}^p |w_j|$ is the L_1-norm of $\boldsymbol{w}$. In the linear separable case, (1) can be thought of as maximizing the geometric margin $\frac{2}{\|\boldsymbol{w}\|_1}$, which is the L_∞-distance between two hyperplanes $\boldsymbol{w}^T\boldsymbol{x} + b = \pm 1$, defined as $\inf_{\boldsymbol{x},\boldsymbol{x}'}\{\|\boldsymbol{x} - \boldsymbol{x}'\|_\infty : \boldsymbol{w}^T\boldsymbol{x} + b = 1, \boldsymbol{w}^T\boldsymbol{x}' + b = -1\}$ with $\|\boldsymbol{x}\|_\infty = \max_{1\le j\le p} |x_j|$ the L_∞-norm.

To extend L1SVM to the multi-class case, we need to generalize the hinge loss as well as the L_1-penalty in the binary case. In the literature, there are several generalizations of the hinge loss in the context of the L_2-norm MSVM (L2MSVM). Vapnik (1998), Weston and Watkins (1998), Bredensteiner and Bennett (1999), and Guermuer (2002) proposed several versions of L2MSVM, which uses a generalized hinge loss

$$V(\boldsymbol{f}, \boldsymbol{z}_i) = \sum_{c\neq y_i} [1 - (f_{y_i}(\boldsymbol{x}_i) - f_c(\boldsymbol{x}_i))]_+; \tag{2}$$

Liu and Shen (2005) suggested

$$V(\boldsymbol{f}, \boldsymbol{z}_i) = [1 - \min_c (f_{y_i}(\boldsymbol{x}_i) - f_c(x_i))]_+; \tag{3}$$

Lee, Lin and Wahba (2004) proposed

$$V(\boldsymbol{f}, \boldsymbol{z}_i) = \sum_{c\neq y_i} [f_c(\boldsymbol{x}_i) + 1]_+. \tag{4}$$

The generalized hinge loss in (2)-(4) are constructed based on different principles, see Zhang (2004) for a detailed discussion. However these losses, as well as the 0-1 loss $l(\boldsymbol{f}, \boldsymbol{z}) = I[y \neq \arg\min_c f_c(\boldsymbol{x})]$, can be expressed in a unified fashion. Define

$\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), y)$ as $(f_y(\boldsymbol{x}) - f_1(\boldsymbol{x}), \ldots, f_y(\boldsymbol{x}) - f_{y-1}(\boldsymbol{x}), f_y(\boldsymbol{x}) - f_{y+1}(\boldsymbol{x}), \ldots, f_y(\boldsymbol{x}) - f_k(\boldsymbol{x}))$, which compares class y against the remaining classes. Then $V(\boldsymbol{f}, \boldsymbol{z})$ can be written as $h(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), y))$, with $h(\boldsymbol{u}) = \sum_{j=1}^{k-1}[1-u_j]_+$ for (2); $h(\boldsymbol{u}) = [1-\min_j u_j]_+$ for (3); $h(\boldsymbol{u}) = \sum_{j=1}^{k-1}[\frac{\sum_{c=1}^{k-1} u_c}{k} - u_j + 1]_+$ for (4); and $l(\boldsymbol{f}, \boldsymbol{z}) = h(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), y))$ with $h(\boldsymbol{u}) = I[\min_j u_j < 0]$. Figure 1 displays the 3D plots of these h functions in three-class classification.

Figure 1 about here.

As suggested in Figure 1, the generalized hinge losses (2)-(4) are upper envelops of the 0-1 loss. However, their risk minimizers with respect to a class of candidate functions may differ dramatically.

2.2 L1MSVM

In k-class linear classification, linear decision functions $f_c(\boldsymbol{x}) = \boldsymbol{w}_c^T \boldsymbol{x} + b_c$; $c = 1, \ldots, k$ are used, with $\boldsymbol{w}_c = (w_{c,1}, \ldots, w_{c,p})^T \in \mathbb{R}^p$ and $b_c \in \mathbb{R}^1$, subject to zero-sum constraints $\sum_{c=1}^k \boldsymbol{w}_c = \vec{0}$ and $\sum_{c=1}^k b_c = 0$. In k-class nonlinear classification, decision functions $f_c(x) = \sum_{j=1}^q w_{c,j} h_j(x) + b_c$; $c = 1, \ldots, k$ involve flexible representations through a basis $\{h_j(x)\}_{j=1}^q$. The representations reduce to the k-class linear case when $\boldsymbol{H} = (h_j(x_i))_{n \times q}$ is treated as the design matrix instead of $\boldsymbol{X} = (x_{ij})_{n \times p}$. For the purpose of variable selection, linear representations are used, particularly in the case of p greatly exceeding n, because non-linear representations are overspecified. The reader refers to Lee, Kim, Lee and Koo (2004) for a discussion of nonlinear component selection in MSVM. With $V(\boldsymbol{f}, \boldsymbol{z})$ representing a generalized hinge loss in (2)-(4), we propose L1MSVM,

$$\min_{\boldsymbol{w}_c, b_c; c=1,\ldots,k} n^{-1} \sum_{i=1}^n V(\boldsymbol{f}, \boldsymbol{z}_i), \text{ subject to } \sum_{c=1}^k \|\boldsymbol{w}_c\|_1 \le s, \sum_{c=1}^k \boldsymbol{w}_c = \vec{0}, \sum_{c=1}^k b_c = 0, \quad (5)$$

where $\boldsymbol{f} = (\boldsymbol{w}_1^T \boldsymbol{x} + b_1, \ldots, \boldsymbol{w}_c^T \boldsymbol{x} + b_c)^T$ is a vector of linear decision functions, $\sum_{c=1}^k \|\boldsymbol{w}_c\|_1 = \sum_{c=1}^k \sum_{j=1}^p |w_{c,j}|$ is an L_1-norm penalty, and s is a non-negative tuning parameter.

L1MSVM defined in (5) can be regarded as structural risk minimization, c.f., Vapnik (1998). Let $\mathcal{F}(p, s) = \{\boldsymbol{f} = (f_1, \ldots, f_k)^T : f_c(\boldsymbol{x}) = \sum_{j=1}^p w_{c,j} x^{(j)} + b_c; \sum_{c,j} |w_{c,j}| \le$

$s; c = 1, \ldots, k, \sum_c f_c = 0\}$, consisting of all the decision function vector $\boldsymbol{f}$'s satisfying the constraints in (5). Then, (5) is equivalent to

$$\hat{\boldsymbol{f}} = \arg \min_{\boldsymbol{f} \in \mathcal{F}(p,s)} n^{-1} \sum_{i=1}^{n} V(\boldsymbol{f}, \boldsymbol{z}_i). \tag{6}$$

The optimization in (5) is solvable through a linear program for each loss in (2)-(4). In the sequel, we shall only implement (4). Write $n^{-1} \sum_{i=1}^{n} V(\boldsymbol{f}, \boldsymbol{z}_i)$ as $l_1(\boldsymbol{W}, \boldsymbol{b})$, where $\boldsymbol{W}$ is a $k \times p$ matrix with $W_{cj} = w_{c,j}$; $c = 1, \ldots, k$, $j = 1, \ldots, p$, and $\boldsymbol{b} = (b_1, \ldots, b_k)^T$. For any given value of $s \leq t^* = \inf\{\sum_{c=1}^{k} \|\boldsymbol{w}_c^*\|_1 : l_1(\boldsymbol{W}^*, \boldsymbol{b}^*) = \inf_{(\boldsymbol{W},\boldsymbol{b})} l_1(\boldsymbol{W}, \boldsymbol{b})\}$, the optimization in (5) with loss (4) is equivalent to

$$\min_{w_{c,j}^+, w_{c,j}^-, b_c^+, b_c^-, \xi_{i,c}} \sum_{i,c\,:\, c \neq y_i} \xi_{i,c} \tag{7}$$

subject to $\sum_{1 \leq j \leq p} (w_{c,j}^+ - w_{c,j}^-) x_{ij} + (b_c^+ - b_c^-) + 1 \leq \xi_{i,c}$; $\sum_{1 \leq c \leq k, 1 \leq j \leq p} (w_{c,j}^+ + w_{c,j}^-) \leq s$; $\sum_{1 \leq c \leq k} (w_{c,j}^+ - w_{c,j}^-) = 0$; $\sum_{1 \leq c \leq k} (b_c^+ - b_c^-) = 0$; $w_{c,j}^+, w_{c,j}^-, b_c^+, b_c^-, \xi_{i,c} \geq 0$; $c = 1, \ldots, k$; $j = 1, \ldots, p$; $(i, c) \in \{(i, c) : c \neq y_i\}$. The solution of (7), denoted as $(\hat{\boldsymbol{W}}^+, \hat{\boldsymbol{W}}^-, \hat{\boldsymbol{b}}^+, \hat{\boldsymbol{b}}^-)$, yields that of (5), $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{b}}) = (\hat{\boldsymbol{W}}^+, \hat{\boldsymbol{b}}^+) - (\hat{\boldsymbol{W}}^-, \hat{\boldsymbol{b}}^-)$, and the corresponding decision function vector $\hat{f}(\boldsymbol{x}) = (\hat{\boldsymbol{w}}_1{}^T \boldsymbol{x} + \hat{b}_1, \ldots, \hat{\boldsymbol{w}}_k{}^T \boldsymbol{x} + \hat{b}_k)$. For $s > t^*$, the solution may not be unique. To overcome this difficulty, define $(\hat{\boldsymbol{W}}(s), \hat{\boldsymbol{b}}(s)) = (\hat{\boldsymbol{W}}(t^*), \hat{\boldsymbol{b}}(t^*))$ to yield a unique solution for all $s > t^*$, because $s = t^*$ yields the global minimal of $l_1(\boldsymbol{W}, \boldsymbol{b})$.

To derive an unconstrained version of (5), let $l_2(\boldsymbol{W}, \boldsymbol{b}, \lambda) = n^{-1} \sum_{i=1}^{n} V(\boldsymbol{f}, \boldsymbol{z}_i) + \lambda(\sum_{c=1}^{k} \|\boldsymbol{w}_c\|_1 - s)$ with $(\boldsymbol{W}, \boldsymbol{b}) \in \mathcal{S} = \{(\boldsymbol{W}, \boldsymbol{b}) : \sum_c \boldsymbol{w}_c = \vec{0}; \sum_c b_c = 0\}$, and let $\lambda \geq 0$ be a Lagrangian multiplier. It then follows from the strong duality theory (c.f., Cristianini and Shaw-Taylor 2000) that (5) is equivalent to

$$\min_{\boldsymbol{W}, \boldsymbol{b}} \sum_{i=1}^{n} V(\boldsymbol{f}, \boldsymbol{z}_i) + \lambda \sum_{c=1}^{k} \|\boldsymbol{w}_c\|_1, \quad \text{s.t.} \quad \sum_c \boldsymbol{w}_c = \vec{0}; \sum_c b_c = 0, \tag{8}$$

for a choice of $\lambda = \arg\min_{\lambda \geq 0}(\inf_{(\boldsymbol{W},\boldsymbol{b}) \in \mathcal{S}} l_2(\boldsymbol{W}, \boldsymbol{b}, \lambda))$, which is now cast into the framework of regularization with a regularization parameter λ and a penalty $\sum_{c=1}^{k} \|\boldsymbol{w}_c\|_1$.

In (8), $\|\boldsymbol{w}_c\|_1$ can be interpreted as the reciprocal of the geometric margin $m_c =$

$\inf\{\|\boldsymbol{x} - \boldsymbol{x}'\|_\infty : f_c(\boldsymbol{x}) = 0, f_c(\boldsymbol{x}') + 1 = 0\}$, which is the L_∞-distance between two hyperplanes $f_c = 0$ and $f_c + 1 = 0$. Here m_c measures separation of class c from the rest classes. Figure 2 displays m_c in three-class classification.

Figure 2 about here.

In our formulation, we use $\sum_{c=1}^k \|\boldsymbol{w}_c\|_1$ instead of $\max_{1\le c\le k} \|\boldsymbol{w}_c\|_1$, because $\sum_{c=1}^k \|\boldsymbol{w}_c\|_1$ plays a similar role as $\max_{1\le c\le k} \|\boldsymbol{w}_c\|_1$ and is easier to work with.

The variable selection aspect of L1MSVM is useful for classification with p greatly exceeding n. Here the L_1-penalty shrinks the estimated coefficients and coerces some small coefficients to be exactly zero. Therefore, for sufficiently small s, many estimated coefficients $\hat{w}_{c,j}$ become exactly zero, which enables L1MSVM to perform variable selection within the framework of classification. The following lemma says that the number of variables selected by L1MSVM never exceeds $(n-1)(k-1)$. This is in contrast to OVA, in which the maximum number of variables allowed is nk.

Lemma 1 *The number of variables with non-zero coefficients in the solution matrix* $\hat{W}$ *is no more than* $(n-1)(k-1)$.

2.3 Tuning and computational issues

The key to the performance of L1MSVM is the choice of tuning parameter s, which controls the trade-off between training and generalization and determines the number of variables used in classification. To obtain a classifier with high generalization accuracy, adaptive selection of s is often performed by minimizing a model selection routine such as cross-validation with respect to s to yield the best performance. In this process, L1MSVM in (5) or (8) solves a linear program of dimension $2(p+1)k + n(k-1)$ repeatedly for each fixed s, which is computationally intensive in addition to a memory concern when p exceeds thousands.

Motivated by Zhu et. al (2004) and Hastie et. al (2004), an algorithm was developed in Wang and Shen (2006), constructing the entire piecewise linear path of $\hat{w}_{c,j}$ by sequentially identifying the joints on it, and therefore computing (5) for all possible

values of s simultaneously. We briefly describe the algorithm.

Step 1: Initialize $(\hat{\boldsymbol{W}}(s), \hat{\boldsymbol{b}}(s))$ at $s = 0$.

Step 2: At the l-th joint s_l, compute the right derivative $\boldsymbol{D}(s)$ of $(\hat{\boldsymbol{W}}(s), \hat{\boldsymbol{b}}(s))$.

Step 3: Given the current right derivative $\boldsymbol{D}(s)$, compute the next joint s_{l+1}.

Step 4: Iterate Steps 2 and 3 until the algorithm terminates.

This algorithm permits rapid computation of adaptive selection of s, and alleviates the memory requirement. This is because L1MSVM selects no more than $(n-1)(k-1)$ variables, and hence that at most $(n-1)(k-1)$ variables are required to be stored in computing, which makes computation of high-dimensional problems feasible.

3 Statistical learning theory

This section derives a novel learning theory for L1MSVM defined in (6) in its generalization error in high-dimensional linear classification, where p is allowed to grow with n at a speed no faster than $\exp(n)$. In the literature, Tarigan and van der Geer (2004) derived rates of convergence for the binary L_1-penalty SVM when $p < n$.

3.1 Framework

First we introduce some notations. Write $\boldsymbol{X}(p) = (X^{(1)}, \ldots, X^{(p)})^T$ as a truncated infinite-dimensional random vector $\boldsymbol{X} = (X^{(1)}, X^{(2)}, \ldots)^T$. For simplicity, assume that $X^{(j)} \in [0,1]$, and define $\mathcal{X} = [0,1]^\infty$, $\mathcal{Y} = \{1, \ldots, k\}$, although our theory is applicable when $X^{(j)} \in [-B, B]$ for $B < \infty$. Let $\boldsymbol{f}^{(p)} = \arg\inf_{\boldsymbol{f} \in \mathcal{F}(p)} EV(\boldsymbol{f}, \boldsymbol{Z})$, where $\mathcal{F}(p) = \cup_{0 \le s < \infty} \mathcal{F}(p, s)$ is the full p-dimensional model, and $\boldsymbol{f}^{(p)}$ may not belong to $\mathcal{F}(p)$.

For any $\boldsymbol{f} \in \mathcal{F}(p)$, its performance is measured in three risks: The first is the excess hinge risk $e_V(\boldsymbol{f}, \boldsymbol{f}^{(p)}) = EV(\boldsymbol{f}, \boldsymbol{Z}) - EV(\boldsymbol{f}^{(p)}, \boldsymbol{Z}) \ge 0$ representing the performance of $\boldsymbol{f}$ under the hinge loss V; the second is the generalization error $|e(\boldsymbol{f}, \boldsymbol{f}^{(p)})| = |El(\boldsymbol{f}, \boldsymbol{Z}) - El(\boldsymbol{f}^{(p)}, \boldsymbol{Z})|$ under the the 0-1 loss when V is the surrogate loss for the 0-1 loss in classification; the third is $|e(\boldsymbol{f}, \boldsymbol{f}^{(\infty)})|$ with $\boldsymbol{f}^{(\infty)} = \arg\inf_{\boldsymbol{f} \in \mathcal{F}(\infty)} EV(\boldsymbol{f}, \boldsymbol{Z})$ representing the optimal decision function vector $\boldsymbol{f}^{(\infty)}$ over the infinite-dimensional space under the

0-1 loss, which can be thought of as the limit of $\boldsymbol{f}^{(p)}$ when $p \to \infty$. Here, the expectation is taken with respect to P, the distribution of $\boldsymbol{Z} = (\boldsymbol{X}, Y)$ on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_o(\mathcal{X}) \times 2^{\mathcal{Y}})$, where $\mathcal{B}_o(\mathcal{X})$ is the σ-algebra generated by the open balls defined by the uniform metric $d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sup_{1 \le j < \infty} |x_1^{(j)} - x_2^{(j)}|$ over $\mathcal{X}$.

We now develop a theory to quantify the magnitude of $e_V(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})$ and $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})|$ as a function of n, k, p and s, which in turn yields $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(\infty)})|$. Our theory uses two components: (1) a rate of convergence of the excess hinge risk $e_V(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})$; (2) a conversion formula that establishes a relation between $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})|$ and $e_V(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})$.

3.2 Preliminary result: Convergence rate of $e_V(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})$

Theorem 1 yields a rate of $e_V(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})$, when $p = p_n$ and $s = s_n$ grow with n, as $n \to \infty$.

Theorem 1 *Assume that* $\tau_n = (n^{-1} \log p_n)^{1/2} \to 0$ *as* $n \to \infty$. *Then,* $e_V(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p_n)}) = O\left(\max(s_n \tau_n \log(\tau_n^{-1}), d_n)\right)$, *a.s. under* P, *where* $d_n = \inf_{\boldsymbol{f} \in \mathcal{F}(p_n, s_n)} e_V(\boldsymbol{f}, \boldsymbol{f}^{(p_n)})$.

In Theorem 1, d_n is the approximation error of $\mathcal{F}(p_n, s_n)$ to $\cup_{0 \le s < \infty} \mathcal{F}(p_n, s)$, which tends to 0 as $s_n \to \infty$. The optimal value of the tuning parameter s_n with respect to the optimal rate of $e_V(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p_n)})$ is roughly determined by $s_n \tau_n \log(\tau_n^{-1}) \sim d_n$, yielding the best trade-off between approximation and estimation.

When an additional assumption of L_1-norm sparseness is made in Assumption A, the rate in Theorem 1 can be greatly simplified.

Assumption A: There exists a finite s^* such that $\boldsymbol{f}^{(p)} \in \mathcal{F}(p, s^*)$ for all p.

Assumption A generally describes an L_1-norm "sparse scenario", which is satisfied when the number of relevant predictors is finite. For example, suppose $\boldsymbol{f}^{(p)}$ has a form of $f_c^{(p)}(x) = \sum_{j=1}^{J} w_{c,j}^* x^{(j)}$; $c = 1, \ldots, k$, for all $p = J, \ldots, \infty$, then Assumption A is met with a choice of $s^* = \sum_{c=1}^{k} \sum_{j=1}^{J} |w_{c,j}^*|$.

Corollary 1 *Under Assumption A, let* $s_n = s^*$ *for all* n. *Then*

$$e_V(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p_n)}) = O\left(\tau_n \log(\tau_n^{-1})\right) = O\left((n^{-1} \log p_n)^{1/2} \log(n (\log p_n)^{-1})\right). \text{ a.s. under } P,$$

In contrast to the classical asymptotic results, where p is usually required to be no greater than n, Corollary 1 yields an error rate tending to zero as long as p_n grows no faster than $\exp(n)$, and yields an error rate of the order $n^{-1/2}(\log n)^{3/2}$ when p_n grows no faster than n^{r_0}, which is of nearly the same order as in the case of $p_n \leq n$. This explains the phenomenon that L1MSVM is less sensitive to an increase of dimension than L2MSVM as described in Section 4.1.2.

3.3 Main result I: Convergence rate of $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})|$

To obtain the convergence rate of $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})|$, we establish a relationship between $e_V(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})$ and $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})|$. Two situations are considered: (1) $\boldsymbol{f}^{(p)} \in \mathcal{F}(p)$ for all p, or (2) $\boldsymbol{f}^{(p^*)} \notin \mathcal{F}(p^*)$ for some p^*. In the first case the relationship greatly depends on the marginal distribution of $\boldsymbol{X}(p)$, whereas in the second case it is independent.

Assumptions B1 and B2 below are made respectively for the cases (1) and (2). Define an L_2 metric over $\mathcal{F}(p)$ as $d(\boldsymbol{f}, \boldsymbol{f}') = (\sum_{c=1}^{k} E(f_c(\boldsymbol{X}) - f_c'(\boldsymbol{X}))^2)^{1/2}$.

Assumption B1: There exist constants $1 \leq r < +\infty$, $0 < \alpha \leq \infty$, and $c_1(p)$ and $c_2(p)$ that may depend on p, such that for all small $\epsilon > 0$, $\boldsymbol{f}^{(p)} \in \mathcal{F}(p)$ and all p,

$$\inf_{\{d(\boldsymbol{f}, \boldsymbol{f}^{(p)}) \geq \epsilon, \boldsymbol{f} \in \mathcal{F}(p)\}} e_V(\boldsymbol{f}, \boldsymbol{f}^{(p)}) \geq c_1(p)\epsilon^r, \tag{9}$$

$$\sup_{\{d(\boldsymbol{f}, \boldsymbol{f}^{(p)}) \leq \epsilon, \boldsymbol{f} \in \mathcal{F}(p)\}} |e(\boldsymbol{f}, \boldsymbol{f}^{(p)})| \leq c_2(p)\epsilon^{\alpha}. \tag{10}$$

Assumption B2: There exists an p^* such that $\boldsymbol{f}^{(p^*)} \notin \mathcal{F}(p^*)$. Assume that for all p, the distribution $P(\boldsymbol{X}(p), Y)$ is regular in the sense that the marginal distribution of $\boldsymbol{X}(p)$ has a finite density $q(\boldsymbol{x}) \leq U$ for some finite $U > 0$, with respect to the Lebesgue measure λ on $[0,1]^p$, and $P(B_c) > 0$; $c = 1, \ldots, k$, where $B_c = \{\boldsymbol{x} \in [0,1]^p : c = \arg\max_{1 \leq j \leq k} \eta_j(\boldsymbol{x})\}$ with $\eta_j(\boldsymbol{x}) = P(Y = j | \boldsymbol{X} = \boldsymbol{x})$.

Assumption B1 describes the nonseparable case, whereas Assumption B2 concerns the separable case, as to be shown in Lemma 5 in Section 6. In Assumption B1, because $\boldsymbol{f}^{(p)} \in \mathcal{F}(p)$, local smoothness of functionals $E[V(\cdot, \boldsymbol{Z})]$ and $E[l(\cdot, \boldsymbol{Z})]$ can

be characterized as in (9) and (10) within a neighborhood of $\boldsymbol{f}^{(p)}$. For example, both $E[V(\boldsymbol{f}, \boldsymbol{Z})]$ and $E[l(\boldsymbol{f}, \boldsymbol{Z})]$ can be parameterized as functions of $(\boldsymbol{W}, \boldsymbol{b})$, $R_V(\boldsymbol{W}, \boldsymbol{b})$ and $R(\boldsymbol{W}, \boldsymbol{b})$, which are generally piecewise differentiable. This yields $r = 2$ in (9), when $R_V(\boldsymbol{W}, \boldsymbol{b})$ is twice differentiable at $(\boldsymbol{W}^*, \boldsymbol{b}^*) = \arg\min_{(\boldsymbol{W},\boldsymbol{b})} R_V(\boldsymbol{W}, \boldsymbol{b})$ with a positive definite Hessian matrix. In (10), a Lipschitz condition is given with the exponent α determined by $R(\boldsymbol{W}, \boldsymbol{b})$ depending on the distribution P.

Lemma 2 *Under Assumption B1, there exists a constant $c(p) > 0$ that may depend on p such that for all $\boldsymbol{f} \in \mathcal{F}(p)$,*

$$|e(\boldsymbol{f}, \boldsymbol{f}^{(p)})| \leq c(p) e_V(\boldsymbol{f}, \boldsymbol{f}^{(p)})^{\frac{\alpha}{r}}. \tag{11}$$

Under Assumption B2, for all $\boldsymbol{f} \in \mathcal{F}(p)$ and $p \geq p^*$,

$$|e(\boldsymbol{f}, \boldsymbol{f}^{(p)})| \leq e_V(\boldsymbol{f}, \boldsymbol{f}^{(p)}). \tag{12}$$

In Lemma 2, (11) is formulated for a general case, where $c(p)$, determined by $c_1(p)$ and $c_2(p)$ in Assumption B1, may depend on p. When $c(p)$ depends on p, it may affect the convergence rate in Corollary 2. In Section 3.5, an example is given with $c(p)$ independent of p, $r = 2$, and α that can be arbitrarily large. In (12), the conversion exponent is 1, which coincides with the result on the binary hinge loss in Bartlett, Jordan and McAuliffe (2006). This is not surprising because in the linearly separable case, $\boldsymbol{f}^{(p)}$ becomes the Bayes rule with respect to all measurable decision function vectors, which reduces to the case of Fisher-consistency (c.f. Lin, 2002, 2004).

Corollary 2 *Assume that $\tau_n = (n^{-1} \log p_n)^{1/2} \to 0$ as $n \to \infty$. If Assumption B1 is met, then $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p_n)})| = O\left(c(p_n)(\max(s_n \tau_n \log(\tau_n^{-1}), d_n))^{\frac{\alpha}{r}}\right)$, a.s. under P. If Assumption B2 is met, then $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p_n)})| = O\left(\max(s_n \tau_n \log(\tau_n^{-1}), d_n)\right)$, a.s. under P.*

Corollary 3 *Under Assumptions A, only case (1) occurs. If assumption B1 is met, let $s_n = s^*$ for all n. Then $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p_n)})| = O\left(c(p_n)((n^{-1} \log p_n)^{1/2} \log(n(\log p_n)^{-1}))^{\frac{\alpha}{r}}\right)$, a.s. under P.*

Corollary 3 yields an error rate depending on $(n^{-1}\log p_n)^{1/2}$, $c(p_n)$ and the conversion exponent α/r. Under a suitable condition of the distribution, α can be arbitrarily large, yielding an error rate that is much faster than n^{-1}, as shown in Section 3.5.

3.4 Main result II: Convergence rate of $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(\infty)})|$

Finally, the rate of convergence of $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(\infty)})|$ can be obtained using the results in Corollaries 3 and 4, and the approximation error of $\mathcal{F}(p_n)$ to $\boldsymbol{f}^{(\infty)}$ $r_n = |e(\boldsymbol{f}^{(p_n)}$.

Corollary 4 *If* $\tau_n = (n^{-1}\log p_n)^{1/2} \to 0$ *as* $n \to \infty$ *and Assumption B1 are met, then* $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(\infty)})| = O\left(\max(c(p_n)(\max(s_n\tau_n \log(\tau_n^{-1}), d_n))^{\frac{\alpha}{r}}, r_n)\right)$, *a.s. under* P. *If the assumptions of Corollary 1 and Assumption B2 are met, then* $r_n = 0$ *for* $p \geq p^*$, $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(\infty)})| = O\left(\max(s_n\tau_n \log(\tau_n^{-1}), d_n)\right)$, *a.s. under* P. *If Assumptions A and B1 are met, let* $s_n = s^*$ *for all* n. *Then* $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(\infty)})| = O\left(\max(c(p_n)(\tau_n \log(\tau_n^{-1}))^{\frac{\alpha}{r}}, r_n)\right)$ *a.s. under* P.

3.5 An illustrative example

This section applies our general theory to one specific binary classification example for illustration. This example is chosen to demonstrate that the rate of convergence of L1MSVM can be arbitrarily fast because of the distribution of $\boldsymbol{Z} = (\boldsymbol{X}, Y)$. This is contrary to a general belief that the n^{-1} rate is typically expected by a large margin classifier.

Predictors $X^{(j)}$; $j = 1, \ldots, \infty$, are i.i.d. according to probability density $q(x) = 2^{\beta}(\beta+1)|x - 1/2|^{\beta}$ for $x \in [0, 1]$ and $\beta \geq 0$, with β corresponding to the uniform distribution. For each $\boldsymbol{X} = (X^{(1)}, \cdots)$, Y is assigned to 1 if $X^{(1)} > 1/2$ and -1 otherwise. Then, Y is randomly flipped with a constant probability $0 \leq \theta < 1/2$. Note that $P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$ is $(1-\theta)$ for $x^{(1)} > 1/2$, and is θ otherwise. Consequently, classification depends only on $X^{(1)}$, and is linearly nonseparable when $0 < \theta < 1/2$, and is linearly separable when $\theta = 0$. Now consider two cases below.

Case 1: (Nonseparable, $0 < \theta < 1/2$**).** In this case, we verify Assumption B1.

Note that $P(X^{(1)}, Y|X^{(2)}, \ldots, X^{(p)}) = P(X^{(1)}, Y)$, because $(X^{(1)}, Y)$ is independent of $X^{(j)}$; $j = 2, \ldots, p$. Then, for any $\boldsymbol{f} \in \mathcal{F}(p)$, $E(V(\boldsymbol{f}, \boldsymbol{Z})|X^{(2)}, \ldots, X^{(p)}) = EV(\boldsymbol{f}, (X^{(1)}, Y)) \geq EV(\boldsymbol{f}^{(1)}, (X^{(1)}, Y))$ and $EV(\boldsymbol{f}, \boldsymbol{Z}) \geq EV(\boldsymbol{f}^{(1)}, (X^{(1)}, Y))$, implying $\boldsymbol{f}^{(p)} = \boldsymbol{f}^{(1)}$. An application of the same conditioning argument yields that there exist constants $c_j > 0; j = 1, 2$ such that $e_V(\boldsymbol{f}, \boldsymbol{f}^{(1)}) \geq c_1 d(\boldsymbol{f}, \boldsymbol{f}^{(1)})^r$ and $e(\boldsymbol{f}, \boldsymbol{f}^{(1)}) \leq c_2 d(\boldsymbol{f}, \boldsymbol{f}^{(1)})^\alpha$ for $\boldsymbol{f} \in \mathcal{F}(p)$ with sufficiently small $d(\boldsymbol{f}, \boldsymbol{f}^{(1)})$, if and only if $e_V(\boldsymbol{f}, \boldsymbol{f}^{(1)}) \geq c_1 d(\boldsymbol{f}, \boldsymbol{f}^{(1)})^r$ and $e(\boldsymbol{f}, \boldsymbol{f}^{(1)}) \leq c_2 d(\boldsymbol{f}, \boldsymbol{f}^{(1)})^\alpha$ for $\boldsymbol{f} \in \mathcal{F}(1)$ with sufficiently small $d(\boldsymbol{f}, \boldsymbol{f}^{(1)})$. Thus, without loss of generality, we restrict our discussion to $\boldsymbol{f} \in \mathcal{F}(1) = \{\boldsymbol{f} : \boldsymbol{f}(\boldsymbol{x}) = (ax^{(1)} + b, -ax^{(1)} - b)^T\}$. Write $EV(\boldsymbol{f}, \boldsymbol{Z})$ as $R_V(a, b)$. It can be verified that $R_V(a, b)$ is piecewisely differentiable and convex. For $0 \leq -(1 + b)a^{-1} \leq 1/2$ and $1/2 \leq (1 - b)a^{-1} \leq 1$,

$$R_V(a, b) = \lambda \left(\frac{\theta}{2^{\beta+2}}(a(\beta + 1) + 2\beta + 4) + \frac{1 - \theta}{2a^{\beta+1}}((1 + \frac{a}{2} + b)^{\beta+2} + (1 - \frac{a}{2} - b)^{\beta+2}) \right),$$

with the minimizer $(a^*, b^*) = (2(\frac{1-\theta}{\theta})^{1/(\beta+2)}, -(\frac{1-\theta}{\theta})^{1/(\beta+2)})$ and positive definite Hessian matrix

$$H_1 = \lambda_1 \begin{pmatrix} (a^*)^{-(\beta+3)} + \frac{1}{4}(a^*)^{-(\beta+1)} & \frac{1}{2}(a^*)^{-(\beta+1)} \\ \frac{1}{2}(a^*)^{-(\beta+1)} & (a^*)^{-(\beta+1)} \end{pmatrix}$$

at (a^*, b^*). This implies that (a^*, b^*) is the global minimizer of $R_V(a, b)$ by convexity, and hence that $\boldsymbol{f}^{(1)} = (a^* x^{(1)} + b^*, -a^* x^{(1)} - b^*)^T$. For any $(a^* + e_1, b^* + e_2)$ in the neighborhood of (a^*, b^*), $e_V(\boldsymbol{f}, \boldsymbol{f}^{(1)}) = R_V(a^* + e_1, b^* + e_2) - R_V(a^*, b^*) \geq \lambda_2(e_1, e_2)H_1(e_1, e_2)^T$. Note that $d(\boldsymbol{f}, \boldsymbol{f}^{(1)})^2 = 2E(f_1(\boldsymbol{X}) - f_1^{(1)}(\boldsymbol{X}))^2 = (e_1, e_2)H_2(e_1, e_2)^T$ with

$$H_2 = \begin{pmatrix} \frac{\beta+2}{\beta+3} & 1 \\ 1 & 2 \end{pmatrix}.$$

Hence there exists a constant $c_1 > 0$ such that $e_V(\boldsymbol{f}, \boldsymbol{f}^{(1)}) \geq c_1 d(\boldsymbol{f}, \boldsymbol{f}^{(1)})^2$. Consequently $r = 2$ in (9). For (10), there exist some constants $\lambda_j > 0$, such that $e(\boldsymbol{f}, \boldsymbol{f}^{(1)}) = \lambda_3(2\theta - 1)|\frac{2e_2 + e_1}{2(a^* + e_1)}|^{\beta+1} \leq \lambda_4|2e_2 + e_1|^{\beta+1}$, while $d(\boldsymbol{f}, \boldsymbol{f}^{(1)}) = (2E(f_1(\boldsymbol{X}) - f_1^{(1)}(\boldsymbol{X}))^2)^{1/2} \geq \sqrt{2}E|f_1(\boldsymbol{X}) - f_1^{(1)}(\boldsymbol{X})| = \sqrt{2}E|e_1 X^{(1)} + e_2| = \sqrt{2}|2e_2 + e_1|E|\frac{e_1}{2e_2 + e_1}(X^{(1)} - 1/2) +$

$1/2| \geq \lambda_5|2e_2 + e_1|$. Therefore there exists a constant $c_2 > 0$ such that $e(\boldsymbol{f}, \boldsymbol{f}^{(1)}) \leq c_2 d(\boldsymbol{f}, \boldsymbol{f}^{(1)})^{(\beta+1)}$. As a result, Assumption B1 is fulfilled with $r = 2$ in (9) and $\alpha = \beta + 1$ in (10), for c_i; $i = 1, 2$, independent of p. It follows from Lemma 2 that $|e(\boldsymbol{f}, \boldsymbol{f}^{(p)})| \leq c e_V(\boldsymbol{f}, \boldsymbol{f}^{(p)})^{(\beta+1)/2}$, and from Corollary 3 that $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p_n)})| = O\left(((n^{-1}\log p_n)^{1/2}\log(n/\log p_n))^{(\beta+1)/2}\right)$ a.s. under P, for $\hat{\boldsymbol{f}}$ in (6) with p_n satisfying $n^{-1}\log(p_n) \to 0$. Note that the sign of $\boldsymbol{f}^{(p_n)} = \boldsymbol{f}^{(1)}$ coincides with $\bar{\boldsymbol{f}} = (\text{sign}(x^{(1)} - 1/2), -\text{sign}(x^{(1)} - 1/2))^T$ the Bayes rule over all functions. Therefore $El(\boldsymbol{f}^{(p_n)}, \boldsymbol{Z}) = El(\bar{\boldsymbol{f}}, \boldsymbol{Z})$ and the Bayesian regret $e(\hat{\boldsymbol{f}}, \bar{\boldsymbol{f}}) = O\left(((n^{-1}\log p_n)^{1/2}\log(n/\log p_n))^{(\beta+1)/2}\right)$ converges to 0 arbitrarily fast as $\beta \to \infty$.

Case 2: (Separable, $\theta = 0$). In this case, we verify Assumption B2. Take $\boldsymbol{g}^{(C)} = C(x^{(1)} - 2^{-1}, -x^{(1)} + 2^{-1})^T \in \mathcal{F}(1)$ with $C \to \infty$. Easily, $EV(\boldsymbol{g}^{(C)}, \boldsymbol{Z}) = O(C^{-(\beta+1)}) \to 0$, and thus $EV(\boldsymbol{f}^{(1)}, \boldsymbol{Z}) = \inf_{\boldsymbol{f}\in\mathcal{F}(1)} EV(\boldsymbol{f}, \boldsymbol{Z}) = 0$. Therefore, $\boldsymbol{f}^{(1)} \notin \mathcal{F}(1)$ and Assumption B2 is met with $p^* = 1$. It follows from Lemma 2 that $|e(\boldsymbol{f}, \boldsymbol{f}^{(p)})| \leq e_V(\boldsymbol{f}, \boldsymbol{f}^{(p)})$ for all $p \geq 1$. By Corollary 2, $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p_n)})| = O\left(\max(s_n\tau_n\log(\tau_n^{-1}), d_n)\right)$, where the optimal rate is achieved when s_n is chosen to satisfy $s_n\tau_n\log(\tau_n^{-1}) \approx d_n$. Note that for given s_n, $d_n = \inf_{\boldsymbol{f}\in\mathcal{F}(p_n,s_n)} EV(\boldsymbol{f}, \boldsymbol{Z}) \leq EV(\boldsymbol{g}^{(s_n)}, \boldsymbol{Z}) = O(s_n^{-(\beta+1)})$. The best trade-off is achieved by $s_n = O((\tau_n\log(\tau_n^{-1}))^{-1/(\beta+2)})$. Consequently, the convergence rate of $|e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p_n)})|$ is $O\left(((n^{-1}\log p_n)^{1/2}\log(n/\log p_n))^{(\beta+1)/(\beta+2)}\right)$, for $\hat{\boldsymbol{f}}$ in (6) with p_n satisfying $n^{-1}\log(p_n) \to 0$. In this case, $El(\bar{\boldsymbol{f}}, \boldsymbol{Z}) = El(\boldsymbol{f}^{(p)}, \boldsymbol{Z}) = 0$. This yields a rate of convergence of $e(\hat{\boldsymbol{f}}, \bar{\boldsymbol{f}}) = El(\hat{\boldsymbol{f}}, \boldsymbol{Z}) = O\left(((n^{-1}\log p_n)^{1/2}\log(n/\log p_n))^{(\beta+1)/(\beta+2)}\right)$ a.s. under P, which is in contrast to the nonseparable case.

In conclusion, this example reveals an important aspect of classification as discussed in Section 1. In the nonseparable case, the distribution of $\boldsymbol{Z}$ plays an important role, as characterized by β in Assumption B1. When β is arbitrarily large, it results in an arbitrarily fast rate of convergence of $e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})$. In the separable case, however, the distribution is less important, which affects only the approximation error of $\mathcal{F}(p, s)$. In

this case, the conversion exponent is exactly 1, which seems to coincide with the result of Bartlett, Jordan and McAuliffe (2006). In general, when the candidate function class $\mathcal{F}$ is not sufficiently large, the risk minimizer is usually not the (global) Bayes rule. As a result, the marginal distribution of $\boldsymbol{X}$ matters, which is contrary to the general belief that only the conditional distribution of Y given $\boldsymbol{X}$ is relevant to classification.

4 Numerical studies

4.1 Simulation

This section examines the performance of L1MSVM with respect to its generalization accuracy and variable selection in both simulated and benchmark examples. We compare it against OVA and L2MSVM in Lee, Lin and Wahba (2004).

4.1.1 L1MSVM versus OVA:

We compare L1MSVM with OVA in two situations characterized by the level of difficulty of classification, as measured by two parameters: the number of classes k and the degree of overlapping among classes d. Consider k-class classification with $k = 4, 8$. First, sample $(u_{i,1}, \ldots, u_{i,100})$ from $N(0, I_{100\times 100})$; $i = 1, \ldots, 20k$. Second, randomly assign instances to k classes with twenty instances in each class. Third, perform linear transformation: $x_{i,j} = u_{i,j} + a_j$; $j = 1, 2$ and $x_{i,j} = u_{i,j}$; $j = 3, \ldots, 100$, with $(a_1, a_2) = (d\cos(2(c-1)\pi/k), d\sin(2(c-1)\pi/k))$ for classes c; $c = 1, \ldots, k$, where three values $d = 1, 2, 3$ are examined. Evidently, only the first two components of $\boldsymbol{x}$ are relevant to classification, whereas the rest 98 components are redundant.

For both L1MSVM and OVA, the tuning parameter λ is optimized over a discrete set of values with respect to the test error on an independent test sample of size $20,000$, which well approximates the generalization error. For OVA, the tuning parameter λ in (1) is selected from 63 values $\lambda = i10^j$; $i = 1, \ldots, 9$, $j = -3, \ldots, 3$. For L1MSVM, it is tuned through the solution path of Wang and Shen (2006) with s chosen from $s = (i-1)t^*/63$; $i = 1, \ldots, 63$, where t^*, as defined in Section 2.2, is the termination

point of the solution path. After tuning, the optimal test errors of L1MSVM and OVA are averaged over 100 simulation replications, as well as the number of selected variables. The results are summarized in Table 1.

Table 1 about here.

With regard to prediction, L1MSVM performs better than OVA in every single case, except in case of $d = 3$ and $k = 4$ where the improvement is not large in view of their standard errors. This may be explained by presence/absence of a dominating class. When d is small and k is large, the classes overlap largely, resulting in a large fraction of instances without a dominating class. In this case, OVA suffers from this difficulty. With respect to variable selection, L1MSVM outperforms OVA in that it removes more redundant variables, and surprisingly, it selects a nearly correct subset of variables even in this difficult situation with largely overlapping classes. In contrast, OVA selects more redundant variables, which seems to agree with the aforementioned discussion regarding OVA in Section 1.

4.1.2 L_1-norm versus L_2-norm:

We compare the performances of L1MSVM and L2MSVM with respect to dimension p. Consider a three-class classification. First, twenty instances are generated for each class as a training sample. For class 1, each instance $\boldsymbol{x} = (x_1, \ldots, x_p)$ is sampled as follows: $(x_1, \ldots, x_p)$ are independent normals with $x_j \sim N(\sqrt{2}, 1)$; $j = 1, 2$, and $x_j \sim N(0, 1)$; $j = 3, \ldots, p$, where $p = 10, 20, 40, 80, 160$ are examined. For classes 2 and 3, instances are generated in the same manner except that the first two components of $\boldsymbol{x}$ are centered at $(-\sqrt{2}, -\sqrt{2})$ and $(\sqrt{2}, -\sqrt{2})$. Evidently, only the first two components of $\boldsymbol{x}$ are relevant to classification, whereas the rest $p - 2$ components are redundant.

To evaluate predictive performance, a test sample of size 30,000 is generated, with 10,000 in each class. For L1MSVM, the tuning parameter s is optimized over a discrete set over $[0, t^*]$ as described in Section 4.1.1. For L2MSVM, the tuning parameter λ

is selected in the same fashion as OVA. The optimal test error is averaged over 100 simulation replications, and is given in Table 2 and displayed in Figure 3.

Table 2 and Figure 3 about here.

As indicated by Table 2 and Figure 3, L1MSVM outperforms L2MSVM in every single case. As the number of redundant variables increases, the performance of L1MSVM appears to be stable, whereas L2MSVM deteriorates faster. This is because L1MSVM is able to remove redundant features, which is in contrast to L2MSVM involving all the variables. This is consistent with our statistical learning theory.

4.2 Application to gene expression microarray data

One important application of L1MSVM is cancer genomics classification. Consider a study of gene expressions in acute leukemia described in Golub et al. (1999). This study examined $n = 72$ samples from three types of acute leukemias with 38 samples in B-cell ALL (acute lymphoblastic leukemia), 9 samples in T-cell ALL and 25 samples in AML (acute myelogenous leukemia), involving $p = 6,817$ genes typed through the Affymetrix technology. After pre-processing consisting of thresholding, filtering, and standardization, 3571 genes are left, c.f., Dudoit et. al. (2002). Detailed descriptions can be found at http://www.genome.wi.mit.edu/MPR. The goal of this study is to predict the types of tumors from their genomics information, and to select the best subset of genes that are most responsible for leukemias. We apply L1MSVM, and compare its predictive performance with OVA and L2MSVM.

In this study, we examine predictive performance through cross-validation, that is, we randomly select two thirds of the 72 samples by stratified random sampling for training, while leaving the remaining one third for testing. This process is repeated 100 times, and in each data partition, the tuning parameter is selected by five-fold cross-validation on the training sample from a set of pre-specified values as described in Section 4.1. Then, the test error and the corresponding number of selected genes

for each method are averaged over 100 data partitions, and are reported in Table 3, as well as the standard errors.

Table 3 about here.

As suggested by Table 3, L1MSVM outperforms OVA in terms of predictive performance. Moreover, L1MSVM performs slightly better than L2MSVM, while yielding a much simpler model involving about 20 genes. This appears appealing as a subset of most informative genes provides a guideline for biologists to perform additional confirmatory experiments.

5. Summary

This article proposes a novel L1MSVM that performs variable selection and classification simultaneously. In particular, the proposed method, together with the solution path, permits a treatment of data with high dimension, low sample size that can be difficult for conventional methods. L1MSVM is shown to perform well as long as p does not grow too fast and a sparse representation of the underlying decision function is obtainable. Compared to the L_2-penalty, the L_1-penalty achieves the desired objective of variable selection, especially in presence of many irrelevant variables. Compared to OVA, L1MSVM uses a single generalized loss function, which addresses the problems of OVA in feature selection and classification. Moreover, the method is readily applicable to other multi-class losses, such as the generalized hinge losses in (2)-(3), the multi-class deviance loss (c.f., Zhu and Hastie 2005), and multi-class exponential loss (c.f., Zhu, Rosset, Zou and Hastie 2005), although a detailed analysis is performed only for a specific generalized hinge loss (4) in this article. Further theoretical investigation is necessary to compare the performances of various losses over different classes of candidate functions.

Our investigation in L1MSVM also provides an insight into regression analysis, particularly LASSO (Tibshirani 1996) with regard to feature selection involving high

dimension, low sample size data. It is possible that a similar result can be established for LASSO with the technique developed in this article.

6. Appendix: technical proofs.

The proof of Theorem 1 and Corollary 1 uses a large deviation empirical process inequality in Theorem 2, which bounds the tail probability of $e_V(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})$ for given n and $\mathcal{F}(p,s)$ based on Lemmas 3 and 4. Lemmas 5-7 are used to establish a conversion formula between $e_V(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})$ and $e(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})$ in Lemma 2, which leads to Corollaries 2-4.

Theorem 2 *Given p, k, n, and $0 < \theta < 1$, assume there exists an $M > 0$ such that*

$$(\log_2 \frac{8\varepsilon_0\sqrt{6}}{\theta M} + 1)(\frac{128\log(e + e(2k(p+1))\varepsilon_0^2)}{n\theta})^{1/2} \le \theta M/8, \tag{13}$$

where ε_0 is given by

$$2\varepsilon_0^{-2}\log(e + e(2k(p+1))\varepsilon_0^2) = \theta n M^2/2. \tag{14}$$

Then for $\hat{\boldsymbol{f}}$ defined in (6) and $d = \inf_{\boldsymbol{f}\in\mathcal{F}(p,s)} e_V(\boldsymbol{f}, \boldsymbol{f}^{(p)})$

$$P(e_V(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)}) \ge 8k(3s+2k)M + d) \le 6[1 - \frac{1}{16nM^2}]^{-1}\exp(-2(1-\theta)nM^2). \tag{15}$$

Before proving Theorem 2, we introduce some notations. First, note that for $\hat{\boldsymbol{f}}$ defined in (6), $\hat{b}_c \ge -\max_{i:\, y_i\neq c}(\hat{\boldsymbol{w}}_c^T\boldsymbol{x}_i + 1)$, because $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{b}})$ minimizes $\sum_{i=1}^n V(\boldsymbol{f}, z_i)$. Hence, $\hat{b}_c \ge -(\|\hat{\boldsymbol{w}}_c\|_1 + 1)$ for $c = 1, \ldots, k$, implying $\|\hat{\boldsymbol{b}}\|_1 = -2\sum_c b_c I[b_c \le 0] \le 2(s+k)$. Therefore, $\hat{\boldsymbol{f}} \in \mathcal{F}(p,s) \cap \{\boldsymbol{f} : \|b\|_1 \le 2(s+k)\}$, and it suffices to consider $\mathcal{F}^b(p,s) = \mathcal{F}(p,s) \cap \{\boldsymbol{f} : \|b\|_1 \le 2(s+k)\}$. Next, define $\boldsymbol{f}_0 = \arg\min_{\mathcal{F}(p,s)} EV(\boldsymbol{f}, \boldsymbol{Z})$, $h_{\boldsymbol{f}}(\cdot) = (2k(3s+2k))^{-1}(V(\boldsymbol{f}, \cdot) - V(\boldsymbol{f}_0, \cdot))$, and $\mathcal{H} = \{h_{\boldsymbol{f}} : \boldsymbol{f} \in \mathcal{F}^b(p,s)\}$. Here the scaling factor $(2k(3s+2k))^{-1}$ is chosen to normalize $h_{\boldsymbol{f}}$ such that the $L_2(Q)$-diameter of $\mathcal{H}$ is no larger than 1 for any distribution Q. Second, define the indexed empirical processes as $h \mapsto v_n(h) = P_n h - Ph$, where $h \in \mathcal{H}$, $Ph = \int h dP$ and

$P_n h = n^{-1}\sum_{i=1}^n h(\boldsymbol{Z}_i)$ with $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ i.i.d. random elements from distribution P.

Proof of Theorem 2: First, we establish a connection between $e_V(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)})$ and the empirical processes $P_n(h) - P(h)$. Because $\boldsymbol{f}_0 = \arg\min_{\mathcal{F}(p,s)} EV(\boldsymbol{f}, \boldsymbol{Z})$, $d = \inf_{\boldsymbol{f}\in\mathcal{F}(p,s)} e_V(\boldsymbol{f}, \boldsymbol{f}^{(p)}) = e_V(\boldsymbol{f}_0, \boldsymbol{f}^{(p)})$ and $e_V(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p)}) > 8k(3s+2k)M + d$ implies $e_V(\hat{\boldsymbol{f}}, \boldsymbol{f}_0)(2k(3s+2k))^{-1} > 4M$. Note that $\hat{\boldsymbol{f}}$ is a minimizer of $n^{-1}\sum_{i=1}^n V(\boldsymbol{f}, \boldsymbol{z}_i)$ over $\mathcal{F}(p,s)$, and $\hat{\boldsymbol{f}}$, $\boldsymbol{f}_0 \in \mathcal{F}^b(p,s)$.

$$
\begin{aligned}
&P(e_V(\hat{\boldsymbol{f}}_n, \boldsymbol{f}^{(p)}) > 8k(3s+2k)M + d) \le P(e_V(\hat{\boldsymbol{f}}_n, \boldsymbol{f}_0)(2k(3s+2k))^{-1} > 4M) \\
&\le\ P^*\left(\sup_{\{\boldsymbol{f}\in\mathcal{F}^b(p,s): e_V(\boldsymbol{f},\boldsymbol{f}_0)(2k(3s+2k))^{-1}>4M\}} n^{-1}\sum_{i=1}^n (V(\boldsymbol{f}_0, \boldsymbol{z}_i) - V(\boldsymbol{f}, \boldsymbol{z}_i)) > 0\right) \\
&\le\ P^*\left(\sup_{\{\boldsymbol{f}\in\mathcal{F}^b(p,s): e_V(\boldsymbol{f},\boldsymbol{f}_0)(2k(3s+2k))^{-1}>4M\}} -n^{-1}(\sum_{i=1}^n h_{\boldsymbol{f}}(Z_i) + Eh_{\boldsymbol{f}}(Z)) \right. \\
&\qquad\qquad \left. > (2k(3s+2k))^{-1}E(V(\boldsymbol{f}, \boldsymbol{z}_i) - V(\boldsymbol{f}_0, \boldsymbol{z}_i))\right) \\
&\le\ P^*(\sup_{h\in\mathcal{H}} |P_n h - Ph| > 4M).
\end{aligned}
$$

The result then follows from lemma 4. □

We are now ready to prove our main result.

Proof of Theorem 1: Take $M = K\tau_n \log(\tau_n^{-1})$ with $K < \infty$ a constant to be determined later on. It remains to verify (13) holds for the choice of M and ε_0 determined by (14). Note that $\log(\tau_n^{-1}) \to \infty$ and ε_0 satisfies $\theta n M^2/2 = (2\varepsilon_0^{-2})\log(e + e(2k(p_n+1))\varepsilon_0^2)$. One can verify that $\varepsilon_0 \to 0$. Note that (13) is equivalently to $(\log_2 \frac{8\varepsilon_0\sqrt{6}}{\theta M} + 1)^2 \le \frac{\theta^3 n M^2}{2^{13}\log(e+e(2k(p_n+1))\varepsilon_0^2)}$, in which the order of the left hand side is no greater than $O(\log^2(\tau_n^{-1}))$, whereas that of the right hand side is no less than $O(\log^2(\tau_n^{-1}))$. Therefore, (13) holds for sufficiently large n. Note that $nM^2 \ge K\log(p_n)\log(\tau_n^{-1}) = (K/2)\log(p_n)\log(n/\log(p_n)) \ge (K/2)\log(n)$ for sufficiently large n. Take $K \ge 2/(1-\theta)$. Then, the result follows from the Borel-Cantelli theorem and Theorem 1. □

Proof of Corollary 1: As $p_n \to \infty$, $d_n = 0$ for sufficiently large n. The result follows from Theorem 1. □

Proofs of Corollaries 2-4: The results follow from Theorem 1 and Corollary 1 and the corresponding conversion formula in Lemma 2. □

Next we prove other technical lemmas.

Proof of Lemma 1: Let $w_{c,0} = b_c$, and define $\mathcal{A}(s) = \{(c,j) : \hat{w}_{c,j}(s) \neq 0; c = 1,\ldots,k; j = 0,1,\ldots,p\}$, $\mathcal{J}(s) = \{j : \hat{w}_{c,j}(s) \neq 0 \text{ for some } c\}$, $\mathcal{E}(s) = \{(i,c) : \sum_{j=0}^{p} \hat{w}_{c,j}(s)x_{ij} + 1 = 0, c \neq y_i\}$. By Theorem 2 in Wang and Shen (2006), $|\mathcal{A}(s)| \leq |\mathcal{E}(s)| + |\mathcal{J}(s)| + 2$ with $|\mathcal{S}|$ denoting the cardinality of $\mathcal{S}$. Note that the number of non-zero columns of $\hat{\boldsymbol{W}}$ is $|\mathcal{J}| - 1$ and $|\mathcal{A}| - k \geq 2(|\mathcal{J}| - 1)$. Then $|\mathcal{J}| - 1 \leq |\mathcal{E}(s)| - k + 3$. In addition, $|\mathcal{E}(s)| \leq n(k-1) - 1$ for $s < t^*$. Hence, $|\mathcal{J}| - 1 \leq (n-1)(k-1)$. This completes the proof. □

Proof of Lemma 2: By Assumption B1, (9) implies $e_V(\boldsymbol{f}, \boldsymbol{f}^{(p)}) \geq c_1(p)d(\boldsymbol{f}, \boldsymbol{f}^{(p)})^r$, and (10) implies $|e(\boldsymbol{f}, \boldsymbol{f}^{(p)})| \leq c_2(p)d(\boldsymbol{f}, \boldsymbol{f}^{(p)})^\alpha$, when $d(\boldsymbol{f}, \boldsymbol{f}^{(p)})$ is small. Therefore, there exists a sufficiently small ϵ, such that $\frac{|e(\boldsymbol{f}, \boldsymbol{f}^{(p)})|}{e_V(\boldsymbol{f}, \boldsymbol{f}^{(p)})^{\alpha/r}} \leq \frac{c_2(p)}{c_1(p)^{\alpha/r}}$ for $\boldsymbol{f} \in \{\boldsymbol{f} : d(\boldsymbol{f}, \boldsymbol{f}^{(p)}) \leq \epsilon\}$. For $\boldsymbol{f} \notin \{\boldsymbol{f} : d(\boldsymbol{f}, \boldsymbol{f}^{(p)}) \leq \epsilon\}$, note that $e_V(\boldsymbol{f}, \boldsymbol{f}^{(p)}) \geq c_1(p)\epsilon^r$, and $|e(\boldsymbol{f}, \boldsymbol{f}^{(p)})|$ is bounded by 1. Then, we have $\frac{|e(\boldsymbol{f}, \boldsymbol{f}^{(p)})|}{e_V(\boldsymbol{f}, \boldsymbol{f}^{(p)})^{\alpha/r}} < \frac{1}{c_1(p)^{\alpha/r}\epsilon^\alpha}$, and (11) follows with the choice of $c(p) = \max\{\frac{c_2(p)}{c_1(p)^{\alpha/r}}, \frac{1}{c_1(p)^{\alpha/r}\epsilon^\alpha}\}$.

To prove (12), let the Bayes rule $\bar{\boldsymbol{f}}$ be $\bar{f}_j(\boldsymbol{x}) = kI(j = \arg\max_c \eta_c(\boldsymbol{x})) - 1$; $j = 1,\ldots,k$. By Lemma 5, $EV(\boldsymbol{f}^{(p)}, \boldsymbol{Z}) = El(\boldsymbol{f}^{(p)}, \boldsymbol{Z}) = 0$. Note that $\bar{\boldsymbol{f}}$ minimizes both $EV(\boldsymbol{f}, \boldsymbol{Z})$ and $El(\boldsymbol{f}, \boldsymbol{Z})$ over all measurable decision functions, which implies that $El(\boldsymbol{f}^{(p)}, \boldsymbol{Z}) = El(\bar{\boldsymbol{f}}, \boldsymbol{Z}) = 0$, and $EV(\boldsymbol{f}^{(p)}, \boldsymbol{Z}) = EV(\bar{\boldsymbol{f}}, \boldsymbol{Z}) = 0$. Then, (12) follows from Lemma 7. This completes the proof. □.

Lemma 3 controls the complexity of $\mathcal{H}$ in terms of the $L_2(Q)$-metric entropy for any class of functions $\mathcal{B}$, which is defined as follows. Given any $\varepsilon > 0$, a set $\{g_i\}_1^m$ is called an ε-net of $\mathcal{B}$, if for any $f \in \mathcal{B}$, there exists a g_k such that $\|g_k - f\|_{Q,2} \leq \varepsilon$, where $\|\cdot\|_{Q,2}$ is the $L_2(Q)$-norm defined as $\|f\|_{Q,2} = (\int f^2 dQ)^{1/2}$. The $L_2(Q)$-covering number $N(\varepsilon, \mathcal{B}, L_2(Q))$ is defined as the minimal size of all ε-nets. The $L_2(Q)$-metric entropy

$H(\varepsilon, \mathcal{B}, L_2(Q))$ is the logarithm of the covering number. Further define the uniform covering number and uniform metric entropy as $N_2(\varepsilon, \mathcal{B}) = \sup_Q N(\varepsilon, \mathcal{B}, L_2(Q))$ and $H_2(\varepsilon, \mathcal{B}) = \sup_Q H(\varepsilon, \mathcal{B}, L_2(Q))$, respectively. The following lemma gives an upper bound of $H_2(\varepsilon, \mathcal{H})$.

Lemma 3 *For any $\varepsilon > 0$, $H_2(\varepsilon, \mathcal{H}) \leq \frac{2}{\varepsilon^2} \log(e + e(2k(p+1))\varepsilon^2)$.*

Proof: For simplicity, expand predictor vector $\boldsymbol{x}$ and coefficient vector $\boldsymbol{w}_c$ to $\dot{\boldsymbol{x}} = (1, \boldsymbol{x}^T)^T$ and $\dot{\boldsymbol{w}}_c = (b_c, \boldsymbol{w}_c^T)^T$. Now consider $\mathcal{F} = \{\boldsymbol{f} : \sum_{c=1}^k \|\dot{\boldsymbol{w}}_c\|_1 \leq 3s + 2k\} \supset \mathcal{F}^b(p, s)$, and $\mathcal{G} = \{V(\boldsymbol{f}, \cdot) : \boldsymbol{f} \in \mathcal{F}\}$. Evidently, an entropy bound of $\mathcal{H}$ can be obtained by that of $\mathcal{G}$. We proceed to bound $\mathcal{G}$. To construct an ε-net on $\mathcal{G}$, first examine the relation between $\mathcal{G}$ and $\mathcal{F}$. For $g(\boldsymbol{z}) = V(\boldsymbol{f}, \boldsymbol{z})$ and $g'(\boldsymbol{z}) = V(\boldsymbol{f}', \boldsymbol{z}) \in \mathcal{G}$,

$$\begin{aligned} \|g - g'\|_{Q,2}^2 &= E(\sum_{c \neq Y} [f_c(X) + 1]_+ - \sum_{c \neq Y} [f_c'(X) + 1]_+)^2 \leq E(\sum_{c=1}^k |f_c(X) - f_c'(X)|)^2 \\ &\leq k \sum_c \|f_c - f_c'\|_{Q,2}^2 \leq k^2 \|\tilde{f} - \tilde{f}'\|_{\tilde{Q},2}^2. \end{aligned} \tag{16}$$

In (16), $\tilde{f}(\tilde{\boldsymbol{x}}) = \sum_{c=1}^k f_c(\boldsymbol{x}_c) = \sum_{c=1}^k \dot{\boldsymbol{w}}_c^T \dot{\boldsymbol{x}}_c$, for each $\tilde{\boldsymbol{x}} = (\dot{\boldsymbol{x}}_1, \dot{\boldsymbol{x}}_2, \ldots, \dot{\boldsymbol{x}}_k)$ with $\dot{\boldsymbol{x}}_c \in \mathbb{R}^{p+1}$, $\tilde{f}'(\tilde{\boldsymbol{x}}) = \sum_{c=1}^k f_c'(\boldsymbol{x}_c)$ is defined in the same manner, and $\tilde{Q}$ is defined as the distribution of $\tilde{\boldsymbol{X}} = (\delta_1 \dot{\boldsymbol{X}}_1, \ldots, \delta_k \dot{\boldsymbol{X}}_k)$, where $\dot{\boldsymbol{X}}_c = (1, \boldsymbol{X}_c^T)^T$ with $\boldsymbol{X}_c$ i.i.d. random vectors with distribution Q, and $(\delta_1, \ldots, \delta_k)$ has a joint distribution $P((\delta_1, \ldots, \delta_k)^T = \mathbf{1}_c) = k^{-1}$ with $\mathbf{1}_c$ a vector consisting of 1 at the c-th entry and 0 at the remaining $k-1$ entries. As a result, the equality $\|g - g'\|_{Q,2} \leq k\|\tilde{f} - \tilde{f}'\|_{\tilde{Q},2}$ establishes a relation between $\mathcal{G}$ and function class $\tilde{\mathcal{F}} = \{\tilde{f} : \tilde{f}(\tilde{\boldsymbol{x}}) = \sum_{c=1}^k \sum_{j=0}^p \dot{w}_{c,j} \dot{x}_c^{(j)}; \sum_{c,j} |\dot{w}_{c,j}| \leq 3s + 2k\}$.

To bound $\tilde{\mathcal{F}}$, let $\tilde{f}_{c,j}(\tilde{\boldsymbol{x}}) = (3s+2k)\, \dot{x}_c^{(j)}$. Then $\mathcal{D} = \{\pm \tilde{f}_{c,j}\}$ consists a basis for $\tilde{\mathcal{F}}$ and each $\tilde{f} = \sum_{c=1}^k \sum_{j=0}^p \dot{w}_{c,j} \dot{x}_c^{(j)} = \sum_{c=1}^k \sum_{j=0}^p \frac{|\dot{w}_{c,j}|}{3s+2k} (\text{sign}(\dot{w}_{c,j}) \tilde{f}_{c,j}(\tilde{\boldsymbol{x}}))$ is a convex combination of $\pm \tilde{f}_{c,j}$; $c = 1, \ldots, k$, $j = 0, \ldots, p$. Thus, $\tilde{\mathcal{F}} = \text{conv}\mathcal{D}$, the convex hull of $\mathcal{D}$. By Lemma 2.6.11 of Van der Vaart and Wellner (2000), $N(\varepsilon \text{diam}\mathcal{D}, \text{conv}\mathcal{D}, L_2(\tilde{Q})) \leq (e + e(2k(p+1))\varepsilon^2)^{2/\varepsilon^2}$, where $\text{diam}\mathcal{D} = \max_{u,v \in \mathcal{D}} \|u - v\|_{\tilde{Q},2} \leq 2(3s + 2k)$.

From (16), a $2(3s + 2k)\varepsilon$-net in $\tilde{\mathcal{F}}$ induces a $2k(3s + 2k)\varepsilon$-net in $\mathcal{G}$, implying

$N(2k(3s+2k)\varepsilon, \mathcal{G}, L_2(Q)) \leq (e + e(2k(p+1))\varepsilon^2)^{2/\varepsilon^2}$. Because $H_2(\varepsilon, \mathcal{H}) \leq H_2(2k(3s+2k)\varepsilon, \mathcal{G})$, the result then follows. □

In Lemma 3, we derive an upper bound of the random entropy instead of the bracketing entropy to quantify the complexity of $\mathcal{H}$. Based on Lemma 3, we establish a probability tail bound of $\sup_{\mathcal{H}} |P_n h - Ph|$ as in Lemma 4. In the literature, similar bounds are obtained under the condition of $H_2(\varepsilon, \mathcal{H}) \leq A\varepsilon^{-W}$ with $W < 2$. To our knowledge, no result is available for random entropy with $W = 2$. Therefore, we need to derive a bound.

Lemma 4 *Assume that n, p, k, M and ε_0 satisfy (13) and (14). Then*

$$P^*(\sup_{\mathcal{H}} |P_n h - Ph| > 4M) \leq 6[1 - \frac{1}{16nM^2}]^{-1} \exp(-2(1-\theta)nM^2),$$

where P^* denotes the outer probability.

Proof: The proof uses conditioning and chaining.

Fist, we bound the probability of interest by a corresponding tail probability of sampling n observations without replacement from the empirical measure of a sample of size $N = mn$, with $m = 2$. Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ be an i.i.d. sample from P, and let $(R_1, \ldots, R_N)$ be uniformly distributed on the set of permutations of $(1, \ldots, N)$. Define $n' = N - n$, $\tilde{P}_{n,N} = n^{-1} \sum_{i=1}^{n} \delta_{\boldsymbol{Z}_{R_i}}$, and $P_N = N^{-1} \sum_{i=1}^{N} \delta_{\boldsymbol{Z}_i}$, with $\delta_{\boldsymbol{Z}_i}$ the Dirac measure at observation $\boldsymbol{Z}_i$. Then, we have the following inequality, which can be treated as an alternative to the classical symmetrization inequality (c.f., Van der Vaart and Wellner 2000, Lemma 2.14.18 with $a = 2^{-1}$ and $m = 2$),

$$P^*(\sup_{\mathcal{H}} |P_n h - Ph| > 4M) \leq [1 - \frac{1}{16nM^2}]^{-1} P^*(\sup_{\mathcal{H}} |\tilde{P}_{n,N} h - P_N h| > M). \qquad (17)$$

Conditioning on $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$, we consider $P^*(\sup_{\mathcal{H}} |\tilde{P}_{n,N} h - P_N h| > M)$, and denote by $P_{|N}$ the conditional distribution given $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$. Let $\varepsilon_0 > \varepsilon_1 > \ldots > \varepsilon_T > 0$, which is to be chosen later on. Denote by $\mathcal{H}_q$ the minimal ε_q-net for $\mathcal{H}$ with

respect to the $L_2(P_N)$ norm. For each h, let $\pi_q h = \arg\min_{g\in\mathcal{H}_q} \|g-h\|_{P_N,2}$. Evidently, $\|\pi_q h - h\|_{P_N,2} \le \varepsilon_q$, and $|\mathcal{H}_q| = N(\varepsilon_q, \mathcal{H}, L_2(P_N))$. Hence

$$\begin{aligned}
P^*_{|N}(\sup_{\mathcal{H}} |\tilde{P}_{n,N}h - P_N h| > M) &\le P^*_{|N}(\sup_{\mathcal{H}} |(\tilde{P}_{n,N} - P_N)(\pi_0 h)| > (1-\frac{\theta}{4})M) \\
&\quad + P^*_{|N}(\sup_{\mathcal{H}} |(\tilde{P}_{n,N} - P_N)(\pi_0 h - \pi_T h)| > \frac{\theta M}{8}) \\
&\quad + P^*_{|N}(\sup_{\mathcal{H}} |(\tilde{P}_{n,N} - P_N)(\pi_T h - h)| > \frac{\theta M}{8}) \\
&\le |\mathcal{H}_0| \sup_{\mathcal{H}} P^*_{|N}(|(\tilde{P}_{n,N} - P_N)(\pi_0 h)| > (1-\frac{\theta}{4})M) \\
&\quad + \sum_{q=1}^{T} |\mathcal{H}_q||\mathcal{H}_{q-1}| \sup_{\mathcal{H}} P^*_{|N}(|(\tilde{P}_{n,N} - P_N)(\pi_q h - \pi_{q-1} h)| > \eta_q) \\
&\quad + \sup_{\mathcal{H}} P^*_{|N}(|(\tilde{P}_{n,N} - P_N)(\pi_T h - h)| > \frac{\theta M}{8}) := \ P_1 + P_2 + P_3,
\end{aligned}$$

where η_q; $q = 1, \ldots, T$, are to be chosen such that

$$\sum_{q=1}^{T} \eta_j \le \theta M/8. \tag{18}$$

Now we bound P_1, P_2 and P_3 separately.

For P_3, take $\varepsilon_T \le \frac{\theta M}{8\sqrt{2(m+1)}}$. Note that $\tilde{P}_{n,N} f \le m P_N f$ for any non-negative f, and $P_N(\pi_T h - h)^2 \le \varepsilon_T^2$ by the definition of π_T. Then we have $|(\tilde{P}_{n,N} - P_N)(\pi_T h - h)|^2 \le 2(\tilde{P}_{n,N} + P_N)(\pi_T h - h)^2 \le 2(m+1)\varepsilon_T^2 \le (\theta M/8)^2$. So $P_3 = 0$.

For P_1, note that $0 \le \pi_0 h(\boldsymbol{Z}_i) \le 1$ for all $h \in \mathcal{H}$ and $i = 1, \ldots, N$. By Hoeffding's inequality (c.f., Hoeffding 1963, Theorem 2, Theorem 4), $P^*_{|N}(|(\tilde{P}_{n,N} - P_N)(\pi_0 h)| > (1-\frac{\theta}{4})M) \le 2\exp(-2n(1-\theta/4)^2 M^2$. Then, take ε_0 as defined in (14). By Lemma 3,

$$H(\varepsilon_0, \mathcal{H}, L_2(P_N)) \le 2\varepsilon_0^{-2} \log(e + e(2k(p+1))\varepsilon_0^2) = \theta(2nM^2)/4. \tag{19}$$

Therefore, $P_1 \le 2\exp(H(\varepsilon_0, \mathcal{H}, L_2(P_N)))\exp(-2n(1-\theta/4)^2 M^2) \le 2\exp(-(1-\theta)2nM^2)$.

For P_2, if $\epsilon_0 \le \frac{\theta M}{8\sqrt{2(m+1)}}$, let $\varepsilon_T = \varepsilon_0$. Then $P_2 = 0$ and the proof is trivial. Otherwise, we consider the case $\epsilon_0 > \frac{\theta M}{8\sqrt{2(m+1)}} > \epsilon_T$ in P_2. Note that $P_N(\pi_q h - \pi_{q-1} h)^2 \le 2(P_N(\pi_q h - h)^2 + P_N(h - \pi_{q-1} h)^2) \le 2\varepsilon_q^2 + 2\varepsilon_{q-1}^2 \le 4\varepsilon_{q-1}^2$. Then by Massart's inequal-

ity (c.f., Van der Vaart and Wellner 2000, Lemma 2.14.19), $P^*_{|N}(|(\tilde{P}_{n,N} - P_N)(\pi_q h - \pi_{q-1}h)| > \eta_q) \leq 2\exp(-n\eta_q^2/(m\sigma_N^2))$ with $\sigma_N^2 = P_N(\pi_q h - \pi_{q-1}h)^2 \leq 4\varepsilon_{q-1}^2$, and it follows that

$$\begin{aligned} P_2 &\leq \sum_{q=1}^{N} |\mathcal{H}_q|^2 \sup_{\mathcal{H}} P^*_{|N}(|(\tilde{P}_{n,N} - P_N)(\pi_q h - \pi_{q-1}h)| > \eta_q) \\ &\leq 2\sum_{q=1}^{T} \exp(2H(\varepsilon_q, \mathcal{H}, L_2(P_N)) - \frac{n\eta_q^2}{4m\varepsilon_{q-1}^2}) \\ &\leq 2\sum_{q=1}^{T} \exp(2\frac{2}{\varepsilon_q^2}\log(e + e(2k(p+1))\varepsilon_0^2) - \frac{n\eta_q^2}{4m\varepsilon_{q-1}^2}). \end{aligned}$$

Take $\varepsilon_q = 2^{-q}\varepsilon_0$; $q = 0, \ldots, T$, where $T = \lceil \log_2 \frac{8\varepsilon_0\sqrt{2(m+1)}}{\theta M} \rceil$ (such that $\varepsilon_T \leq \frac{\theta M}{8\sqrt{2(m+1)}} \leq \varepsilon_{T-1}$). Then take $\eta_q = (\frac{16m\varepsilon_{q-1}^2 \log(e+e(2k(p+1))\varepsilon_0^2)}{\varepsilon_q^2 n\theta})^{1/2} = (\frac{64m\log(e+e(2k(p+1))\varepsilon_0^2)}{n\theta})^{1/2}$. By (13), M satisfies

$$\sum_{q=1}^{T} \eta_j = T\eta_1 \leq (\log_2 \frac{8\varepsilon_0\sqrt{2(m+1)}}{\theta M} + 1)(\frac{64m\log(e + e(2k(p+1))\varepsilon_0^2)}{n\theta})^{1/2} \leq \theta M/8.$$

Then, we have

$$\begin{aligned} P_2 &\leq 2\sum_{q=1}^{T} \exp((4 - 4/\theta)\frac{1}{\varepsilon_q^2}\log(e + e(2k(p+1))\varepsilon_0^2)) \\ &\leq 2\sum_{q=1}^{T} \exp((4 - 4/\theta)\frac{4^q}{\varepsilon_0^2}\log(e + e(2k(p+1))\varepsilon_0^2)) \\ &\leq 2\sum_{q=1}^{\infty} \exp((2 - 2/\theta)4^q\theta(2nM^2)/4) \leq 4\exp(-(1-\theta)(2nM^2)). \end{aligned}$$

Therefore, $P^*_{|N}(\sup_{\mathcal{H}} |\tilde{P}_{n,N}h - P_N h| > M) \leq 6\exp(-(1-\theta)(2nM^2))$. Take expectation with respect to $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ on both sides and use (17), we have

$$P^*(\sup_{\mathcal{H}} |P_n h - Ph| > 4M) \leq 6[1 - \frac{1}{16nM^2}]^{-1}\exp(-2(1-\theta)nM^2).$$

This completes the proof. □

Lemma 5 *For any given p, assume that $P(\boldsymbol{X}(p), Y)$ is regular in the sense that the marginal distribution of $\boldsymbol{X}(p)$ has a finite density $q(\boldsymbol{x}) \leq U$ with respect to the Lebesgue*

measure λ *on* $[0,1]^p$*, and* $P(B_c) > 0$*;* $c = 1, \ldots, k$*, where* $B_c = \{\boldsymbol{x} \in [0,1]^p : c = \arg\max_{1 \le j \le k} \eta_j(\boldsymbol{x})\}$ *with* $\eta_j(\boldsymbol{x}) = P(Y = j | \boldsymbol{X} = \boldsymbol{x})$*. If there exists a* p^* *such that* $\boldsymbol{f}^{(p^*)} \notin \mathcal{F}(p^*)$*, then* $\inf_{\boldsymbol{f} \in \mathcal{F}(p)} EV(\boldsymbol{f}, \boldsymbol{Z}) = \inf_{\boldsymbol{f} \in \mathcal{F}(p)} El(\boldsymbol{f}, \boldsymbol{Z}) = 0$ *for all* $p \ge p^*$*.*

Proof of Lemma 5: First prove the case of $p = p^*$.

For any $\boldsymbol{f} \in \mathcal{F}(p)$, define $\|\boldsymbol{f}\|_1 = \sum_{c=1}^k (\sum_{j=1}^p |w_{c,j}| + |b_c|)$ and $\|f_c\|_1 = \sum_{j=1}^p |w_{c,j}| + |b_c|$; $c = 1, \ldots, k$. Because $\boldsymbol{f}^{(p)} \notin \mathcal{F}(p)$, by the definition of $\boldsymbol{f}^{(p)}$ and $\mathcal{F}(p)$, there exists a sequence $\{\boldsymbol{g}^{(n)}\}_{n=1}^{\infty} \in \mathcal{F}(p)$, such that $\|\boldsymbol{g}^{(n)}\|_1 \to \infty$ and $EV(\boldsymbol{g}^{(n)}, \boldsymbol{Z}) \to EV(\boldsymbol{f}^{(p)}, \boldsymbol{Z}) < \infty$ as $n \to \infty$. We will prove that $\lim_{n\to\infty} P(\cup_{c=1}^k M_c(\boldsymbol{g}^{(n)})) = 0$, where $M_c(\boldsymbol{g}^{(n)}) = \{\boldsymbol{z} = (\boldsymbol{x}, y) : y \ne c \text{ and } g_c^{(n)}(\boldsymbol{x}) \ge 0\}$ can be thought of as the misclassification set for $g_c^{(n)}$, $c = 1, \ldots, k$.

We prove by contradiction. Suppose $\lim_{n\to\infty} P(\cup_{c=1}^k M_c(\boldsymbol{g}^{(n)})) = 0$ does not hold. Then there exist constants $\epsilon_0 > 0$, c^* and a subsequence $\{\boldsymbol{g}^{(n_l)}\}_{l=1}^{\infty}$ such that $P(M_{c^*}(\boldsymbol{g}^{(n_l)})) > \epsilon_0$ for all l. Without loss of generality, assume $c^* = 1$ and $P(M_1(\boldsymbol{g}^{(n)})) > \epsilon_0$ for all n. By the definition of $V(\boldsymbol{f}, \boldsymbol{z})$ in (4) and the definition of $M_1(\boldsymbol{g}^{(n)})$,

$$
\begin{aligned}
EV(\boldsymbol{g}^{(n)}, \boldsymbol{Z}) &= E(\sum_{c \ne Y} [1 + g_c^{(n)}(\boldsymbol{X})]_+) \\
&= E(\sum_{c=1}^k [1 + g_c^{(n)}(\boldsymbol{X})]_+ I[Y \ne c]) \ge E([1 + g_1^{(n)}(\boldsymbol{X})]_+ I[Y \ne 1]) \\
&\ge E([1 + g_1^{(n)}(\boldsymbol{X})]_+ I[\boldsymbol{Z} \in M_1(\boldsymbol{g}^{(n)})]) \ge (1+t) E(I[g_1^{(n)}(\boldsymbol{X}) \ge t] I[\boldsymbol{Z} \in M_1(\boldsymbol{g}^{(n)})]) \\
&= (1+t)(P(M_1(\boldsymbol{g}^{(n)})) - P(\{g_1^{(n)}(\boldsymbol{x}) < t\} \cap M_1(\boldsymbol{g}^{(n)}))) \\
&\ge (1+t)(\epsilon_0 - P(0 \le g_1^{(n)}(\boldsymbol{x}) < t)), \qquad (20)
\end{aligned}
$$

for any $t > 0$. We discuss two cases: $\|g_1^{(n)}\|_1 \to \infty$ and $\|g_1^{(n)}\|_1 \nrightarrow \infty$.

Suppose $\|g_1^{(n)}\|_1 \to \infty$. By Lemma 6, $\lambda(\boldsymbol{x} \in [0,1]^p : 0 \le g_1^{(n)}(\boldsymbol{x}) < t) \le \lambda(\boldsymbol{x} \in [0,1]^p : |g_1^{(n)}(\boldsymbol{x})| < t) \le At/(\|g_1^{(n)}\|_1 - t) \to 0$, which implies $P(0 \le g_1^{(n)}(\boldsymbol{x}) < t) \to 0$ in (20). Hence $EV(\boldsymbol{f}^{(p)}, \boldsymbol{Z}) = \lim_{n\to\infty} EV(\boldsymbol{g}^{(n)}, \boldsymbol{Z}) \ge (t+1)\epsilon_0$, which contradicts to the fact that $EV(\boldsymbol{f}^{(p)}, \boldsymbol{Z}) < \infty$, because t can be arbitrarily large.

Suppose $\|g_1^{(n)}\|_1 \nrightarrow \infty$. Without loss of generality, assume there exists an $0 < D < \infty$, such that $\|g_1^{(n)}\|_1 \leq D$ for all n. Note that

$$\begin{aligned} EV(\boldsymbol{g}^{(n)}, \boldsymbol{Z}) &= E(\sum_{c=1}^{k}(1-\eta_c(\boldsymbol{X}))[1+g_c^{(n)}(\boldsymbol{X})]_+) \geq E(\sum_{c\neq 1}(1-\eta_c(\boldsymbol{X}))[1+g_c^{(n)}(\boldsymbol{X})]_+) \\ &\geq E(\sum_{c\neq 1} 1/2[1+g_c^{(n)}(\boldsymbol{X})]_+ I[\boldsymbol{X}\in B_1]) \geq 2^{-1}E(\max_{c\neq 1}[1+g_c^{(n)}(\boldsymbol{X})]_+ I[X\in B_1]) \\ &\geq 2^{-1}(t+1)E(I[\max_{c\neq 1} g_c^{(n)}(\boldsymbol{X}) \geq t] I[\boldsymbol{X}\in B_1]) \\ &\geq 2^{-1}(t+1)(P(B_1) - P(\max_{c\neq 1} g_c^{(n)}(\boldsymbol{X}) < t)), \end{aligned} \tag{21}$$

where the second inequality holds because $1-\eta_c(\boldsymbol{x}) \geq 1/2$ for all $c \neq 1$ and $\boldsymbol{x} \in B_1$ by the definition of B_1. Now we prove $P(\max_{c\neq 1} g_c^{(n)}(\boldsymbol{X}) < t) \to 0$ in (21). Take $t > D$. Because $|g_1^{(n)}(\boldsymbol{x})| \leq \|g_1^{(n)}\|_1 \leq D < t$ for all $\boldsymbol{x} \in [0,1]^p$, $\{\max_{c\neq 1} g_c^{(n)}(\boldsymbol{x}) \leq t\} \subset \{\max_c g_c^{(n)}(\boldsymbol{x}) \leq t\}$. Note that through the zero-sum constraint $\sum_{c=1}^k g_c^{(n)}(\boldsymbol{x}) = 0$, $\sum_{c=1}^k |g_c^{(n)}(\boldsymbol{x})| = 2\sum_{c=1}^k [g_c^{(n)}(\boldsymbol{x})]_+ \leq 2k \max_c g_c^{(n)}(\boldsymbol{x})$. By Lemma 6,

$$\begin{aligned} &\lambda(\boldsymbol{x}\in[0,1]^p : \max_{c\neq 1} g_c^{(n)}(\boldsymbol{x}) \leq t) \leq \lambda(\boldsymbol{x}\in[0,1]^p : \max_c g_c^{(n)}(\boldsymbol{x}) \leq t) \\ &\leq \lambda(\boldsymbol{x}\in[0,1]^p : \sum_{c=1}^k |g_c^{(n)}(\boldsymbol{x})| \leq 2kt) \leq \min_c \lambda(\boldsymbol{x}\in[0,1]^p : |g_c^{(n)}(\boldsymbol{x})| \leq 2kt) \\ &\leq A\frac{2kt}{\max_c \|g_c^{(n)}\|_1 - 2kt} \leq A\frac{2k^2 t}{\|\boldsymbol{g}^{(n)}\|_1 - 2k^2 t} \to 0, \end{aligned}$$

where the last inequality holds because $\max_c \|g_c^{(n)}\|_1 \geq \|\boldsymbol{g}^{(n)}\|_1/k$. Hence in (21), $P(\max_{c\neq 1} g_c^{(n)}(\boldsymbol{X}) < t) \to 0$, which implies $EV(\boldsymbol{f}^{(p)}, \boldsymbol{Z}) = \lim_{n\to\infty} EV(\boldsymbol{g}^{(n)}, \boldsymbol{Z}) \geq 2^{-1}(t+1)P(B_1) \to \infty$ as $t \to \infty$, which contradicts the fact that $EV(\boldsymbol{f}^{(p)}, \boldsymbol{Z}) < \infty$. Combining the two cases, $\lim_{n\to\infty} P(\cup_{c=1}^k M_c(\boldsymbol{g}^{(n)})) = 0$ is proved.

Now prove that $\inf_{\boldsymbol{f}\in\mathcal{F}(p)} El(\boldsymbol{f}, \boldsymbol{Z}) = 0$ and $\inf_{\boldsymbol{f}\in\mathcal{F}(p)} EV(\boldsymbol{f}, \boldsymbol{Z}) = 0$. To prove $\inf_{\boldsymbol{f}\in\mathcal{F}(p)} El(\boldsymbol{f}, \boldsymbol{Z}) = 0$, note that $\max_c g_c^{(n)}(x) \geq 0$. Then $\{\boldsymbol{z} : y \neq \arg\max_c g_c^{(n)}(\boldsymbol{x})\} \subset \cup_{c=1}^k M_c(\boldsymbol{g}^{(n)})$, and $\inf_{\boldsymbol{f}\in\mathcal{F}(p)} El(\boldsymbol{f}, \boldsymbol{Z}) \leq El(\boldsymbol{g}^{(n)}, \boldsymbol{Z}) = P(Y \neq \arg\max_c g_c^{(n)}(\boldsymbol{X})) \leq P(\cup_{c=1}^k M_c(\boldsymbol{g}^{(n)})) \to 0$. Therefore, $\inf_{\boldsymbol{f}\in\mathcal{F}(p)} El(\boldsymbol{f}, \boldsymbol{Z}) = 0$. For $\inf_{\boldsymbol{f}\in\mathcal{F}(p)} EV(\boldsymbol{f}, \boldsymbol{Z})$, note that the sequence $\{\boldsymbol{g}^{(n)}/\|\boldsymbol{g}^{(n)}\|_1\}_{n=1}^{\infty}$ is contained in a bounded closed set $\{\boldsymbol{f} : \|\boldsymbol{f}\|_1 \leq 1\}$

with respect to the norm $\|\boldsymbol{f}\|_1$. Consequently, there exists a limit point denoted by $\boldsymbol{f}^*$. Let $\boldsymbol{u}^{(n)} = n\boldsymbol{f}^*$. Because $P(\boldsymbol{Z} \in \cup_{c=1}^k M_c(\boldsymbol{f}^*)) = 0$, equivalently $P(\boldsymbol{Z} \notin \cup_{c=1}^k M_c(\boldsymbol{f}^*)) = 1$, and for all c, $\boldsymbol{z} \notin \cup_{c=1}^k M_c(\boldsymbol{f}^*)$ and $y \neq c$ implies $f_c^*(\boldsymbol{x}) < 0$,

$$\begin{aligned} EV(\boldsymbol{u}^{(n)}, \boldsymbol{Z}) &= \sum_{c=1}^k E([1 + nf_c^*(\boldsymbol{X})]_+ I[Y \neq c]) \\ &= \sum_{c=1}^k E((1 + nf_c^*(\boldsymbol{X}))I[1 + nf_c^*(\boldsymbol{X}) \geq 0]I[Y \neq c]) \\ &= \sum_{c=1}^k E((1 + nf_c^*(\boldsymbol{X}))I[1 + nf_c^*(\boldsymbol{X}) \geq 0]I[Y \neq c]I[\boldsymbol{Z} \notin \cup_{c=1}^k M_c(\boldsymbol{f}^*)]) \\ &\leq \sum_{c=1}^k E(I[0 \leq 1 + nf_c^*(\boldsymbol{X}) < 1]I[Y \neq c]I[\boldsymbol{Z} \notin \cup_{c=1}^k M_c(\boldsymbol{f}^*)]). \end{aligned}$$

Note that for all c, $P(0 \leq 1 + nf_c^*(\boldsymbol{X}) < 1) \leq P(|1 + nf_c^*(\boldsymbol{X})| < 1) \to 0$ by Lemma 6. We have $\inf_{\boldsymbol{f} \in \mathcal{F}(p)} EV(\boldsymbol{f}, \boldsymbol{Z}) \leq \lim_{n\to\infty} EV(\boldsymbol{u}^{(n)}, \boldsymbol{Z}) = 0$.

It remains to prove the case $p > p^*$. Note that $\mathcal{F}(p) \subset \mathcal{F}(p^*)$ for $p > p^*$, implying $\boldsymbol{g}^{(n)}, \boldsymbol{u}^{(n)} \in \mathcal{F}(p)$. Therefore, the arguments still hold. This completes the proof. □.

Lemma 6 *Let $f(\boldsymbol{x}) = \boldsymbol{\theta}^T\boldsymbol{x} + b$ with $\boldsymbol{\theta} \in \mathbb{R}^p$, $b \in \mathbb{R}$, and $\boldsymbol{x} \in [0,1]^p$, and define $\|f\|_1 = \|\theta\|_1 + |b|$. Then, there exists a constant $A > 0$ such that $\lambda(\boldsymbol{x} \in [0,1]^p : |f(\boldsymbol{x})| \leq t) \leq At/(\|f\|_1 - t)$ for all f and $t > 0$ with $\|f\|_1 - t > 0$.*

Proof of Lemma 6: For any $|b| > \|\boldsymbol{\theta}\|_1 + t$, $|f(\boldsymbol{x})| = |\boldsymbol{\theta}^T\boldsymbol{x} + b| \geq |b| - \|\boldsymbol{\theta}\|_1 > t$, which implies $\lambda(\boldsymbol{x} \in [0,1]^p : |f(\boldsymbol{x})| \leq t) = 0$, and thus the result holds. For any $|b| \leq \|\boldsymbol{\theta}\|_1 + t$, $\|\boldsymbol{\theta}\|_1 \geq (\|f\|_1 - t)/2$, when $\|f\|_1 - t > 0$. Note that $\|\boldsymbol{\theta}\|_2 \geq p^{-1/2}\|\boldsymbol{\theta}\|_1 = p^{-1/2}(\|f\|_1 - t)/2$. Then, $\lambda(\boldsymbol{x} \in [0,1]^p : |f(\boldsymbol{x})| \leq t)$ is

$$\begin{aligned} &\lambda(\boldsymbol{x} \in [0,1]^p : |(\boldsymbol{\theta}^T\boldsymbol{x} + b)|/\|\theta\|_2 \leq t/\|\theta\|_2) \\ \leq\ &\lambda\left(\boldsymbol{x} \in [0,1]^p : |(\boldsymbol{\theta}^T\boldsymbol{x} + b)| \leq \tfrac{2\|\theta\|_2 t}{p^{-1/2}(\|f\|_1 - t)}\right). \end{aligned} \tag{22}$$

Without loss of generality, assume $\boldsymbol{\theta}$ satisfies $\boldsymbol{\theta}^T\boldsymbol{\theta} = 1$. Then, for all $r > 0$, $\lambda(\boldsymbol{x} : -r \leq \boldsymbol{\theta}^T\boldsymbol{x} + b \leq r) = \int_{\boldsymbol{x}\in[0,1]^p} I[-r \leq \boldsymbol{\theta}^T\boldsymbol{x} + b \leq r]\lambda(d\boldsymbol{x})$. Now construct an orthogonal matrix $\boldsymbol{U}$ with $\boldsymbol{\theta}$ as the first column. The integration becomes $\int_{\boldsymbol{z}\in\{Q^T\boldsymbol{x}:\boldsymbol{x}\in[0,1]^p\}} I[-r \leq$

$z_1+b \le r]\lambda(d\boldsymbol{z})$ using the substitution $\boldsymbol{z} = \boldsymbol{Q}^T x$. Here, $\{Q^T\boldsymbol{x} : \boldsymbol{x} \in [0,1]^p\}$ is contained by $[-M, M]^p$ for sufficiently large M. Then, $\int_{\boldsymbol{z}\in\{Q^T\boldsymbol{x}:\boldsymbol{x}\in[0,1]^p\}} I[-r \le z_1+b \le r]\lambda(d\boldsymbol{z}) \le \int_{\boldsymbol{z}\in[-M,M]^p} I[-r \le z_1+b \le r]\lambda(d\boldsymbol{z}) \le (2M)^{p-1}2r$, and by (22) $\lambda(\boldsymbol{x} \in [0,1]^p : |f(\boldsymbol{x})| \le t) \le 4(2M)^{p-1}p^{1/2}t/(\|f\|_1 - t)$. The result holds with $A = 4(2M)^{p-1}p^{1/2}$. □

Lemma 7 *Define the Bayes rule* $\bar{\boldsymbol{f}} = (\bar{f}_1, \ldots, \bar{f}_k)$ *as* $\bar{f}_j(\boldsymbol{x}) = kI(j = arg\max_c \eta_c(\boldsymbol{x})) - 1$*; j* $= 1, \ldots, k$*. Then, for all measurable* $\boldsymbol{f}$*,* $e(\boldsymbol{f}, \bar{\boldsymbol{f}}) \le e_V(\boldsymbol{f}, \bar{\boldsymbol{f}})$.

Proof: The main treatment is to apply the conditioning argument to X to show that $E(l(\boldsymbol{f}, \boldsymbol{Z}) - l(\bar{\boldsymbol{f}}, \boldsymbol{Z})|\boldsymbol{X} = \boldsymbol{x}) \le E(V(\boldsymbol{f}, \boldsymbol{Z}) - V(\bar{\boldsymbol{f}}, \boldsymbol{Z})|\boldsymbol{X} = \boldsymbol{x})$ for all $\boldsymbol{x}$.

For any $\boldsymbol{x}$, write $\eta_c = P(Y = c|\boldsymbol{X} = \boldsymbol{x})$, $f_c = f_c(\boldsymbol{x})$, $\bar{f}_c = \bar{f}_c(\boldsymbol{x})$, $c^* = \arg\max_c f_c$ and $\bar{c} = \arg\max_c \eta_c$, for simplicity. Note that $\bar{f}_c = kI(c = \bar{c}) - 1$. Then we express both $E(l(\boldsymbol{f}, \boldsymbol{Z}) - l(\bar{\boldsymbol{f}}, \boldsymbol{Z})|\boldsymbol{X} = \boldsymbol{x})$ and $E(V(\boldsymbol{f}, \boldsymbol{Z}) - V(\bar{\boldsymbol{f}}, \boldsymbol{Z})|\boldsymbol{X} = \boldsymbol{x})$ in η_c, f_c, $\bar{f}_c$, c^* and $\bar{c}$ as follows. $E(l(\boldsymbol{f}, \boldsymbol{Z}) - l(\bar{\boldsymbol{f}}, \boldsymbol{Z})|\boldsymbol{X} = \boldsymbol{x}) = \sum_{c=1}^k \eta_c I(c \ne c^*) - \sum_{c=1}^k \eta_c I(c \ne \bar{c}) = \eta_{\bar{c}} - \eta_{c^*}$;

$$\begin{aligned} E(V(\boldsymbol{f}, \boldsymbol{Z}) - V(\bar{\boldsymbol{f}}, \boldsymbol{Z})|\boldsymbol{X} = \boldsymbol{x}) &= \sum_{c=1}^k \eta_c \sum_{j\ne c}[f_j + 1]_+ - \sum_{c=1}^k \eta_c \sum_{j\ne c}[\bar{f}_j + 1]_+ \\ &= \sum_{c=1}^k (1 - \eta_c)[f_c + 1]_+ - (1 - \eta_{\bar{c}})k. \end{aligned} \tag{23}$$

In (23), let $[f_c + 1]_+$ be t_c; $c = 1, \ldots, k$, and let t be $\sum_{c=1}^k t_k$. Then it follows from $(1 - \eta_c) \ge (1 - \eta_{\bar{c}})$ that

$$\sum_{c=1}^k (1 - \eta_c)[f_c + 1]_+ = (1 - \eta_{c^*})t_{c^*} + \sum_{c\ne c^*}(1 - \eta_c)t_c \ge (1 - \eta_{c^*})t_{c^*} + (1 - \eta_{\bar{c}})(t - t_{c^*}). \tag{24}$$

In (24), $t = \sum_{c=1}^k [f_c + 1]_+ \ge k[\frac{1}{k}\sum_{c=1}^k (f_c + 1)]_+ = k$ by convexity of $[x]_+$. Note that $\eta_{c^*} \le \eta_{\bar{c}}$ and $t_{c^*} \ge t_c$ for all c, implying $t_{c^*} \ge \frac{t}{k} \ge 1$. Thus $(1 - \eta_{c^*})t_{c^*} + (1 - \eta_{\bar{c}})(t - t_{c^*}) = (\eta_{\bar{c}} - \eta_{c^*})t_{c^*} + (1 - \eta_{\bar{c}})t \ge (\eta_{\bar{c}} - \eta_{c^*}) + (1 - \eta_{\bar{c}})k$. Therefore, by (23) and (24), $E(V(\boldsymbol{f}, \boldsymbol{Z}) - V(\bar{\boldsymbol{f}}, \boldsymbol{Z})|\boldsymbol{X} = \boldsymbol{x}) \ge \eta_{\bar{c}} - \eta_{c^*} = E(l(\boldsymbol{f}, \boldsymbol{Z}) - l(\bar{\boldsymbol{f}}, \boldsymbol{Z})|\boldsymbol{X} = \boldsymbol{x})$. The desired result then follows by taking the expectation with respect to $\boldsymbol{X}$. □

References

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

[2] Bradley, P. S. and Mangasarian, O. L. (1998) Feature selection via concave minimization and support vector machines. in J. Shaclik (ed.), *Machine Learing Proceedings of the Fifteenth International Conference*, Morgan Kaufmann, San Francisco, California, pp. 82-90.

[3] Bredensteiner, E. J. and Bennett, K. P. (1999) Multicategory classification by support vector machines. *Computaional Optimizations and Applications*, **12**, 35-46.

[4] Cristianini, N. and Shawe-Taylor, J. (2000) An introduction to support vector machines. Cambridge University Press, Cambridge, UK.

[5] Dudoit, S., Fridlyand, J. and Speed T. P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77-87.

[6] Guermeur, Y. (2002) Combining discriminant models with new multi-class SVMs. *Pattern Analysis and Applications*, **5**, 168-179.

[7] 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.

[8] Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, **58**, 13-30.

[9] Lee, Y., Kim, Y., Lee, S., and Koo, J. Y.. (2006) Structured Multicategory Support Vector Machine with ANOVA decomposition. *Biometrika*. To appear.

[10] Lee, Y. and Lee, C.-K. (2003) Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, **19**, 1132-1139.

[11] 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.

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

[13] Lin, Y., (2004) A Note on Margin-based Loss Functions in Classification. *Statistics and Probability Letters*, **68**, 73-82.

[14] Liu, Y. and Shen, X. (2005) Multicategory psi-learning. *Journal of the American Statistical Association*. To appear.

[15] Ng, A. Y. (2004) Feature selection, L1 vs. L2 regularization, and rotational invariance. *Proceedings of the Twenty-first International Conference on Machine Learning*.

[16] Shen, X. and Wong, W. (1994) Convergence rate of sieve estimates. *The Annals of Statistics*, **22**, 580-615.

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

[18] Szedmak, S., Shawe-Taylor, J., Saunders, C.J. and Hardoon, D.R. (2004) Multiclass classification by L1 norm support Vector Machine, In *Pattern Recognition and Machine Learning in Computer Vision Workshop*, 02-04 May 2004, Grenoble, France.

[19] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society Sery B*, **58**, 267-288.

[20] Tarigan, B and van de Geer, SA (2004) Adaptivity of support vector machines with l_1 penalty. Technical Report MI 2004-14, University of Leiden.

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

[22] Van der Vaart, A. W. and Wellner, J. A. (2000) *Weak Convergence and Empirical Processes with Application to Statistics*, Springer, New York.

[23] Wahba, G. (1999) Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In B. Scholkopf, C. J. C. Burges, and A. J. Smola (Eds.), *Advances in Kernel Methods: Support Vector Learning*, MIT Press 1999, 69-88.

[24] Wang, L., and Shen, X. (2005) Multi-category Support vector machines, feature selection, and solution path. *Statistica Sinica*, **16**, 617-634.

[25] Weston, J. and Watkins, C. (1999) Support vector machines for multiclass pattern recognition, *Proceedings of the Seventh European Symposium on Artificial Neural Networks*.

[26] Zhu, J., Hastie, T., Rosset, S. and Tibshirani, R. (2003) 1-norm support vector machines. *Neural Information Processing Systems*, **16**.

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

[28] Zhu, J., Rosset, S., Zou, H. and Hastie, T. (2005) A multi-class forward stage-wise generalization of AdaBoost. Technical Report 430, University of Michigan.

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

Table 1: Bayes errors, Averaged test errors in percentage and their standard errors (in parenthesis) as well as averaged number of variables selected and their standard errors (in parenthesis), over 100 simulation replications for L1MSVM and OVA, with $d = 1, 2, 3$ and $k = 1, 2$.

			L1MSVM		OVA	
classes	distance	Bayes	Test error	Variables	Test error	Variables
$k = 4$	$d = 1$	36.42%	42.20(0.09)%	2.20(0.05)	56.87(0.25)%	67.17(1.93)
	$d = 2$	14.47%	15.18(0.04)%	2.06(0.02)	16.21(0.09)%	5.72(0.38)
	$d = 3$	3.33%	3.35(0.02)%	2.02(0.01)	3.50(0.02)%	2.51(0.13)
$k = 8$	$d = 1$	64.85%	70.47(0.10)%	3.51(0.16)	79.76(0.07)%	98.18(0.29)
	$d = 2$	43.82%	46.86(0.12)%	3.02(0.12)	66.72(0.11)%	95.43(0.25)
	$d = 3$	25.06%	27.95(0.13)%	2.75(0.17)	55.84(0.12)%	93.37(0.21)

Table 2: Bayes errors, Averaged test errors in percentage and their standard errors (in parenthesis) over 100 simulation replications for L1MSVM and L2MSVM, with different values of p.

	Bayes	L1MSVM	L2MSVM
$p = 10$	10.81%	13.61(0.12)%	15.44(0.17)%
$p = 20$	10.81%	14.06(0.14)%	17.81(0.22)%
$p = 40$	10.81%	14.94(0.14)%	20.01(0.22)%
$p = 80$	10.81%	15.68(0.15)%	21.81(0.14)%
$p = 160$	10.81%	16.58(0.17)%	27.54(0.17)%

Table 3: Leukemia data: averaged test errors in percentage and their standard errors (in parenthesis) as well as averaged number of variables selected and their standard errors (in parenthesis), over 50 simulation replications for L1MSVM, OVA and L2MSVM.

	Test error	# Variables
L1MSVM	3.76(.51)%	20.92(.71)
OVA	6.24(.45)%	26.73(.80)
L2MSVM	4.20(.33)%	3571(.00)

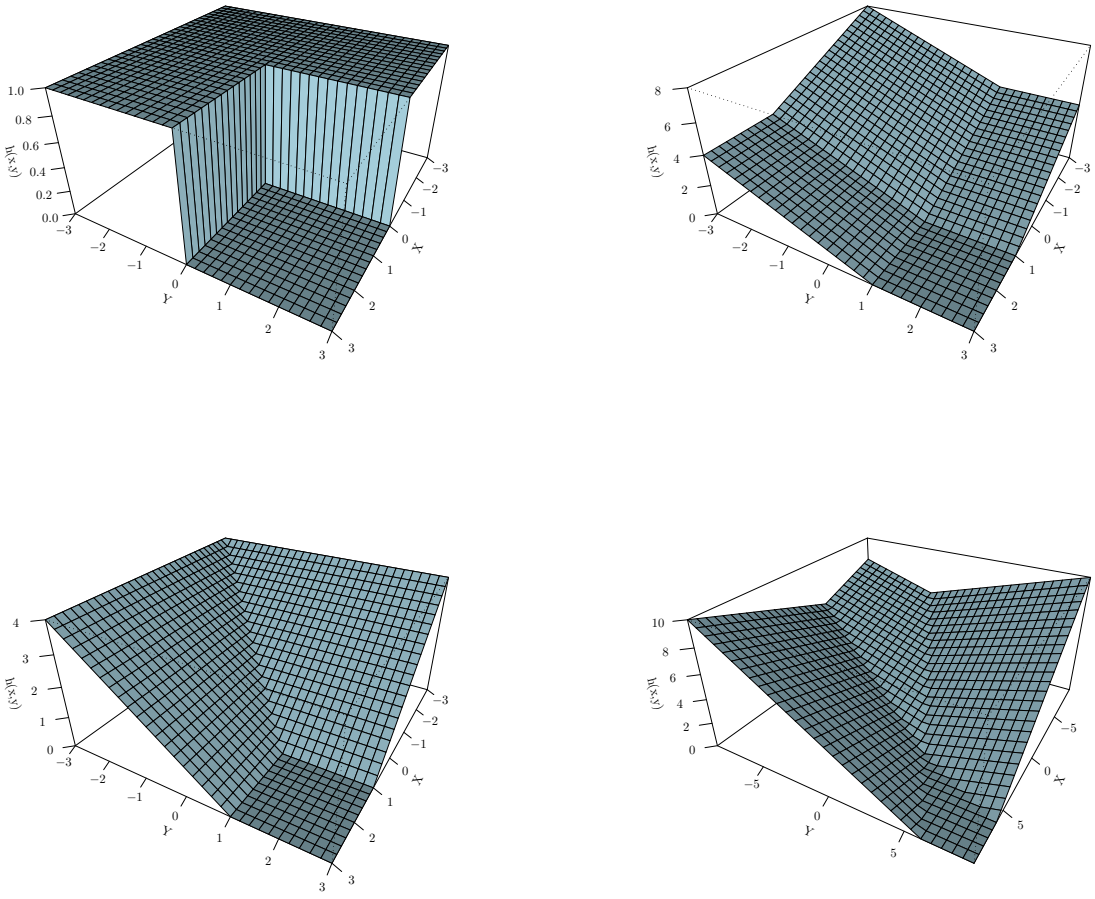

Figure 1: Perspective plots of h functions defined for 0-1 loss and generalized hinge losses (2)-(4) in three-class classification. Upper left: 0-1 loss; Upper right: hinge loss defined in (2); Bottom left: hinge loss defined in (3); Bottom right: hinge loss defined in (4).

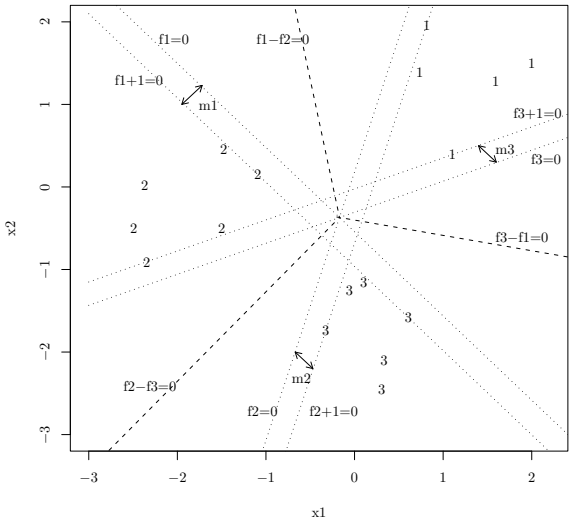

Figure 2: Illustration of the concept of margins in three-class separable classification: f_c separates class c from the rest classes; $c = 1, 2, 3$. The zero-sum constraint implies $f_c = 0$; $c = 1, 2, 3$ have a common intersection. With this constraint in place, the margin $\min_c m_c$ is maximized to obtain the decision functions f_i, which yields the separating boundary (dashed lines) $f_i - f_j = 0$; $i, j = 1, 2, 3; i \neq j$.

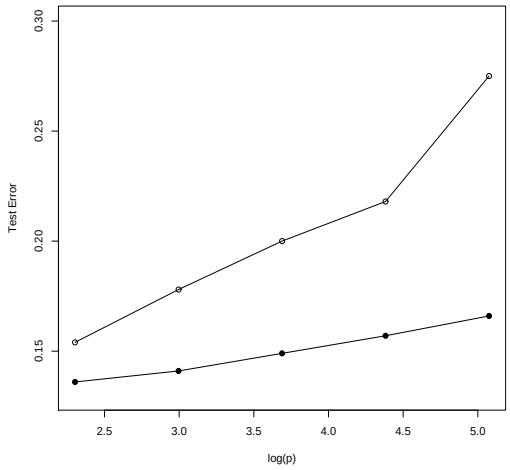

Figure 3: Averaged test errors of L1MSVM and L2MSVM as a function of $\log(p)$ with $p = 10, 20, 40, 80$, and 160. Solid circle and circle represent L1MSVM and L2MSVM, respectively.